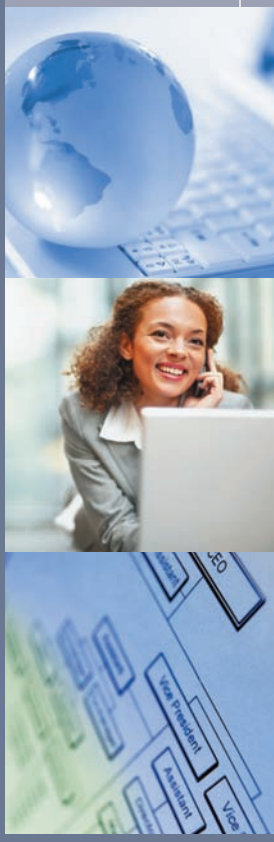
THE
Professional Practice
SERIES

John C. Scott
Douglas H. Reynolds
EDITORS

Handbook of Workplace Assessment

Evidence-Based Practices for Selecting and Developing Organizational Talent



A Publication of the Society for
Industrial and Organizational Psychology

"With the increased use of employee testing and other employee assessment devices and the increased legal challenges to those tests and assessments, this *Handbook* provides an extremely timely and enormously valuable resource for HR professionals and assessment professionals as well as an indispensable and unique reference for HR counsel who advise and defend employers in conjunction with their use of employee tests and other assessments."

—*Mark S. Dichter,* chair, Labor and Employment Practice,
Morgan, Lewis & Bockius LLP

"The *Handbook* is remarkably complete in addressing the complexities of selection research and practice over an exceptionally broad range of contexts and issues faced by those charged with organizational staffing."

—*Neal Schmitt,* chair, Department of Psychology,
Michigan State University

"This volume provides human resource professionals and executives with leading-edge and innovative approaches to assessment that will enhance organizational effectiveness."

—*Ben E. Dowell,* vice president of Talent Management
(retired), Bristol-Myers Squibb

"This is an invaluable resource, with sound, practical guidelines steeped in empirical research for implementing an assessment process that will effectively drive an organization's critical talent decisions."

—*David A. Rodriguez,* executive vice president,
Global Human Resources, Marriott International, Inc.

"This is the only 'go-to' guide for decision makers who need to plan for their current and future workforce to remain competitive on a global basis."

—*Peter M. Fasolo,* Ph.D., chief talent
officer, Portfolio Companies
Kohlberg Kravis Roberts & Company

"The editors' stated purpose for the *Handbook* was to present technically sound, research-based assessment procedures that engage the full spectrum of individual assessment objectives that organizations face when attempting to maximize their human talent. They succeeded. The coverage is broad, deep, and accessible to a wide audience. It examines our most fundamental assessment issues from a variety of perspectives and in a variety of contexts. It covers the landscape, and the differences across perspectives are informative, even for a hard-core academic. *Read* it."

—*John Campbell*, professor of Psychology and Industrial Relations, University of Minnesota

Handbook of Workplace Assessment

The Professional Practice Series

The Professional Practice Series is sponsored by The Society for Industrial and Organizational Psychology, Inc. (SIOP). The series was launched in 1988 to provide industrial and organizational psychologists, organizational scientists and practitioners, human resources professionals, managers, executives, and those interested in organizational behavior and performance with volumes that are insightful, current, informative, and relevant to *organizational practice*. The volumes in the Professional Practice Series are guided by five tenets designed to enhance future organizational practice:

1. Focus on practice, but grounded in science
2. Translate organizational science into practice by generating guidelines, principles, and lessons learned that can shape and guide practice
3. Showcase the application of industrial and organizational psychology to solve problems
4. Document and demonstrate best industrial and organizational-based practices
5. Stimulate research needed to guide future organizational practice

The volumes seek to inform those interested in practice with guidance, insights, and advice on how to apply the concepts, findings, methods, and tools derived from industrial and organizational psychology to solve human-related organizational problems.

Previous Professional Practice Series volumes include:

Published by Jossey-Bass

Going Global: Practical Applications and Recommendations for HR and OD Professionals in the Global Workplace
Kyle Lundby with Jeffrey Jolton

Strategy-Driven Talent Management: A Leadership Imperative
Rob Silzer, Ben E. Dowell, Editors

Performance Management: Putting Research into Practice
James W. Smither, Manuel London, Editors

Alternative Validation Strategies: Developing New and Leveraging Existing Validity Evidence
S. Morton McPhail

Getting Action from Organizational Surveys: New Concepts, Technologies, and Applications
Allen I. Kraut

Customer Service Delivery
Lawrence Fogli, Editor

Employment Discrimination Litigation
Frank J. Landy, Editor

The Brave New World of eHR
Hal G. Gueutal, Dianna L. Stone, Editors

Improving Learning Transfer in Organizations
Elwood F. Holton III, Timothy T. Baldwin, Editors

Resizing the Organization
Kenneth P. De Meuse, Mitchell Lee Marks, Editors

Implementing Organizational Interventions
Jerry W. Hedge, Elaine D. Pulakos, Editors

Organization Development
Janine Waclawski, Allan H. Church, Editors

Creating, Implementing, and Managing Effective Training and Development
Kurt Kraiger, Editor

The 21st Century Executive: Innovative Practices for Building Leadership at the Top
Rob Silzer, Editor

Managing Selection in Changing Organizations
Jerard F. Kehoe, Editor

Evolving Practices in Human Resource Management
Allen I. Kraut, Abraham K. Korman, Editors

Individual Psychological Assessment: Predicting Behavior in Organizational Settings
Richard Jeanneret, Rob Silzer, Editors

Performance Appraisal
James W. Smither, Editor

Organizational Surveys
Allen I. Kraut, Editor

Employees, Careers, and Job Creating
Manuel London, Editor

Published by Guilford Press

Diagnosis for Organizational Change
Ann Howard and Associates

Human Dilemmas in Work Organizations
Abraham K. Korman and Associates

Diversity in the Workplace
Susan E. Jackson and Associates

Working with Organizations and Their People
Douglas W. Bray and Associates

Handbook of Workplace Assessment

The Professional Practice Series

Handbook of Workplace Assessment

Evidence-Based Practices for Selecting and Developing Organizational Talent

John C. Scott
Douglas H. Reynolds, Editors

Foreword by Allan H. Church and Janine Waclawski

JOSSEY-BASS
A Wiley Imprint
www.josseybass.com

Published by Jossey-Bass
A Wiley Imprint
989 Market Street, San Francisco, CA 94103-1741—www.josseybass.com

Library of Congress Cataloging-in-Publication Data

Handbook of workplace assessment : evidence-based practices for selecting and developing organizational talent / John C. Scott, Douglas H. Reynolds, editors ; foreword by Allan H. Church. — 1st ed.
p. cm. — (The professional practice series)
Includes bibliographical references and index.
ISBN 978-0-470-40131-6
1. Employees—Rating of. 2. Needs assessment. 3. Organizational change. 4. Personnel management. I. Scott, John C. (John Carlson), 1955– II. Reynolds, Douglas H.
HF5549.5.R3H28 2010
658.3'124—dc22

2010003886

Printed in the United States of America
FIRST EDITION
HB Printing 10 9 8 7 6 5 4 3 2 1

Contents

Figures, Tables, and Exhibits

Figures

Tables

Exhibits

Foreword

Welcome to the newest volume in the Professional Practice book series of the Society for Industrial and Organizational Psychology (SIOP). We are very excited about this volume and the contribution that we believe it will make not only to the series overall but also to the field in general.

The idea for this book came out of one of our first editorial board meetings at an annual SIOP meeting about six or seven years ago. The approach during our years as series coeditors was to call our board together (since we typically had a quorum at the annual conference) to meet and discuss the trends and practices we were seeing in the field. We talked about sessions we had seen at the conference that were good, bad, or ugly and used these thoughts as fodder to brainstorm ideas for what we hoped would be great future volumes for this series. For the most part, the output of those brainstorming sessions came to fruition in the form of several volumes of which we are very proud. This book is one that we have had a lot of passion and anticipation for since those early days. However, we also recognized that completing this task would require a lot of effort, insight, and dedication to put together under the right volume editors. Luckily for us, it all fell into place under the editorship of John Scott and Doug Reynolds. They have done a fantastic job of surveying the simultaneously broad and deep field of assessment and putting it all together in one place under a simple yet elegant framework.

Talent identification and assessment is one of the most critical issues facing organizations today. From our vantage point as practitioners (one of us as an organization development specialist and the other as a human resource generalist), we see this as a major challenge. A good or bad hire in isolation can have a long-lasting organizational impact (think about your personal

experiences here), and in the aggregate, its impact is profound: it determines not only the organizational culture but also ultimately its success or failure. In this way, assessment is key to our practice as I-O professionals. The concept behind this volume is to provide internal and external practitioners with a much-needed compendium of tools and techniques for effective and accurate assessment.

Our previous volume examined talent management. This time the focus is on the assessment itself and truly understanding what works and for whom. We believe this book will be helpful not only to I-O practitioners working in the assessment arena but also to other professionals who are engaged in assessing or hiring activities in corporations. As with previous volumes, our aim is to provide practical solutions grounded in research and applied experience. We believe this volume does just that. The Appendix alone is a gold mine of information for anyone interested in assessment—not to mention the main content of the volume. In our opinion, John and Doug have made a major contribution to the field with their efforts. We sincerely appreciate their dedication to making this edition a reality. Thanks, guys!

Pound Ridge, New York
May 2010

Janine Waclawski
Allan H. Church

Preface

There has been a marked trend over the past few years for organizations of all sizes to streamline their workforces and focus on selecting and retaining only the "best and the brightest" employees. Couple this with the skills gap that will soon emerge due to the magnitude of baby boomer retirements, and it is no surprise that organizational priorities have been steadily shifting toward talent acquisition and retention. As organizational consultants, we are continually engaged in dialogue about how assessments can best be leveraged to achieve a company's talent management objectives. Specifically, human resource (HR) and line leaders want to know if assessments should be used and, if so, what specific instruments would be applicable, whether they can be administered online, whether they need to be proctored, what the costs are, whether there are specific legal constraints, whether they can be implemented in multiple languages in multiple countries, how an assessment program should be managed, how to know if the process is working, and what the expected return on investment is. And these are just a few of the questions that need to be answered to ensure that an assessment program meets stakeholder needs, achieves the organization's goals, and has a positive impact on its bottom line.

The field of assessment has advanced rapidly over the past decade due in part to advancements in computer technology. By leveraging technology, organizations can reach across the boundaries of language, literacy, and geography to reliably assess a vast catalogue of candidate skills and abilities. Organizations can now harness the capabilities of sophisticated, Web-based assessment tools to simulate actual work environments—effectively measuring candidates' ability to perform under real-life conditions. Technological advances have also fostered a number of assessment methodologies such as adaptive testing that have led to significant improvements in measurement precision and efficiency.

Despite these advances, there remain some fundamental questions and decisions that each organization must grapple with to ensure it is maximizing the potential of its assessment program and taking advantage of well-researched theories and state-of-the-art practice. This book presents sound, practical guidelines that are steeped in empirical research for implementing an assessment process that will effectively drive an organization's critical talent decisions.

The Audience

This book is designed for a broad readership, from HR professionals who are tasked with implementing an assessment program to assessment professionals and practitioners of industrial-organizational (I-O) psychology, who advise, build, validate, and implement assessments. In addition, this book is intended for the users of assessments, including hiring managers and organizational leaders, who are looking for direction on what to assess, what it will take, and how to realize the benefits. This book is also intended for assessment researchers as well as instructors and graduate students in disciplines such as I-O psychology, HR management and organizational behavior, consulting psychology, and organizational development.

Overview of the Book

This book is divided into four parts: it examines frameworks for organizational assessment; assessment for selection, promotion, and development; strategic assessment programs; and advances, trends and issues. The Appendix provides examples of the types of tests and assessments currently available for use in the workplace.

The foundational chapters contained in Part One are designed to provide readers with a thorough understanding of what should be assessed and why and how to ensure that assessment programs are of the highest quality and reflect the latest thinking and practice in the field. Part Two is devoted to the specific applications of workplace assessment and covers a variety of positions where high-volume or high-stakes decisions need to be made. The chapters in this part emphasize

examples of current best practices in assessment to help practitioners understand, apply, and evaluate the success of these practices in their own work contexts. The focus is on assessment systems in place today and that are needed in the future as business needs change. The chapters address the application of assessments to clerical, professional, technical, sales, supervisory and early leadership, and managerial and executive positions. In addition, a chapter addresses the special case of police and firefighter selection.

Part Three highlights some of the key strategic applications of assessment that organizations rely on to boost their competitive edge. The chapters focus on succession management, staffing for organizational change (downsizing, mergers, and reorganizations), assessing for potential, and global selection. The chapters in Part Four cover a wide range of advances, trends, and issues: technology-based delivery of assessment, the legal environment, alternative validation strategies, addressing flaws in assessment decisions, and the strategic use of evaluation to link assessment to bottom-line organizational priorities.

A brief description of each of the chapters follows.

Part One: Framework for Organizational Assessment

Kevin Murphy sets the stage in Chapter One by discussing broad dimensions of individual differences that are likely to be relevant for understanding performance effectiveness and development in the workplace and delineates two general strategies for determining what to assess in organizations. In Chapter Two, Fritz Drasgow, Christopher Nye, and Louis Tay outline the characteristics and features that differentiate outstanding assessment programs from mediocre systems and provide information that practitioners can use to move toward state-of-the-art measurement in their organizations. The next six chapters examine the most commonly assessed characteristics in the workplace: cognitive ability, personality, background and experience, knowledge and skill, physical performance, and competencies. These chapters highlight the challenges faced in accurately and fairly assessing these characteristics and detail advances in the field and the state of practice for their measurement.

Michael McDaniel and George Banks kick off these topics in Chapter Three with a review of the research and practice in the use of general cognitive ability tests in workplace assessment. They trace the history of intelligence testing from its roots to modern applications and detail the merits of cognitive ability assessment for selecting and developing top talent. In Chapter Four Robert Hogan and Robert Kaiser provide a compelling look at the use of personality assessment, why it is so misunderstood, and how it can be leveraged to predict significant outcomes. Leaetta Hough follows in Chapter Five on the assessment of background and experience; she addresses factors affecting this tool's validity and provides empirically based recommendations for improving its accuracy in predicting behavior. In Chapter Six Teresa Russell highlights the different types of knowledge and skill measures and offers some innovative ideas for measuring both declarative and procedural knowledge and skills.

Deborah Gebhardt and Todd Baker focus in Chapter Seven on assessments used for selecting candidates for strenuous jobs. There are many critical applications of these assessments in both the public and private sectors where failure to meet physical demands can have a significant impact on job performance and safety. Finally, Jeffery Schippmann rounds out Part One with a groundbreaking and forthright portrayal of the evolution of the role of competencies in assessment programs.

Part Two: Assessment for Selection, Promotion, and Development

Judith Komaki opens this part with a fictional but very realistic account of an HR manager who is asked to produce a valid test of managerial skills on a shoestring budget. The frustrations and complexities of finding an off-the-shelf test that maps onto the required skills are brought to light in this engaging and perceptive chronicle. Wanda Campbell follows in Chapter Nine by drawing on her experience leading nationwide testing consortia to detail the use of assessment procedures for selecting, promoting, and developing individuals across a variety of technical

roles. In Chapter Ten, Lia Reed, Rodney McCloy, and Deborah Whetzel describe the evolution of responsibilities in both clerical and professional jobs over the past twenty years and provide an insightful analysis of the resulting impact on assessment decisions associated with these typically high-volume hiring jobs. Steven Brown focuses in Chapter Eleven on practical techniques and unique challenges associated with sales assessment (for example, dispersed locations and unproctored testing). He provides particularly valuable recommendations for how to properly define success criteria for salespersons and ensure that the assessment tools are validated. In Chapter Twelve, Mark Schmit and Jill Strange show how assessments can be leveraged to stem the tide of supervisory derailment in organizations and demonstrate how these assessments are important to the bottom line.

Ann Howard and James Thomas delve into the arena of high-stakes decision making in Chapter Thirteen on executive and managerial assessment. They address the unique characteristics and challenges associated with working at this level and show how the effective design and implementation of managerial and executive assessment programs can provide significant benefits to organizations. Chapter Fourteen by Gerald Barrett, Dennis Doverspike and Candice Young describes the special case of public sector assessment, with a specific focus on police and firefighter selection. The authors detail the challenges and outline strategies for successfully navigating in this highly contentious and heavily litigated area.

Part Three: Strategic Assessment Programs

In Chapter Fifteen, Matthew Paese outlines six fundamental actions required for organizations to shift from a traditional replacement-focused succession management system to a more contemporary growth-focused system, which is required to close the ever-widening leadership capability gaps. Rob Silzer and Sandra Davis follow with an incisive chapter that leverages a new integrated model of potential for making long-term predictions of performance. They describe a variety of assessment strategies and tools in the context of this model for assessing

potential and fit and show how the proper measurement of these critical elements links to an organization's competitive edge and bottom line.

In Chapter Seventeen, John Scott and Kenneth Pearlman outline the strategies necessary to build a legally defensible staffing model under various reduction-in-force (RIF) conditions, including mergers and acquisitions, restructuring, and targeted or across-the-board RIFs. Ann Marie Ryan and Nancy Tippins close Part Three with an astute analysis of issues faced by practitioners who must refine the methods used for single-country assessment and confront issues of validation, cultural differences, country-specific legal issues, consistency of use, measurement equivalence, and the impact of culture on design and implementation.

Part Four: Advances, Trends, and Issues

Douglas Reynolds and Deborah Rupp begin Part Four with Chapter Nineteen on technology-based delivery of assessment. They examine the conditions that have enabled the growth of technology-facilitated assessment, as well as the applicable professional standards, guidelines, and legal considerations that are specific to technology-based assessments. They emphasize the context for system deployment, options for system design, and the major issues that arise as these systems are implemented in organizations. In Chapter Twenty, Lawrence Ashe and Kathleen Lundquist focus on the legal environment in which workplace assessments operate, highlighting not only relevant government regulations and case law but also the current priorities of agencies enforcing federal equal employment opportunity law. They explore the future of employment litigation and provide a comprehensive approach for building legal defensibility into the workplace assessment process.

Morton McPhail and Damian Stelly in Chapter Twenty-One provide a summary of the alternative approaches for validating workplace assessments and cover the development of new validity evidence where traditional techniques are not applicable. In Chapter Twenty-Two James Outtz explores how the common flaws in deciding what to assess can have a major impact on both

the organization and its employees and candidates, particularly in high-stakes testing situations. He provides a number of solutions for consideration, including the use of an evidence-based approach and broadening the range of assessments under consideration.

In Chapter Twenty-Three, Jane Davidson helps readers understand how to leverage the evaluation of workplace assessment as a strategic tool for driving business success and achieving competitive advantage. Finally, in Chapter Twenty-Four, Paul Sackett offers concluding thoughts and future directions for the assessment field.

Appendix

The Appendix offers practical suggestions for assessments across the full range of applications that are covered in this book. It is designed as a user-friendly resource to help readers make decisions about which assessments they should consider for their needs. The Appendix, organized into sections related to four types of assessments—construct targeted, position targeted, management and leadership targeted, and job analysis support—provides test and publisher names of popular or commonly used instruments, along with a brief description of each.

Orientation

The chapters in this book provide a range of perspectives on how best to apply the science of people assessment to the workplace. The gracious experts who have contributed to this book were asked to blend the best of the common base of scientific knowledge with the unique demands of the workplace applications with which they are most familiar.

A tour of these topics could be considered in a similar light to a tasting tour across a wide range of cuisine. Just as different chefs draw selectively from the available ingredients and techniques to meet the tastes and expectations of their local culture, the experts here focus on the use of human characteristics and proven measurement techniques to meet the demands of a wide range of workplace environments. By understanding the

ingredients (Part One), how they are combined in different contexts (Parts Two and Three), and new techniques and emergent issues (Part Four), readers should be better prepared to assemble their own unique recipe. We hope the tour is both informative and enjoyable.

■ ■ ■

Darien, Connecticut John C. Scott
Pittsburgh, Pennsylvania Douglas H. Reynolds
May 2010

Acknowledgments

We are pleased to have had such a renowned group of globally recognized authors agree to devote their knowledge, experience, and time to the creation of this handbook. Their contributions reflect cutting-edge theory and practice, and it is clear why they are at the pinnacle of their profession. We greatly appreciate all of the effort and commitment that went into bringing these chapters to life. Without the dedication of the chapter authors, this book would not have been possible.

In addition, we express our sincere gratitude to a second set of authoritative experts who provided in-depth chapter reviews and astute feedback. Their contributions significantly improved this handbook, and we are extremely thankful for their efforts. These reviewers were Seymour Adler, Herman Aguinis, Ronald Ash, Dave Bartram, Milton Blood, Joan Brannick, Eric Dunleavy, Charlotte Gerstner, Arthur Guttman, Monica Hemmingway, Cal Hoffman, Joyce Hogan, Lawrence James, Jerard Kehoe, Rich Klimoski, Deirdre Knapp, Elizabeth Kolmstetter, Manny London, Joyce Martin, Jennifer Martineau, James Outtz, Neal Schmitt, Jeffery Stanton, Garnett Stokes, George Thornton, and Mike Zickar.

We also thank Janine Waclawski and Allan Church, the editors of Jossey-Bass's Professional Practice Series, for their invaluable guidance throughout the preparation of this book. Special thanks go out as well to Matt Davis and Lindsay Morton and the editorial staff at Jossey-Bass for keeping us on track.

Of course, we are particularly grateful to our respective families: Kimberly, Justin, and Jeremy Scott, and Jennifer Cooney, Sam, and Caleb Reynolds. Through their love, patience, tolerance, and generosity, each contributed mightily to this project.

The Editors

John C. Scott is chief operating officer and cofounder of APT *Metrics*, Inc., a global human resource consulting firm that designs sophisticated talent management solutions for Fortune 100 companies and market innovators. He has more than twenty-five years of experience designing and implementing human resource systems across a variety of high-stakes global settings. For the past fifteen years, he has directed APT's talent management practice areas to serve a broad range of client sectors: retail, pharmaceutical, telecommunications, entertainment, insurance, technology, hospitality, aerospace, utilities, and financial services.

John is coeditor of *The Human Resources Program-Evaluation Handbook* and coauthor of *Evaluating Human Resources Programs: A Six-Phase Approach for Optimizing Performance*. He has also authored numerous chapters and articles in the areas of assessment, selection, and organizational surveys and serves on the editorial board of Wiley-Blackwell's Talent Management Essentials series.

John was the 2009 conference program chair for the Society for Industrial and Organizational Psychology (SIOP) and was recently appointed as SIOP's representative to the United Nations. He received his Ph.D. in industrial-organizational psychology from the Illinois Institute of Technology.

■ ■ ■

Douglas H. Reynolds is vice president of assessment technology at Development Dimensions International, where he leads the development and deployment of assessment and testing products. His work has been implemented in many Fortune 500 companies and several federal agencies. In the 1990s, he designed some of the first Internet-based assessments used for large-scale corporate recruiting. More recently, his work has focused on the

use of computer-delivered simulations for executive and leadership evaluation. He is also an expert witness on personnel selection practices, and his articles, book chapters, and presentations often focus on the intersection of technology and assessment.

Recently Doug coauthored *Online Recruiting and Selection*, a book on the integration of technology with personnel selection practices. He also serves on the editorial boards of the *Journal of Management* and Wiley-Blackwell's Talent Management Essentials series.

Doug is active in SIOP leadership, currently serving on the executive board as the communications officer and in the past as chair of the Visibility and Professional Practice committees. In prior roles, he was a senior scientist at the Human Resources Research Organization and adjunct faculty at George Washington University. He earned his Ph.D. in industrial-organizational psychology from Colorado State University.

The Contributors

R. Lawrence Ashe Jr., the chair of Ashe, Rafuse & Hill in Atlanta, Georgia, is in his forty-third year of advising on and litigating employment law, test validity, and civil rights issues. He is nationally recognized for his class action and test validation expertise and experience. He has tried more employment selection class actions to judgment than any other management lawyer in the country, including some of the largest cases tried to date. His civil rights practice is 10 to 15 percent representation of plaintiffs with the balance defendants. Test validity and other employment selection issues are over one-third of his practice. Lawrence is a founding board member of Atlanta's Center for Civil and Human Rights and a board and executive committee member of the National Council for Research on Women. He is a Fellow in the American College of Trial Lawyers and the College of Labor and Employment Lawyers. He graduated from Princeton University and Harvard Law School.

■ ■ ■

Todd A. Baker is a senior research scientist at Human Performance Systems, Inc. and has twenty years of experience developing and validating physical performance and cognitive assessments for public, private, and military organizations. He has conducted job analyses for hundreds of physically demanding jobs and developed and validated numerous physical performance and cognitive test batteries for evaluation of applicant and incumbent personnel for public safety and private sector positions. In 2006, Todd was part of a team that was awarded the Society for Industrial and Organizational Psychology M. Scott Myers Award for Applied Research in the Workplace for developing and validating the assessments and medical guidelines used for

selecting transportation security officers. In 2003, he was part of a team that was awarded the International Public Management Association–Assessment Council Innovations Award. Todd has litigation experience, providing testimony in the areas of job analysis, physical performance tests, promotional tests, and the Fair Labor Standards Act.

■ ■ ■

George C. Banks holds an M.A. in industrial-organizational psychology from the University of New Haven and is pursuing a Ph.D. at Virginia Commonwealth University. His research focuses on employment testing, applicant attraction, and team development. George is a member of the Academy of Management and the Society for Industrial and Organizational Psychology.

■ ■ ■

Gerald V. Barrett, president of Barrett and Associates since 1973, has been involved in the development and validation of employment tests for numerous jobs, including firefighter and police officer. He has consulted with numerous public and private organizations and has been engaged as an expert witness in over 160 court cases, usually dealing with issues of alleged age, race, national origin, or sex discrimination in selection, promotion, termination, reduction in force, or compensation. Gerald received his Ph.D. in industrial psychology from Case Western Reserve University and his J.D. from the University of Akron's School of Law. He is both a licensed psychologist and a licensed attorney in the State of Ohio. He is a Fellow of the American Psychological Association and the American Psychological Society. The Society for Industrial and Organizational Psychology presented him with the Distinguished Professional Contributions Award in 1992 in recognition of his outstanding contributions to the practice of industrial-organizational psychology. He also received the Life Time Achievement Award from the Industrial Organizational Behavior Group.

■ ■ ■

Steven H. Brown is president of SHB Selection Consulting and senior consultant, assessment solutions, for LIMRA International. He consults internationally in the areas of selection and assessment. Previously he was vice president and director of LIMRA International's Assessment Solution Group, a professional staff engaged in human resource research, selection product development, and recruiting and selection process consultation. Steve holds a B.A. from DePauw University and a Ph.D. in industrial-organizational psychology from the University of Minnesota. He is a Fellow of the American Psychological Association, the Society for Industrial and Organizational Psychology (SIOP), and the American Psychological Society. He has published numerous articles in professional journals about personnel and sales selection. He has served on the editorial board of *Personnel Psychology* and was a member of the task force that wrote SIOP's *Principles for the Validation and Use of Personnel Selection Procedures.*

■ ■ ■

Wanda J. Campbell is the senior director of employment testing for Edison Electric Institute, the trade association of investor-owned electric utility companies. She manages a nationwide testing program that includes nine employee selection test batteries, seven of which are for technical jobs, as well as a career assessment and diagnostic instrument. EEI tests have become the industry standard for the electric utility industry. She is a member of the Society for Industrial and Organizational Psychology (SIOP), the Society of Consulting Psychology, the American Psychological Association, and the Maryland Psychological Association. She served on the committee responsible for the 2003 revision of the SIOP *Principles for the Validation and Use of Personnel Selection Procedures* and is currently serving on the SIOP Workshop Committee. She has made over thirty presentations at professional conferences and coauthored four book chapters. She is licensed as a psychologist in the State of Maryland and is currently serving as the treasurer for the Maryland Psychological Association. She earned her Ph.D. in industrial-organizational psychology from Old Dominion University. Prior to becoming a

psychologist, Wanda worked for five years as an equal opportunity specialist for the Equal Employment Opportunity Commission.

■ ■ ■

E. Jane Davidson runs an evaluation consulting business (Real Evaluation Ltd.), working across a range of domains including leadership development, human resources, health, education, and social policy. Her work includes evaluation training and development, facilitated self-evaluation and capacity building, independent evaluation, and formative and summative meta-evaluation (advice, support, and critical reviews of evaluations). Previously she served as associate director of the Evaluation Center at Western Michigan University. There, she launched and directed the world's first fully interdisciplinary Ph.D. in evaluation. She has presented numerous keynote addresses and professional development workshops internationally. Jane is the author of *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation* (2004). She was the 2005 recipient of the American Evaluation Association's Marcia Guttentag Award, awarded to a promising new evaluator within five years of completing the doctorate. She received her Ph.D. from Claremont Graduate University in organizational behavior, with substantial emphasis on evaluation.

■ ■ ■

Sandra L. Davis is chief executive officer of MDA Leadership Consulting, a talent management and leadership development firm. She cofounded the company in 1981 and currently focuses her consulting work on senior executive talent evaluation. She is widely known as an executive coach and thought leader in the industry, counting numerous Fortune 500 companies among her clients. She has contributed chapters and articles to professional books and journals related to assessment, leadership development, coaching, and succession. She served on the Strong Interest Inventory Advisory panel for Consulting Psychologists Press; her book *Reinventing Yourself* was based on her work in the practical use of tests. She is a member of the American Psychological Association and the Society for Industrial and Organizational Psychology. Prior to founding MDA, she served

on the faculty of the University of Minnesota and worked for Personnel Decisions. Sandra earned her B.S. from Iowa State University and her Ph.D. in counseling psychology with an emphasis in industrial-organizational psychology from the University of Minnesota.

■ ■ ■

Dennis Doverspike is a professor of psychology at the University of Akron, senior fellow of the Institute for Life-Span Development and Gerontology, and director of the Center for Organizational Research. He holds a Diplomate in industrial-organizational psychology and in organizational and business consulting from the American Board of Professional Psychology (ABPP) and is a licensed psychologist in the State of Ohio. Dennis has over thirty years of experience working with consulting firms and with public and private sector organizations, including as executive vice president of Barrett & Associates. He is the author of two books and over one hundred refereed journal publications. Current major additional positions include president of the ABPP specialty board in organizational and business consulting. He received his Ph.D. in psychology in 1983 from the University of Akron. His M.S. in psychology is from the University of Wisconsin–Oshkosh and his B.S. is from John Carroll University. His areas of specialization include job analysis, testing, and compensation.

■ ■ ■

Fritz Drasgow is a professor of psychology and labor and industrial relations at the University of Illinois at Urbana-Champaign. Previously he was an assistant professor at Yale University's School of Organization and Management. He has also provided consultation on testing and measurement issues to a variety of organizations in the private and nonprofit sectors. Drasgow's research focuses on psychological measurement, computerized testing, and the antecedents and outcomes of sexual harassment. He is a former chairperson of the American Psychological Association's Committee on Psychological Tests and Assessments, the U.S. Department of Defense's Advisory Committee on Military

Personnel Testing, the Department of Defense and Department of Labor's Armed Services Vocational Aptitude Battery Norming Advisory Group, the American Psychological Association's Taskforce on Internet Testing, and the American Institute of Certified Public Accountants' Psychometric Oversight Committee. Drasgow is a member of the editorial review board of eight journals, including *Applied Psychological Measurement, Journal of Applied Psychology*, and the *International Journal of Selection and Assessment*. He is a former president of the Society for Industrial and Organizational Psychology and received its Distinguished Scientific Contributions Award. He received his Ph.D. from the University of Illinois at Urbana-Champaign.

■ ■ ■

Deborah L. Gebhardt is president of Human Performance Systems, Inc., and has over twenty-five years of experience developing and validating physical performance tests, fitness programs, and medical guidelines and standards for public, private, and military organizations. She holds Fellow status in the Society for Industrial and Organizational Psychology (SIOP) and the American College of Sports Medicine (ACSM). She has published research in the areas of job analysis, physical test development and standards, medical guidelines, injury analysis, and biomechanics. She has conducted over one hundred physical performance test development and validation projects in the federal, public, private, and military sectors. In 2006, Gebhardt was part of a team that was awarded the SIOP M. Scott Myers Award for Applied Research in the Workplace for developing and validating the assessments and medical guidelines used for selecting transportation security officers. In 2003, she was part of a team that was awarded the International Public Management Association–Assessment Council Innovations Award. She has served as an expert witness in class action (Title VII) and Americans with Disabilities Act litigation, and arbitrations regarding the physical performance tests, job analysis, validation, and medical guidelines used in the selection and retention of workers.

■ ■ ■

Robert Hogan, president of Hogan Assessment Systems, is an international authority on personality assessment, leadership, and organizational effectiveness. He was McFarlin Professor and chair of the Department of Psychology at the University of Tulsa for fourteen years. Prior to that, he was professor of psychology and social relations at The Johns Hopkins University. He has received a number of research and teaching awards, is the author of *Personality and the Fate of Organizations* and the Hogan Personality Inventory, and is the editor of the *Handbook of Personality Psychology* (1997). Robert received his Ph.D. from the University of California, Berkeley, specializing in personality assessment. He is the author of more than three hundred journal articles, chapters, and books. He is widely credited with demonstrating how careful attention to personality factors can influence organizational effectiveness in a variety of areas—ranging from organizational climate and leadership to selection and effective team performance. Robert is a Fellow of the American Psychological Association and the Society for Industrial and Organizational Psychology.

■ ■ ■

Leaetta M. Hough is founder and president of the Dunnette Group, an adjunct professor in the psychology department at the University of Minnesota, past president of the Society for Industrial and Organizational Psychology (SIOP), and past president of the Federation of Associations in Behavioral and Brain Sciences, a coalition of twenty-two scientific societies. She is coeditor of the four-volume *Handbook of I-O Psychology*, lead author of chapters in the *Annual Review of Psychology*, the *International Handbook of Work and Organizational Psychology*, the I-O volume of the *Comprehensive Handbook of Psychology*, the *Handbook of Personnel Selection*, and the *Biodata Handbook*, as well as dozens of articles in refereed journals. She has developed new methods of work analysis, performance appraisal systems, and selection methods, including hundreds of valid and defensible personnel selection and performance measures, many of which are innovative, nontraditional measures that have minimal, if any, adverse impact against protected groups. She is an

internationally recognized expert in the measurement of personality, creativity, and global mind-set.

■ ■ ■

Ann Howard is chief scientist for Development Dimensions International (DDI), a global talent management company. She leads the Center for Applied Behavioral Research (CABER), DDI's hub for research to support evidence-based talent management. She directs research that measures the effectiveness and organizational impact of DDI interventions and investigates global workplace practices and issues. Previously, as DDI's manager of assessment technology integrity, she designed, implemented, and evaluated assessment center platforms and set quality standards for DDI's assessment technologies. During twelve years at AT&T she codirected two longitudinal studies of the lives and careers of managers. She is the senior author (with Douglas W. Bray) of *Managerial Lives in Transition: Advancing Age and Changing Times*, which received the George R. Terry Award of Excellence from the Academy of Management. She has written more than one hundred book chapters, monographs, and papers on topics such as assessment centers, executive selection, managerial careers, and leadership development. She has also edited several books on the changing workplace, including *The Changing Nature of Work* and *Diagnosis for Organizational Change: Methods and Models*. She is a Fellow and past president of the Society for Industrial and Organizational Psychology.

■ ■ ■

Robert B. Kaiser is a partner with Kaplan DeVries, an executive development firm, and was previously at the Center for Creative Leadership. He has written over one hundred publications and presentations, including three books. His work on leadership, development, and assessment has appeared in *Harvard Business Review* and *Sloan Management Review*, as well as top-tier scholarly journals. He is the coauthor, along with Bob Kaplan, of the Leadership Versatility Index, a 360-degree feedback tool that received three U.S. patents. Rob also has a consulting practice in

which he grooms high potentials for the executive suite, and in his talent management work for global corporations, he provides research-based services that include developing custom leadership models and assessment tools as well as statistical analysis of performance data to inform talent management strategy. He has an M.S. in industrial-organizational psychology from Illinois State University.

■ ■ ■

Michael R. Kemp works in Development Dimensions International's Assessment and Selection Analytics Group, where he designs, develops, and ensures the ongoing effectiveness of DDI's screening, testing, and assessment solutions worldwide. His major areas of focus are test and assessment content design, validation and documentation, development of local and global norms, and the analysis of operational data. He is also a Ph.D. candidate in industrial-organizational psychology at Central Michigan University. For his doctoral work, his research focuses on leadership development and multisource feedback. Other research interests are applicant assessment and selection, employee engagement, occupational stress, and leadership derailment. He is an affiliate member of the Society for Industrial and Organizational Psychology and has presented research on leadership theory at past conferences.

■ ■ ■

Judith L. Komaki, a former professor of industrial-organizational psychology at Purdue University and City University of New York's Baruch College, initially set up motivational programs. But she quickly learned that without proper management support, the program, no matter how well designed, would be doomed to failure. Hence, she shifted to leadership, identifying what effective leaders did aboard racing sailboats and in darkened theaters. While watching stage directors in connection with an Army Research Institute contract, she noticed what was onstage and what was missing. Rarely did she find characters resembling herself—a professional woman of color—so she began writing plays.

One play forced her to come to terms with the insidious effects of race, something she had assiduously avoided. But realizing her arsenal of professional skills, she began using them to pursue social and economic justice, taking to heart the management adage "We treasure what we measure." She is the author of a leadership book (*Leadership from an Operant Perspective,* Routledge, 1998), an off-off Broadway play, and an article about pursuing the dreams of Martin Luther King Jr. ("Daring to Dream: Promoting Social and Economic Justice at Work," 2007).

■ ■ ■

Kathleen K. Lundquist is CEO and cofounder of APT Metrics, a global human resource consulting firm that designs talent management solutions for Fortune 100 clients. An organizational psychologist, she testifies frequently as an expert witness in employment discrimination class action lawsuits on behalf of both defendants and plaintiffs and has provided invited testimony before the U.S. Equal Employment Opportunity Commission. Following settlements of high-profile class actions, the courts have appointed her to design and implement revised HR processes for organizations such as the Coca-Cola Company, Morgan Stanley, Abercrombie & Fitch, Ford Motor Company, and the Federal Bureau of Investigation. In consulting with clients ranging from multinational corporations to government and nonprofit employers, she designs proactive measures to improve the fairness, validity, and legal defensibility of HR processes before they are challenged. She is a former research associate with the National Academy of Sciences, a fellow in psychometrics with the Psychological Corporation, and a summer research fellow with the Educational Testing Service. Kathleen is a member of the corporate advisory board of the National Council for Research on Women.

■ ■ ■

Rodney A. McCloy is a principal staff scientist at the Human Resources Research Organization (HumRRO), serving as an

in-house technical expert and a mentor to junior staff. He is well versed in several multivariate analytical techniques and has applied them to numerous research questions, particularly those involving personnel selection and classification, job performance measurement and modeling, and attrition and turnover. His assessment and testing experience has spanned both cognitive and noncognitive domains and has involved several large-scale assessment programs, including the Armed Services Vocational Aptitude Battery, National Assessment of Educational Progress, and General Aptitude Test Battery. He directs HumRRO's internal research and development program and is active in the academic community, having served as an adjunct faculty member of the psychology departments at George Mason University and the George Washington University. He currently serves on the advisory board for the Masters of I-O Psychology Program at Northern Kentucky University and is a Fellow of the American Psychological Association and the Society for Industrial and Organizational Psychology. He received his B.S. in psychology from Duke University and his Ph.D. in industrial-organizational psychology from the University of Minnesota.

■ ■ ■

Michael A. McDaniel is a professor of management and research professor of psychology at Virginia Commonwealth University. He is internationally recognized for his research and practice in personnel selection system development and validation. He is also known for his applications of meta-analysis in employment testing, management, and other fields. McDaniel has published in several major journals, including the *Journal of Applied Psychology*, *Personnel Psychology, International Journal of Selection Assessment*, and *Intelligence*. He is a member of the Academy of Management and a Fellow of the Society for Industrial and Organizational Psychology, the American Psychological Association, and the Association for Psychological Science. He received his Ph.D. in industrial-organizational psychology from George Washington University.

■ ■ ■

S. Morton McPhail, a senior vice president and managing principal with Valtera Corporation, received his master's and doctoral degrees in industrial-organizational psychology from Colorado State University. He has served as a consultant for over thirty years to a wide variety of public and private sector clients on issues including employee selection and promotion, test validation, training and development, performance assessment, and termination. He has served as an expert in litigation involving such diverse issues as job analysis, test development and validation, violence in the workplace, equal employment opportunities, compensation, and reductions in force. He has published in professional journals and presented on numerous topics at professional meetings. Mort serves as adjunct faculty for both the University of Houston and Rice University and is on the editorial board of *Personnel Psychology*. A Fellow of the Society for Industrial and Organizational Psychology, Mort is currently its financial officer/secretary. He is a licensed psychologist and serves on a committee of the Texas Psychology Board regarding development and validation of the state's jurisprudence and ethics examination for licensure.

■ ■ ■

Kevin R. Murphy is a professor of psychology and information sciences and technology at Pennsylvania State University. He is a Fellow of the American Psychological Association, American Psychological Society, and the Society for Industrial and Organizational Psychology (SIOP). He has served as president of SIOP (1997–1998) and as associate editor and then editor of the *Journal of Applied Psychology* (1991–2002), as well as a member of the editorial boards of *Human Performance, Personnel Psychology, Human Resource Management Review, International Journal of Management Reviews, Journal of Industrial Psychology*, and *International Journal of Selection and Assessment*. He is the recipient of the SIOP's 2004 Distinguished Scientific Contribution Award. He is the author of over 150 articles and book chapters, and author or editor of eleven books, in areas ranging from psychometrics and statistical analysis to individual differences, performance assessment, gender, and honesty in the workplace.

Kevin's main areas of research are personnel selection and placement, performance appraisal, and psychological measurement. His current work focuses on understanding the validation process.

■ ■ ■

Christopher D. Nye is a Ph.D. candidate at the University of Illinois at Urbana-Champaign. His research primarily involves personnel selection and assessment, organizational research methods, and workplace deviance. His master's thesis examined the prevalence of cheating in multistage testing programs and was later published in the *International Journal of Selection and Assessment*. His dissertation is focused on developing new methods for interpreting the results of studies examining bias in psychological measures. He has also conducted psychometric research for several large organizations, including the Department of Defense, the College Board, and the National Council of State Boards of Nursing.

■ ■ ■

James L. Outtz is president of Outtz and Associates, a consulting firm in Washington, D.C. He develops selection procedures for a wide variety of organizations, from Alcoa to the Federal Deposit Insurance Corporation. His interests include selection, training, performance management, job analysis and work design, workforce diversity, and equal employment opportunity. His professional service includes membership on the Ad Hoc Committee on Revision of the Principles for the Validation and Use of Personnel Selection Procedures. In addition, he has served as consulting editor to the *Journal of Applied Psychology*. He is nationally recognized for his work in the area of adverse impact and alternative selection procedures, subjects about which he has written extensively. He is the editor of *Adverse Impact: Implications for Organizational Staffing and High Stakes Selection* (2010). He received his Ph.D. in industrial-organizational psychology from the University of Maryland. He is a Fellow in the Society for Industrial and Organizational

Psychology, the American Psychological Association, and the American Educational Research Association.

■ ■ ■

Matthew J. Paese is vice president of executive succession management at Development Dimensions International (DDI) and holds his Ph.D. in industrial-organizational psychology from the University of Missouri–St. Louis. Coauthor of *Grow Your Own Leaders* (2002) and numerous articles in the areas of talent management, executive assessment, succession, and development, he has spent more than fifteen years consulting with CEOs and senior teams from leading organizations around the world. His work includes the design and implementation of strategic talent initiatives, including organizational talent strategy, succession management, CEO succession, executive assessment, executive coaching, and executive team building. Paese has consulted extensively with executives from many organizations including Wal-Mart, Microsoft, and BP. Prior to joining DDI in 1994, he worked for Anheuser-Busch, where he was responsible for executive assessment and selection initiatives.

■ ■ ■

Kenneth Pearlman is in independent consulting practice following a twenty-seven-year career employed in both the public and private sectors, including the U.S. Office of Personnel Management, AT&T, and Lucent Technologies. He has specialized in research and applications in the areas of personnel selection and assessment, work and skill analysis, organizational and employee survey development, leadership assessment and development, and productivity measurement and enhancement. He has authored or coauthored over one hundred journal articles, technical reports, book chapters, papers, and presentations in these areas. He is a Fellow of the American Psychological Association, the Association for Psychological Science, and the Society for Industrial and Organizational Psychology (SIOP). He is on the editorial boards of *Personnel Psychology*, *Industrial and Organizational Psychology*, and the *International Journal of Selection*

and Assessment, and served for eight years on the editorial board of SIOP's Professional Practice book series. He is coholder of a U.S. patent on a work analysis software tool. He has served as a member of the National Research Council's Board on Testing and Assessment. He received his B.A. in psychology from the Catholic University of America and his Ph.D. in industrial-organizational psychology from the George Washington University.

■ ■ ■

Lia M. Reed is an industrial-organizational psychologist in the Selection, Evaluation and Recognition Department of the U.S. Postal Service. She helps manage over twenty preemployment tests for the Postal Service, covering over two hundred job titles across four unions. Previously she worked for consulting firms where she assisted clients in developing, validating, and administering a variety of assessments. For the past ten years, Lia's work has focused primarily on the development and validation of preemployment and promotional tests for public, private, and government organizations, and she has worked on a wide variety of assessments. She received her M.A. and Ph.D. from DePaul University in Chicago. She is a member of the Society for Industrial and Organizational Psychology, the American Psychological Association, and the Society for Human Resource Management and has served as a board member of the Personnel Testing Council of Metropolitan Washington.

■ ■ ■

Deborah E. Rupp is an associate professor of labor/employment relations, psychology, and law at the University of Illinois at Urbana-Champaign. Her research related to assessment focuses on assessment center validity, the use of the method to foster professional development, and the use of technology to enhance assessment and development. She has coauthored the new edition of *Assessment Centers in Human Resource Management* with George C. Thornton and was the first recipient of the Douglas W. Bray and Ann Howard Award (for research on leadership assessment and development). She also conducts research on organizational

justice, corporate social responsibility, and emotions at work, and has published over fifty scholarly papers. She is currently an associate editor at *Journal of Management*, recently cochaired the International Congress on Assessment Center Methods, and co-led the International Taskforce on Assessment Center Guidelines in publishing a revision to the *Guidelines and Ethical Considerations for Assessment Center Operations*. Her assessment center research was also cited by U.S. Supreme Court Justice J. Ginsburg in proceedings surrounding the decision of the employment discrimination case *Ricci* v. *DeStefano et al.*

■ ■ ■

Teresa L. Russell is a principal staff scientist at the Human Resources Research Organization and has more than twenty-five years of experience in personnel selection and classification. Early in her career, she gained a broad base of experience as a part of the U.S. Army's Project A research team, developing spatial and perceptual tests, content-analyzing critical incidents, developing performance rating scales, and collecting and analyzing data. Since that time, she has been involved in the development and validation of predictor measures for a wide variety of military and civilian organizations. She is currently the project director for a series of projects to develop and validate a measure of information and communication technology literacy for inclusion on the Armed Services Vocational Aptitude Battery. She has authored book chapters on cognitive ability measurement, measurement plans and specifications, experimental test battery psychometrics, and career planning, as well as dozens of technical reports and conference papers.

■ ■ ■

Ann Marie Ryan is a professor of organizational psychology at Michigan State University. Her major research interests are improving the quality and fairness of employee selection methods and topics related to diversity and justice in the workplace. In addition to publishing extensively in these areas, she regularly consults with organizations on improving assessment processes. She is

a past president of the Society for Industrial and Organizational Psychology and past editor of the journal *Personnel Psychology*. She received her B.S. with a double major in psychology and management from Xavier University and her M.A. and Ph.D. in psychology from the University of Illinois at Chicago.

■ ■ ■

Paul R. Sackett is the Beverly and Richard Fink Distinguished Professor of Psychology and Liberal Arts at the University of Minnesota. He received his Ph.D. in industrial-organizational psychology at the Ohio State University in 1979. His research interests revolve around various aspects of testing and assessment in workplace, educational, and military settings. He has served as editor of the journals *Industrial and Organizational Psychology Perspectives on Science and Practice* and *Personnel Psychology*. He has also served as president of the Society for Industrial and Organizational Psychology, cochair of the committee producing the Standards for Educational and Psychological Testing, a member of the National Research Council's Board on Testing and Assessment, and chair of the American Psychological Association's Committee on Psychological Tests and Assessments and its Board of Scientific Affairs.

■ ■ ■

Jeffery S. Schippmann is the senior vice president of human resources and chief talent officer for Balfour Beatty Construction in Dallas, Texas. Previously he was vice president of global talent management for the Hess Corporation, where he was responsible for succession planning, performance management, talent assessment, and management development and training activities. Previously, he was the director of Organization and Management Development for PepsiCo. In this role he was responsible for a broad range of talent management programs and initiatives over a six-year period, including significant work to refocus managers on people development activities and restructuring the Pepsi "employment deal." Before Pepsi, Jeff was in consulting with Personnel Decisions International in a variety of roles focusing

on selection and staffing solutions, executive assessment and development, assessment centers, and competency modeling for a broad range of clients including Texas Instruments, Bank One, Memorial Sloan-Kettering, American Express, Boeing, and Ford. He received his Ph.D. in industrial-organizational psychology at the University of Memphis in 1987 and is the author of two books and numerous book chapters and articles, including work appearing in the *Journal of Applied Psychology* and *Personnel Psychology*.

■ ■ ■

Mark J. Schmit is the western regional vice president for APT, with an office in Denver. He has more than twenty years of experience in the field of human resources (HR). He has spent time as an HR generalist, academic, applied researcher, and internal and external consultant to both public and private organizations. He has developed recruitment, selection, promotion, performance management, organizational effectiveness, and development tools and systems for numerous organizations. Most recently he has been involved in employment discrimination litigation, serving as an expert witness and consultant from the field of industrial-organizational psychology. Schmit earned a Ph.D. in industrial-organizational psychology from Bowling Green State University in 1994. He has published more than twenty-five professional journal articles and book chapters and delivered more than forty-five presentations at professional meetings on HR and industrial-organizational psychology topics. He is an active member of the American Psychological Association, Society for Industrial and Organizational Psychology, and Society for Human Resource Management.

■ ■ ■

Rob Silzer is managing director of HR Assessment and Development, a corporate consulting firm, and is on the I-O psychology doctoral faculty at Baruch College and Graduate Center, City University of New York. He has consulted with

leaders in over 150 organizations, focusing on leadership assessment, selection, succession, development, coaching, and talent management. Rob is also a fellow of the American Psychological Association, the Association for Psychological Science, the Society for Industrial and Organizational Psychology, and the Society of Consulting Psychology. After receiving his Ph.D. in industrial-organizational psychology and counseling psychology from the University of Minnesota, he served as director of personnel research at Fieldcrest-Cannon and president of PDI–New York. He has taught doctoral courses at the University of Minnesota, New York University, and Baruch–CUNY. Rob has served on the editorial boards of *Personnel Psychology, Industrial and Organizational Psychology*, and *The Industrial-Organizational Psychologist* and as president of the Metropolitan New York Association of Applied Psychology. Rob has edited several books, including *Strategy-Driven Talent Management* (with Ben Dowell), *The 21st Century Executive*, and *Individual Psychological Assessment* (with Dick Jeanneret). He has authored over one hundred articles, book chapters, professional workshops, and presentations. He enjoys global adventure travel, mountain trekking, alpine skiing, and scuba diving.

■ ■ ■

Damian J. Stelly, organization development director at JCPenney, leads the company's performance management and employee selection programs and consults with leaders regarding organizational development and research initiatives. As an internal and external consultant, he has also held positions at Anheuser-Busch and Valtera Corporation. Damian has managed a broad range of projects, including the development of selection and placement systems, survey programs, performance management programs, organizational development initiatives, and employee development programs. He has presented or published on a variety of topics, including validation methods, test development, and leadership behavior. He is a member of the American Psychological Association and the Society for Industrial and Organizational Psychology. Damian holds a B.A. in psychology from Louisiana State University and received his M.A. and Ph.D.

in industrial-organizational psychology from the University of Missouri–St. Louis. He is licensed as a psychologist in Texas.

■ ■ ■

Jill M. Strange, a project manager at APT with ten years of experience in the field of industrial-organizational psychology, specializes in the design, development, and implementation of competency models and competency-based tools and provides consulting services in the areas of job analysis, employee selection development and validation, and litigation support. She has led several large-scale efforts in the areas of competency modeling, job analysis, and selection assessment design for industry, government, and military clients. In addition, she worked with several clients to design and validate selection assessments as well as performance assessment systems and workforce planning strategies. Strange is the author of over fifteen peer-reviewed articles and book chapters on the subjects of leadership, assessment, and competency modeling and frequently presents at professional meetings and conferences on these areas. She earned her Ph.D. in industrial-organizational psychology from the University of Oklahoma and is Senior Professional in Human Resources (SPHR) certified. She is a member of the American Psychological Association and the Society for Industrial and Organizational Psychology.

■ ■ ■

Louis Tay is a doctoral student in industrial-organizational psychology at the University of Illinois at Urbana-Champaign. His research interests include subjective well-being, emotions, and culture. He is actively working on conceptual and methodological advances in psychological measurement, encompassing the synthesis of item-response theory, latent class modeling, and hierarchical linear modeling. He has conducted research with several large organizations, including the American Dental Association, the College Board, and the Gallup Organization. He has published in several journals, including *International Journal of Testing*, *Journal of Applied Psychology*, and *Organizational Research Methods*.

■ ■ ■

James N. Thomas, vice president of consulting services at Development Dimensions International, heads its Northeastern Regional Consulting Group headquartered in New York City. Previously he led its Southeastern Regional Consulting team located in Atlanta. Thomas has developed and deployed talent management solutions for many Fortune 500 companies and public sector entities in the Americas, Europe, Asia, and Australia. He has recognized expertise in recruitment, selection, assessment, and executive development and has published on topics ranging from job analysis, and behavioral interviewing to psychological assessment. He has also presented at numerous national and international human resource conferences.

■ ■ ■

Nancy T. Tippins is a senior vice president and managing principal of Valtera Corporation, where she is responsible for the development and execution of firm strategies related to employee selection and assessment. She has extensive experience in the development and validation of tests and other forms of assessment that are designed for purposes of selection, promotion, development, and certification and used for all levels of management and for hourly employees. She has designed and implemented global test and assessment programs, as well as designed performance management programs and leadership development programs. Prior to joining Valtera, she worked as an internal consultant in large Fortune 100 companies managing the development, validation, and implementation of selection and assessment tools. She is active in professional affairs and is a past president of the Society for Industrial and Organizational Psychology (SIOP). She is a Fellow of SIOP, the American Psychological Association, and the Association for Psychological Science. Nancy received M.S. and Ph.D. degrees in industrial-organizational psychology from the Georgia Institute of Technology.

■ ■ ■

Deborah L. Whetzel is a program manager of the Personnel Selection and Development program at the Human Resources Research Organization. She has over twenty years of experience

in personnel selection research and development in both the public and private sectors. Her areas of expertise include job analysis, leadership competency models, performance appraisal systems, development of structured interviews and situational judgment tests, and developing and validating assessment processes. She has conducted several meta-analyses; most recently she analyzed the validity of various measures (cognitive ability and personality) for predicting performance in clerical occupations. She has coedited two books, *Applied Measurement: Industrial Psychology in Human Resources Management* and *Applied Measurement Methods in Industrial Psychology*. Deborah has served as an adjunct professor in the graduate program at George Mason University and in the undergraduate program at Virginia Commonwealth University. She earned her Ph.D. at George Washington University specializing in industrial-organizational psychology.

■ ■ ■

Candice M. Young holds an M.A. in industrial-organizational psychology (I-O) and is a senior associate in human resource consulting with Barrett and Associates. Young has extensively studied the content areas of personnel selection, training, and performance appraisal, as well as organizational behavior and motivation. Furthermore, she has been trained in psychometrics, advanced research methods, and statistics. She has provided litigation support for race, sex, and age discrimination lawsuits and has performed test development activities for selection and promotional purposes in the public and private sector. Candice is a doctoral candidate at the University of Akron. She received her M.A. in industrial-organizational psychology from Xavier University and her B.A. in psychology from Spelman College.

Handbook of Workplace Assessment

Part One

Framework for Organizational Assessment

CHAPTER 1

INDIVIDUAL DIFFERENCES THAT INFLUENCE PERFORMANCE AND EFFECTIVENESS

What Should We Assess?

Kevin R. Murphy

Assessment in organizations can be carried out for a variety of purposes, many with high stakes for both individuals and organizations. The stakes can be particularly high when assessments are used to make decisions about personnel selection and placement or about advancement and development of individuals once they have been hired. If assessments focus on traits, attributes, or outcomes that are not relevant to success and effectiveness, both organizations and individuals may end up making poor decisions about the fit between people and jobs. If assessments are appropriately focused but poorly executed (perhaps the right attributes are measured, but they are measured with very low levels of reliability and precision), these assessments may lead to poor decisions on the parts of both organizations and individuals.

In this chapter, I focus on broad questions about the content of assessments (for example, What sorts of human attributes should assessments attempt to measure?) and say very little about the execution of assessments (the choice of specific tests

or assessment methods, for example) or even the use of assessment data. My discussion is general rather than specific, focusing on general dimensions of assessment (whether to assess cognitive abilities or broad versus narrow abilities, for example) rather than on the specifics of assessment for a particular job (say, the best set of assessments for selecting among applicants for a job as a firefighter).

This chapter provides a general foundation for many of the chapters that follow. It sets the stage by discussing broad dimensions of individual differences that are likely to be relevant for understanding performance, effectiveness, and development in the workplace. The remaining chapters in Part One start addressing more specific questions that arise when attempting to assess these dimensions. Chapter Two reviews the range of methods that can be used to assess the quality of measures, and Chapters Three through Eight provide a more detailed examination of specific domains: cognitive abilities, personality, background and experience, knowledge and skill, physical and psychomotor skills and abilities, and competencies.

Part Two of this book discusses assessment for selection, promotion, and development, and Parts Three and Four deal with strategic assessment programs and with emerging trends and issues.

I begin this chapter by noting two general strategies for determining what to assess in organizations: one that focuses on the work and the other that focuses on the person. The person-oriented approaches are likely to provide the most useful guidance in determining what to assess for the purpose of selection and placement in entry-level jobs, and work-oriented assessments might prove more useful for identifying opportunities for and challenges to development and advancement.

Two Perspectives for Determining What to Assess

A number of important decisions must be made in determining what to assess, but the first is to determine whether the focus should be on the person or the work. That is, it is possible to build assessment strategies around the things people do in organizations in carrying out their work roles (work oriented) or

around the characteristics of individuals that influence what they do and how well they do it in the workplace (person oriented). For example, it is common to start the process of selecting and deciding how to use assessments with a careful job analysis on the assumption that a detailed examination of what people do, how they do it, and how their work relates to the work of others will shed light on the knowledge, skills, abilities, and other attributes (KSAOs) required to perform the job well. An alternative strategy is to start by examining the individual difference domains that underlie most assessments and to use knowledge about the structure and content of those domains to drive choices about what to assess.

The choice of specific assessments is a three-step process that starts with choosing between a broadly person-oriented or work-oriented approach, then making choices about the domains within each approach to emphasize (for example, whether to focus on cognitive ability or on personality), and finally narrowing down the choice of specific attributes (say, spatial ability) and assessment methods (perhaps computerized tests). As I noted earlier, this chapter focuses on the first two of these steps.

Work-Oriented Strategies

Different jobs involve very different tasks and duties and may call on very different sorts of knowledge or skill, but it is possible to describe the domain of work in general terms that are relevant across a wide range of jobs and organizations; such a wide-ranging description provides the basis for worker-oriented strategies for determining what to assess. Starting in the late 1960s, considerable progress was made in the development of structured questionnaires and inventories for analyzing jobs (for example, the Position Analysis Questionnaire; McCormick, Jeanneret, & Mecham, 1972). These analysis instruments in turn helped to define the contents and structure of the O*NET (Occupational Information Network; Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999) Generalized Work Activities Taxonomy, arguably the most comprehensive attempt to describe the content and nature of work. Table 1.1 lists the major dimensions of this taxonomy.

Table 1.1. O*NET Generalized Work Activities

Information input	Looking for and receiving job-related information
	Identifying and evaluating job-relevant information
Mental processes	Information and data processing
	Reasoning and decision making
Work output	Performing physical and manual work activities
	Performing complex and technical activities
Interacting with others	Communicating and interacting
	Coordinating, developing, managing, and advising
	Administering

If you were to ask, "What do people do when they work?" Table 1.1 suggests that the answer would be that they gather information, process and make sense of that information, make decisions, perform physical and technical tasks, and interact with others. The specifics might vary across jobs, but it is reasonable to argue that Table 1.1 provides a general structure for describing jobs of all sorts and for describing, in particular, what it is that people do at work. Each of these major dimensions can be broken down into subdimensions (which are shown in this table), most of which can be broken down even further (for example, administering can be broken down into performing administrative activities, staffing organizational units, and monitoring and controlling resources) to provide a more detailed picture of the activities that make up most jobs.

In the field of human resource (HR) management, the detailed analysis of jobs has largely been replaced with assessments of competencies. The term *competency* refers to an individual's demonstrated knowledge, skills, or abilities (Shippmann et al., 2000). The precise definition of competencies and the similarities and differences between traditional job analysis and competency modeling are matters that have been sharply debated (Shippmann et al., 2000),

and it is not clear whether competency modeling is really anything other than unstructured and informal job analysis. Nevertheless, the business world has adopted the language of competencies, and competency-based descriptions of work are becoming increasingly common.

Some competency models are based on careful analysis and compelling data, most notably the Great Eight model (Bartram, 2005):

Great Eight Competency Model

- Leading and deciding
- Supporting and cooperating
- Interacting and presenting
- Analyzing and interpreting
- Creating and conceptualizing
- Organizing and executing
- Adapting and coping
- Enterprising and performing

Bartram summarizes evidence of the validity of a range of individual difference measures for predicting the Great Eight. Unlike some other competency models, assessment of these particular competencies is often done on the basis of psychometrically sound measurement instruments.

Drilling Deeper

Work can be described in general terms such as the competencies detailed in the previous section. A more detailed analysis of what people do at work is likely to lead to an assessment of more specific skills and an evaluation of background and experience factors that are likely to be related to these skills. In this context, *skill* has a specific meaning: the consistent performance of complex tasks with a high level of accuracy, effectiveness, or efficiency. Skills are distinct from abilities in three ways: (1) they involve the performance of specific tasks, (2) they involve automatic rather than controlled performance, and (3) they are the result of practice. These last two features of skills are especially critical. The acquisition and mastery of skills usually requires a substantial amount of

practice or rehearsal, which suggests a link between assessment of skills and assessments of background and experience. In the past two decades, considerable progress has been made in assessments of background and experience (Mael, 1991), but it is fair to say that there are not well-established taxonomies of job-related skills or of background and experience factors, making it difficult to describe these domains in a great deal of detail.

Inferring Job Requirements

One of the most difficult challenges that proponents of worker-oriented approaches face is to convincingly translate information about what people do at work into judgments about the KSAOs required for performing well in particular jobs. This is sometimes done on an empirical basis (for example, the Position Analysis Questionnaire provides data that can be used to determine the predicted validity of a range of ability and skill tests), but it is most often done on the basis of subjective judgments. Virtually all methods of job analysis and competency modeling involve inferences about the attributes required for successful performance, but these judgments are rarely themselves validated. Indeed, there is little scientific evidence that given a good description of the job, analysts can make valid inferences about what attributes are required for successful performance beyond a handful of obvious prerequisites; knowing that electricians work with wires that are often color-coded, it is not hard to infer that color vision is required for this job, for example. Usually inferences of this sort are based on the assumption that if the content of the test matches the content of the assessments, those tests will be valid predictors of performance on the job.

Murphy, Dzieweczynski, and Yang (2009) reviewed a large number of studies testing the hypothesis that the match between job content and test content influences the validity of tests and found little support for this hypothesis. Nevertheless, an analysis of the job, whether it is done in terms of competencies, generalized work activities, or detailed questionnaires, is often the first step in making a decision about the content and the focus of workplace assessments.

Work-oriented approaches to assessment are likely to be particularly useful as part of the process of making decisions about

placement and development. In particular, comparisons between the content of previous and current jobs and the content of future jobs are useful for identifying developmental needs and gaps between the knowledge, skills, and experiences developed in previous jobs and those required in future assignments.

Person-Oriented Analyses

A very different strategy for making decisions about what attributes should or should not be included in assessments starts from the perspective of differential psychology: using what we know about individual differences to drive what we assess. In particular, this approach takes our knowledge of the dimensions and structure of human cognitive ability, normal personality, and interests and value orientations as a starting point for determining what to assess.

Cognitive Ability

There are enduring and stable individual differences in performance on virtually all tasks that involve the active processing of information; these individual differences form the core of the domain we refer to as cognitive ability.

The key to understanding the structure of human cognitive abilities is the fact that scores on almost any reliable measure that calls for active information processing will be positively correlated with any other reliable measure that also involves cognitive activity. That is, scores on virtually all cognitively demanding tasks exhibit positive manifold (Carroll, 1993). Thus, scores on paragraph comprehension measures will be correlated with scores on numerical problem solving, which will be correlated with scores on spatial relations tests and so on. The existence of positive manifold virtually guarantees that the structure of human abilities will be hierarchically arranged, with virtually all specific abilities (or groups of abilities) positively correlated with more general ability factors. Theories of cognitive ability that give little emphasis to *g* or deny the utility of a general factor do not seem to provide any convincing explanation for positive manifold.

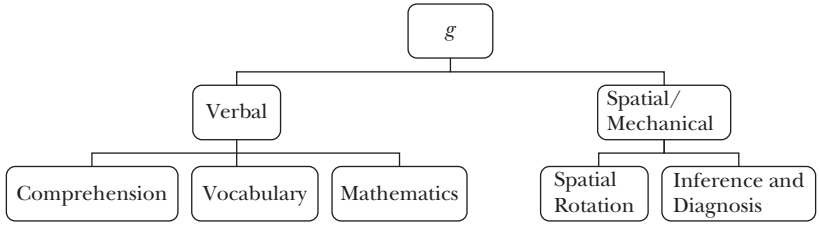
Carroll's (1993) three-stratum model of cognitive ability (based on the results of a large number of factor-analytic studies) nicely

illustrates the nature of modern hierarchical models. The essential features of this model are shown in Figure 1.1. At the most general level, there is a *g* factor, which implies stable differences in performance on a wide range of cognitively demanding tasks. At the next level (the broad stratum) are a number of areas of ability, which imply that the rank ordering of individuals' task performance will not be exactly the same across all cognitive tasks, but rather will show some clustering. Finally, each of these broad ability areas can be characterized in terms of a number of more specific abilities (the narrow stratum) that are more homogeneous still than those at the next highest level.

The hierarchical structure of the domain of cognitive abilities has important implications for understanding three key aspects of cognitive ability tests: (1) the validity of these tests as predictors of job performance and effectiveness, (2) the relationships among abilities and the relative importance of general versus specific abilities for predicting performance, and (3) adverse impact. First, abundant evidence shows that cognitive ability is highly relevant in a wide range of jobs and settings and that measures of general cognitive ability represent perhaps the best predictors of performance (Schmidt & Hunter, 1998). The validity of measures of general cognitive ability has been established in all sorts of jobs and settings, and it is reasonable to believe that a good ability test will be a valid predictor of performance in virtually any application of testing.

The hierarchical structure of the cognitive domain is almost certainly a key to the widespread evidence of the validity of cognitive tests. All jobs require active information processing (such as retrieving and processing information, making judgments), and

Figure 1.1. The Cognitive Domain



even when the content of the job focuses on very specific tasks or types of ability (a job might require spatial visualization abilities, for example), the strong intercorrelations among abilities virtually guarantee that measures of general ability will predict performance. This intercorrelation among cognitive abilities also has important implications for evaluating the importance of general versus specific abilities.

A good deal of evidence exists that the incremental contribution of specific abilities (over and above general ability) to the prediction of performance or training outcomes is often minimal (Ree, Earles, & Teachout, 1994). Because of the correlation among measures of general and specific abilities, payoff for the specific abilities required in a job is usually small. Measures of general ability will usually be as good as, and often better than, measures of specific abilities as a predictor of performance and effectiveness.

The strong pattern of intercorrelation among cognitive abilities poses a strong challenge to the hypotheses that many types of intelligence exist (Gardner, 1999) or that important abilities have not yet been fully uncovered. In particular, the overwhelming evidence of positive correlations among virtually all abilities raises important questions about the nature of emotional intelligence.

Organizations have shown considerable interest in the concept of emotional intelligence (EI: Murphy, 2006). There are many different definitions and models of EI, but the claim that it is a distinct type of intelligence is at the heart of the debate over its meaning and value. On the whole, little evidence exists that emotional intelligence is related to other cognitive abilities, casting doubts on its status as an "intelligence." Some evidence suggests that EI is related to a variety of organizationally relevant criteria, but on the whole, the claim that EI is a distinct type of intelligence and an important predictor of performance and effectiveness does not hold up to close scrutiny (Murphy, 2006). More generally, the idea that there are distinct types of intelligence does not square with the evidence.

Finally, the hierarchical structure of the cognitive domain has important implications for the likelihood that ability measures will lead to different outcomes for members of different ethnic and racial groups. Black (and, to a lesser extent, Hispanic) examinees

consistently receive lower scores on cognitive ability tests than white examinees, and the use of cognitive ability tests in personnel selection or placement will normally lead to adverse impact against black and Hispanic examinees (Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). Some differences in the amount of racial disparity are expected with measures of different specific abilities (in general, the stronger the correlation of specific abilities with *g*, the larger the racial disparities), but one consequence of the positive manifold among measures of various abilities is that adverse impact will be expected almost regardless of what specific abilities are measured. The hierarchical structure of the cognitive ability domain has several implications for research and practice in personnel assessment, including:

- General abilities have broad relevance in most settings.
- Identifying the right specific abilities is not necessarily important.
- The faults of general abilities will be shared with specific ones.
- The belief in multiple types of intelligence or newly discovered intelligences is not consistent with the data.

First, the hierarchical structure of cognitive abilities means that general abilities are more likely to be useful for predicting and understanding behavior in organizations than more narrowly defined specific abilities. This structure guarantees that even if it is the specific ability that is important, general abilities will also turn out to be good predictors in most settings. Because general abilities are usually measured with more reliability and more precision, making the case for focusing on specific rather than on general abilities is often hard.

Second, if the goal is predicting future performance and effectiveness, this structure suggests a diminishing payoff for getting it exactly right when drawing inferences about the abilities required by a job. For example, the spatial-perceptual branch of most hierarchical models of cognitive ability includes a number of specific abilities (say, three-dimensional spatial visualization versus the ability to estimate distance and range). The further down the chain of related abilities one goes (from general to spatial to

three-dimensional spatial visualization), the less difference choices among branches of the ability tree are likely to make in determining the eventual value and criterion-related validity of ability tests.

Third, the use of ability tests in making decisions about people in organizations such as personnel selection or placement will lead to adverse impact against members of a number of racial and ethnic groups, and the use of specific rather than general ability measures will rarely change this fundamentally. Specific ability measures do show slightly lower levels of adverse impact than general ones, but they also typically show lower levels of criterion-related validity. The decision to use cognitive ability tests in organizations is necessarily also a decision to accept a certain level of adverse impact; the decision to refrain from using such tests is almost always also a decision to sacrifice validity.

Finally, the long-standing assumption and hope of many researchers and practitioners (especially in educational settings) that we can identify many separate types of intelligence is exactly that: an assumption and an aspiration. Models that posit multiple intelligences or suggest that specific types of content such as emotional information require their own type of intelligence are popular but not well supported. In the case of emotional intelligence, which has attracted a great deal of attention in both educational and organizational settings, improvements in the models and measures of this construct may eventually lead to the acceptance of EI as a distinct and important domain of human cognitive ability, but there are few data on the immediate horizon to lead us to believe that current conclusions about the structure and nature of human cognitive ability will need to be radically changed to accommodate separate intelligences such as EI.

Personality

The link between personality and behavior in organizations has a long history of interest. In a highly influential review, Guion and Gottier (1965) cast considerable doubt on the value of personality measures, especially as predictors of job performance. They concluded that "there is no generalizable evidence that personality measures can be recommended as good or practical tools for employee selection" (p. 159) and that "it is difficult to advocate, with a clear conscience, the use of personality measures in most

situations as a basis for making employment decisions about people" (p. 160). This review led to a long period of skepticism about the relevance of personality in understanding performance and effectiveness in the workplace. Not until the 1990s did personality reemerge as a viable tool for understanding and predicting performance and effectiveness (Barrick & Mount, 1991). It is now widely accepted that measures of normal personality have some value as predictors of performance, but the validities of these measures are often low. Nevertheless, they are also viewed as useful measures for helping to structure and manage development and placement programs.

As with cognitive ability, one of the keys to understanding the relevance and value of personality measures is to examine the structure and the contents of this domain. The Five Factor Model, often referred to as the "Big Five," has emerged as a dominant model for describing normal personality. This model has been replicated across a number of methods, settings, and cultures, and it provides a good starting point for describing what exactly *personality* means. This model suggests that normal personality can be described largely in terms of five broad factors that are at best weakly related to one another and (with the exception of Openness to Experience) with cognitive abilities:

- Neuroticism: emotional instability, tendency to experience negative emotions easily
- Extraversion: outgoing, energetic, tending toward positive emotions
- Agreeableness: cooperates with, is compassionate and considerate toward others
- Conscientiousness: reliability, self-discipline, achievement oriented, planfulness
- Openness to Experience: curiosity, imagination, appreciation for new ideas and experiences, appreciation of art, emotion, adventure

The two structural aspects of the domain of normal personality that are most important for understanding the ways broad personality dimensions might be used in assessment are the relatively

weak correlations among the dimensions of normal personality and the relatively weak relationships between personality and cognitive ability. The weak correlations among the Big Five mean that different dimensions of personality really do convey different information and that all sorts of personality profiles are possible. The weak correlations between personality and cognitive ability have three very different and very important implications. First, personality measures will contribute unique information not captured by cognitive ability. That is, whatever variance in performance, behavior, or effectiveness is explained by personality will almost certainly be distinct from variance explained by cognitive ability. Second, personality measures will not share some of the characteristics common to ability measures. In particular, measures of normal personality are typically unrelated to the respondent's race, ethnicity, or gender.

Whereas the use of cognitive ability tests is a major cause of adverse impact in personnel selection, the use of personality measures can reduce adverse impact. Unfortunately, the reduction in adverse impact when ability and personality measures are combined is not as large as one might expect; the combination of ability tests (which have adverse impact) and personality inventories (which do not) leads to some reduction in adverse impact, but it will not cut it in half (Ryan, Ployhart, & Friedel, 1998). Third, the weak relationships between personality and cognitive ability are consistent with one of the most contentious issues in research on personality assessment in organizations: the validity of broad personality dimensions as predictors of performance and effectiveness. Although there is considerable interest in the use of personality assessments in organizations, studies of the validity of personality measures as predictors of performance have consistently shown that the correlations between personality and performance are small (Morgeson et al., 2007). If the goal is to predict performance and effectiveness, it is unlikely that measures of broad personality dimensions will help very much.

The two alternatives to using broad personality dimensions in assessment might yield higher levels of validity. First, it is possible to use finely grained measures. For example, measures of the Big Five often provide separate assessments of multiple facets

of each major dimension. For example, the NEO-PI (Costa & McCrae, 1995) yields scores on the Big Five and on several facets of each dimension; these are shown in Table 1.2. For example, Conscientiousness can be broken down into Competence, Order, Dutifulness, Achievement-striving, Self-discipline, and Deliberation. It is possible that different facets are relevant in different jobs or situations and from assessment of specific facets will yield different levels of validity from those that have been exhibited by measures of the Big Five.

An alternative to the use of finely grained measures is the use of composite measures. For example, there is evidence that

Table 1.2. Facets of the Big Five

Neuroticism	Extraversion
Anxiety	Warmth
Hostility	Gregariousness
Depression	Assertiveness
Self-consciousness	Activity
Impulsiveness	Excitement seeking
Vulnerability	Positive emotions
Conscientiousness	Agreeableness
Competence	Trust
Order	Straightforwardness
Dutifulness	Altruism
Achievement-striving	Compliance
Self-discipline	Modesty
Deliberation	Tender-mindedness
Openness	
Fantasy	
Aesthetics	
Feelings	
Actions	
Ideas	
Values	

integrity tests capture aspects of Conscientiousness, Neuroticism, and Agreeableness (Ones, Viswesvaran, & Schmidt, 1993); the breadth of the domain these tests cover may help to explain their validity as a predictor of a fairly wide range of criteria. In principle, there might be no effective limit to the types of composite personality tests that might be created, and some of these might plausibly show very respectable levels of validity. However, this strategy almost certainly involves a trade-off between the potential for validity and interpretability.

The use of personality assessments to make high-stakes decisions about individuals is controversial (Morgeson et al., 2007), in large part because most personality inventories are self-reports that are potentially vulnerable to faking. The research literature examining faking in personality assessment is broad and complex (Ones, Viswesvaran, & Reiss, 1996), but there is consensus about a few key points. First, people can fake, in the sense that they can often identify test responses that will paint them in the most favorable light. Second, while faking can influence the outcomes of testing, it often does not greatly affect the validity of tests. This is because positive self-presentation biases are often in play when job applicants and incumbents respond to personality inventories, meaning that everyone's scores might be inflated. Although faking is a legitimate concern, it is probably more realistic to be worried about the possibility of differential faking. That is, if some people inflate their scores more than others, faking could change both the mean score and the rank order of respondents. In other words, if everyone fakes, it might not be a big problem, but if some people fake more or better than others, faking could seriously affect the decisions based on personality inventories.

As with cognitive ability, the structure and nature of the domain of normal personality have important implications for research and practice in organizational assessment:

- The relative independence of major personality dimensions puts a greater premium on identifying the right dimensions and the right rules for combining information from separate dimensions.

- Personality measures provide information that is distinct from that provided by ability measures.
- The relatively low correlations with ability suggest that personality measures will be poor predictors of performance and effectiveness; the available evidence seems to confirm this prediction.
- Narrow dimensions of personality are easiest to interpret, but are often similarly narrow in terms of what they predict. The broadest dimensions show more predictive power but are hard to sensibly interpret.

First, the broad dimensions that characterize the Big Five are relatively distinct, which poses both opportunities and challenges. On the opportunity side, it is more likely that the complex models (for example, configural models, in which the meaning of a score on one dimension depends on a person's score on other dimensions) will pay off in the domain of personality than in the domain of cognitive ability. In the ability domain, the pervasive pattern of positive correlations among virtually all ability measures means it is hard to go too far wrong. Even if you fail to identify the exact set of abilities that is most important, you can be pretty certain of capturing relevant variance with measures of general abilities. In the personality domain, choices of which dimensions to assess and how to combine them are likely to matter. This also means that identifying the best way to use personality information is likely to be a much more difficult challenge than identifying the best way to use information about abilities.

Second, personality and ability seem to be largely independent domains. There are some broad personality dimensions that may be related to *g*, but most are not. This means that potential exists for personality measures to contribute to the prediction of performance and effectiveness above and beyond the contributions of ability measures. Unfortunately, as noted in our third point, this often does not happen. The validities of personality measures are statistically different from zero but are often not much greater than zero (Morgeson et al., 2007).

Finally, personality assessment often poses trade-offs. One trade-off is often between predictive power and interpretability

and another between ease of use and trustworthiness. Personality measures are usually self-reports, and they are not necessarily hard to develop. They are, however, vulnerable to faking. Ability tests have many defects, but at least it is hard to "fake smart." A personality inventory that shows an applicant to be high on Conscientiousness and Agreeableness might mean exactly what it appears to mean—or it might mean that the respondent knows that high scores on these dimensions are viewed favorably, and is faking.

Interests and Value Orientations

Organizational assessments are used not only to predict performance and efficiency but also to evaluate the fit between people and environments or jobs. Ability and personality measures can be very useful in assessing fit, but many discussions of fit focus on interests and value orientation, based on the argument that the congruence between the interests and the values of an individual and the environment in which he or she functions is an important determinant of long-term success and satisfaction. There are important questions about the extent to which fit can be adequately measured and about the importance of person-environment fit (Tinsley, 2000), but the idea of congruence between individuals and environments is widely accepted in areas such as career development and counseling. Numerous models have been used to describe the congruence between individuals and environments; Lofquist and Dawis's (1969) Theory of Work Adjustment represents the most comprehensive and influential model of fit. The theory examines the links between the worker's needs and values and the job's ability to satisfy those needs, and it also considers the match between the skills an individual brings to the job and the skills required for effective performance in that job.

Assessments of interests have long been an important part of matching individuals with jobs. Strong (1943) defined an interest as "a response of liking" (p. 6). It is a learned affective response to an object or activity. Things in which we are interested elicit positive feelings, things in which we have little interest elicit little affect, and things in which we are totally disinterested elicit apathy or even feelings of aversion. Interest measures are widely used to

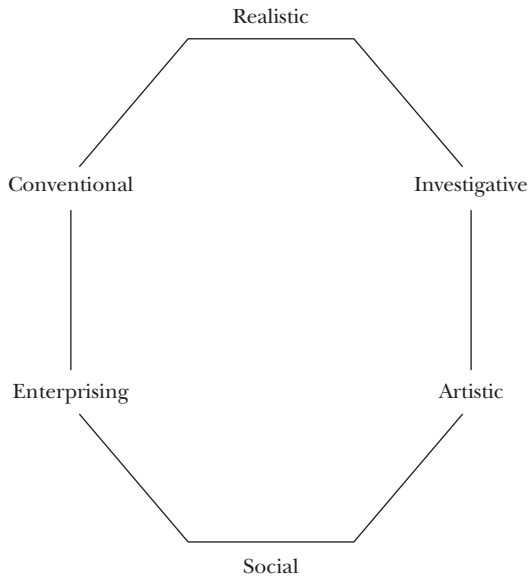
help individuals identify vocations and jobs that are likely to satisfy and engage them.

The dominant theory of vocational choice was developed by Holland (1973), who suggested that vocational interests can be broken down into six basic types: realistic (interest in things), investigative (interest in ideas), artistic (interest in creating), social (interest in people), enterprising (interest in getting ahead), and conventional (interest in order and predictability). The Holland RIASEC model is shown in Figure 1.2.

The hexagonal structure of Figure 1.2 reflects one of the key aspects of the Holland model. Interests that are close together on the Holland hexagon, such as Realistic and Investigative, are more likely to co-occur than interests that are far apart such as Realistic and Social. The great majority of measures of vocational interests and theories of vocational choice are based on the Holland model.

Unlike the field of interest measurement, there is no single dominant model of work-related values. Probably the best-researched

Figure 1.2. Holland Taxonomy of Vocational Interests



model is that proposed by Lofquist and Dawis (1969). Their taxonomy of work-related values, shown in Table 1.3, was adopted by O*NET as a way of characterizing the values most relevant to various occupations.

Like many other taxonomies, the O*NET Work Value Taxonomy is hierarchically structured. At the highest level of abstraction, jobs can be characterized in terms of the extent to which they are likely to satisfy value related to opportunities for achievement, favorable working conditions, opportunities for recognition, emphasis on relationships, support, and opportunities for independence. One of the many uses of O*NET is to match jobs to people's values. For example, individuals who value achievement and recognition can use O*NET to identify jobs that are likely to satisfy those preferences. The lower level of the taxonomy helps to clarify the meaning of each of the higher-order values and provides a basis

Table 1.3. O*NET Work Value Taxonomy

Achievement	Relationships
Ability utilization	Coworkers
Achievement	Social service
	Moral values
Working conditions	Support
Activity	Company policies and practices
Independence	Supervision, human relations
Variety	Supervision, technical
Compensation	
Security	
Working conditions	
Recognition	Independence
Advancement	Creativity
Recognition	Responsibility
Authority	Autonomy
Social status	

for a more finely grained assessment of person-job fit. For example, good working conditions might refer to almost any combination of opportunities for activity, independence, variety, compensation, or job security.

Assessments of cognitive abilities and personality traits are often used to predict criteria such as performance and effectiveness. Assessments of interests and values are not likely to reveal as much about performance, but are related to criteria such as satisfaction, burnout, and retention. Good person-job fit is thought to enhance the attractiveness and the motivational potential of a job, and in theory these assessments can be used for both individual counseling and placement. In practice, systematic placement (hiring an individual first and deciding afterward what job or even what occupational family to assign that person to) is rarely practiced outside the armed services. However, interest measures might be quite useful for career planning activities at both the individual and the organizational levels. For example, executive development programs often involve a sequence of jobs or assignments, and the use of interest and value assessments might help in fine-tuning the sequence of assignments that is most likely to lead to successful development.

Implications for Assessment in Organizations

Individual differences in cognitive ability, personality, values, and interests are likely to affect the performance, effectiveness, motivation, and long-term success of workers at all levels in an organization. A general familiarity with the structure and the content of each of these domains provides a good starting point for designing organizational assessments.

The essential feature of cognitive abilities is their interrelatedness. This presents both opportunities and challenges when using ability tests in organizations. Because virtually all abilities are correlated (often substantially) with general abilities, it is hard to go seriously wrong with the choice of ability measures; jobs that require one ability also tend to require the constellation of other related abilities. Because virtually all jobs require and involve the core activities that define cognitive ability (the acquisition, manipulation, and use of information), it is generally a safe bet

that ability measures will turn out to be valid predictors of performance. Unfortunately, the interconnectedness of abilities also implies that any of the shortcomings of general cognitive ability as a predictor will be broadly shared by more specific measures. In particular, ability measures of all sorts are likely to show substantial adverse impact on the employment opportunities of black and Hispanic applicants and employees, and this impact has both legal and ethical implications. Depending on the weight you give to predictive validity versus the social impact of using ability tests to make high-stakes decisions, you might come to very different conclusions about whether including these measures in organizational assessments makes sense (Murphy, 2010).

The domain of normal personality has a much different structure. The Big Five personality factors are interrelated, but the correlations among dimensions are generally quite weak, and no general factor describes human personality. Like many other taxonomic structures, the Big Five can be broken down into facets, or they can be combined into composites, but moving from the level of the Big Five to either higher (composite) or lower (facet) levels of abstraction often involves trade-offs between interpretability and predictive value.

Two issues seem especially important when using personality measures as part of assessment in organizations. First, these are usually self-reports and are vulnerable to manipulation and misrepresentation. There are important debates about the actual effects of faking on validity and the outcomes of selection (Morgeson et al., 2007), but the possibility that respondents might be able to consciously inflate their scores on high-stakes assessments is likely to be a realistic barrier to their use in many settings. More important, the validity of these measures as predictors of criteria such as performance or effectiveness is often disappointing, and the value of obtaining these assessments is not always clear.

Vocational interests are well understood and are captured nicely by Holland's hexagonal model. This model posits relationships among interests that can be captured by the distance between any pair of interests on the hexagon; this model has been applied with considerable success in vocational counseling. However, it is not always clear how to use assessments of interests or values to make

more detailed predictions of judgments. There are many models of person-job fit, and different models often depend on different sets of values. No single agreed-on taxonomy adequately captures the universe of organizationally relevant values. Nevertheless, the general proposition that some jobs are more likely than others to fit an individual's values and that some individuals are more likely than others to fit any specific job seems well established, and the measurement of work-related values has potential for both research and practice.

This chapter has been intentionally broad in its focus, and the implications for assessment laid out in the preceding paragraphs are similarly broad. Chapters Two through Eight examine more specific issues in assessments of domains ranging from abilities to personality to background and experience. Chapters Nine through Fourteen show how assessments of these domains are used in making decisions in occupations ranging from hourly or skilled work to executive and managerial positions. Chapters Fifteen through Twenty-Four discuss a wide range of questions encountered when developing and using assessments in a range of organizational contexts.

References

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64*, 21–50.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century*. New York: Basic Books.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18*, 135–164.

Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Upper Saddle River, NJ: Prentice Hall.

Lofquist, L. H., & Dawis, R. V. (1969). *Adjustment to work*. New York: Appleton–Century–Crofts.

Mael, F. A. (1991). A conceptual rationale for the domain of attributes of biodata items. *Personnel Psychology, 44*, 763–792.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 56*, 347–368.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729.

Murphy, K. R. (2000). What constructs underlie measures of honesty or integrity? In R. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: A festschrift to Douglas N. Jackson at seventy* (pp. 265–284). Norwell, MA: Kluwer.

Murphy, K. R. (2006). *A critique of emotional intelligence*. Mahwah, NJ: Erlbaum.

Murphy, K. (2010). How a broader definition of the criterion domain changes our thinking about adverse impact. In J. Outtz (Ed.), *Adverse impact* (pp. 137–160). San Francisco: Jossey-Bass.

Murphy, K. R., Dzieweczynski, J. L., & Yang, Z. (2009). Positive manifold limits the relevance of content-matching strategies for validating selection test batteries. *Journal of Applied Psychology, 94*, 1018–1031.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660–679.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L (1993). Comprehensive meta-analysis of integrity test validities. *Journal of Applied Psychology, 78*, 679–703.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (1999). *An occupational information system for the 21st century: The development of O*NET*. Washington, DC: American Psychological Association.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79*, 518–524.

Ryan, A. M., Ployhart, R. E., & Friedel, L. A. (1998). Using personality testing to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83*, 298–307.

Schippmann, J. S., Ash, R. A., Carr, L., Hesketh, B., Pearlman, K., Battista, M. et al. (2000). The practice of competency modeling. *Personnel Psychology, 53*, 703–740.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*, 719–730.

Strong, E. K. (1943). *Vocational interests of men and women.* Stanford, CA: Stanford University Press.

Tinsley, H. E. (2000). The congruence myth: An analysis of the efficacy of the person-environment fit model. *Journal of Vocational Behavior, 56*, 147–179.

CHAPTER 2

INDICATORS OF QUALITY ASSESSMENT

Fritz Drasgow, Christopher D. Nye, Louis Tay

Assessment, whether for selection or development, can play a critical role in elevating an organization from mediocrity to excellence. However, this is true only if the assessment is excellent. In this chapter, we describe the characteristics and features that differentiate outstanding assessment programs from mediocre systems. With this information, organizational practitioners can thoughtfully consider how assessments can be implemented in their organizations, evaluate any current uses of tests and assessments, and move toward state-of-the-art measurement.

When an organization decides to begin an assessment program, its first decision concerns whether to purchase a test from a test publisher or consulting firm or develop the assessment tool in-house. We begin the chapter by reviewing the issues to consider when making this important decision. We next discuss the test construction process, which begins with the question, "What should the test measure?" and addresses item writing, pretesting, and psychometric analyses. The next two sections examine the critical quality issues of reliability and validity. Obviously organizations want their assessments to be reliable and valid, but there are some subtleties that test users should understand in order to make informed judgments; we summarize these issues.

We then discuss operational models for assessment programs. With advances in computer technology and the Internet, organizations have a dizzying array of choices. Some of the topics discussed

include testing platform (paper-and-pencil versus computer), unproctored Internet testing, cheating, and score reporting.

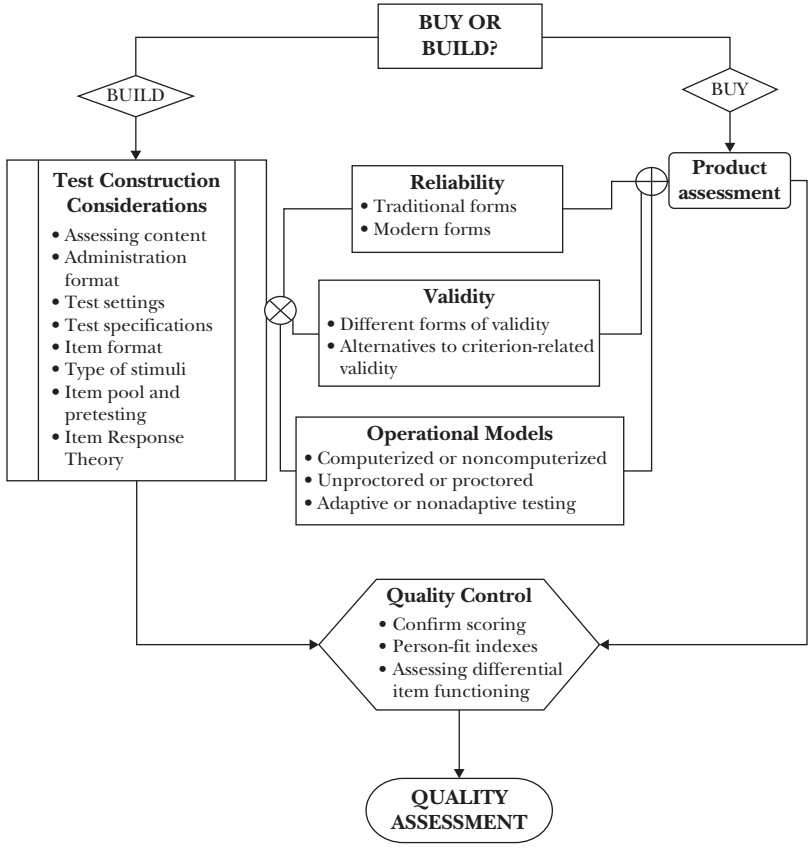
The next section of the chapter addresses quality control, a topic that receives little attention in many testing programs. There have been several highly publicized fiascos in high-profile testing programs in recent years and undoubtedly many other problems that were kept under wraps. Consequently, we discuss issues to consider and steps organizations can take to ensure high quality. Finally, we end with a few brief conclusions.

We expect that people with diverse backgrounds will read this chapter. We encourage those with psychometric training, including classical test theory (CTT) and item response theory (IRT), to dig into the technical details that are needed to fully address the quality of a testing program. To this end, the equations that are referenced throughout the chapter are in Table 2.1 for convenience. For those who do not have this background, we encourage looking at the big picture to gain an understanding of critical issues. We have attempted to give conceptual descriptions of each topic so that all readers can understand important problems; a flowchart of the key processes related to quality assessment is shown in Figure 2.1. Organizational leaders can then consult with either internal or external measurement professionals for guidance on technical concerns.

Buy Versus Build

If a decision is made to implement an assessment program, the organization must decide whether to purchase a commercially available test or develop a measure in-house. To make this decision, organizations need to weigh the costs and benefits of each approach to determine which will be more appropriate, and a number of questions must be addressed. First, do any currently available tests meet the needs of the organization? Specifically, do the commercially available tests validly measure the requisite knowledge, skills, and abilities (KSAs) and have rigorous empirical support for their validity? Commercial tests frequently assess constructs that are broadly focused and applicable across a wide range of jobs. Although the predictive validity of many general

Figure 2.1. Flowchart of Key Processes in Quality Assessment



Note: ⊗ represents the confluence of decision factors associated with test construction; ⊕ represents the confluence of decision factors associated with product assessment.

constructs such as general cognitive ability has been established by meta-analytic research, these tests may be criticized for their apparent lack of job relevance and face validity. In contrast, a homegrown test can be developed to measure the specific knowledge, skills, or ability obviously required in the target job.

A related concern involves the empirical evidence supporting the validity of a measure. Although most practitioners are aware of legal and professional guidelines stressing the importance of validity, organizations might not employ a sufficient number of people in a particular job category to obtain an accurate estimate of the correlation between a predictor and an outcome. As a result, irrelevant characteristics of the available sample (known as sampling error) may severely affect the magnitude of the observed relationship. This effect is largely responsible for the variation in the size of the relationship between predictor and criterion across organizational settings (Schmidt & Hunter, 1977). In contrast, the best commercial tests have substantial empirical evidence about validity; an evaluation of the extent and rigor of validity evidence is a key consideration in choosing which commercial test to purchase. Thus, in situations where organizations lack the sample size for appropriate validation studies, commercially available tests may be the only viable alternative.

Another consideration in the buy-versus-build decision concerns whether an organization seeking to develop a test in-house has the necessary skills and resources to do so. Expertise in test development, test administration, and statistical analysis may not be available. Test development, for example, requires carefully defining the KSAs to be assessed, developing a test blueprint that specifies the content areas to be measured and the number of items from each content area to be included, and then writing a sufficient number of items for each content area. Thus, a specialized knowledge of the subject matter is required, and a substantial review of relevant literature will frequently be necessary. Even if a sufficient understanding of the construct can be obtained, it may still be difficult to write discriminating items at the appropriate difficulty levels.

Once the items have been written, psychometric knowledge is required to ensure that the test has appropriate measurement properties. For example, statistical analysis using IRT is often used for this purpose. Briefly, IRT is a psychometric method that can be used to predict an individual's performance on a test item by relating characteristics of the item to the ability of the person on the latent trait being measured (here, the term *latent* is used to reference a characteristic such as intelligence or personality that

is not directly observable). However, many organizations do not have expertise in IRT. Additional difficulties may be encountered when using computerized or computer-adaptive tests (CATs). (See Drasgow, Luecht, & Bennett, 2006, for a description of the complexities of a technologically sophisticated testing program.)

Other decision factors are the time lines and the breadth of the assessment program. Organizations with an immediate need may benefit from purchasing a commercially available test because test development and validation can be time-consuming. An organization should also consider how frequently the test will be used. For example, if a brief unproctored Internet test is used as an initial screening for tens of thousands of job applicants, a commercially available test may become very expensive as costs accrue with each administration. In contrast, organizations that use assessments only intermittently may not recoup the cost of developing the test in-house.

Finally, test security should also be considered. Are cheating conspiracies likely? Test security is enhanced by using multiple forms of paper-and-pencil tests. Even more effective is the use of a CAT with a large item pool and an effective item exposure control algorithm. To illustrate the significance of this problem, note that there have been significant cheating conspiracies involving college entrance and licensure exams. Even multiple conventional forms and CATs can be susceptible to large-scale cheating conspiracies, such as online sharing sites and companies devoted to cracking the tests. One benefit of commercially developed assessments is that the developers are well positioned to ensure the security of the exam because their business success is severely affected by cheating conspiracies. Some professional test developers may even employ individuals with the sole responsibility of searching for and eliminating item-sharing sites and companies.

In sum, the buy-versus-build decision involves considerations of availability, feasibility, timeliness, in-house expertise, cost, and so forth. Clearly this is a critical and complex choice. Regardless of the buy-versus-build decision, a quality assessment must be created by a careful process. In the remainder of this chapter, we provide more details about this process and note criteria for evaluating quality. Before deciding to build a test, an organization should evaluate whether it has the resources necessary to

perform the steps we describe. And before buying, the organization should examine documentation from the test publisher to ascertain whether the criteria we describe next are satisfied.

Test Construction Considerations

Several steps in the development process have a critical impact on the quality of an assessment. Integrating these steps provides a systematic approach to test development and ensures a high-quality result. A less-systematic approach may produce a test that misses important aspects of the KSAs to be assessed, which is likely to reduce the effectiveness of the assessment.

The first step in test development lies in identifying what a test is intended to measure. Here, test developers establish the content that will be assessed. In an employment setting, this is most frequently done with a thorough job analysis. Test developers may survey or interview subject matter experts, examine critical incidents, or rely on expert judgment. Because it is usually impossible to assess all important KSAs for a particular job or job family, the criteria for including content should be based on information provided by the job analysis regarding the importance of each dimension. For psychological phenomena such as intelligence, personality, and attitudes, inclusion criteria should also be based on a careful definition of the trait to be assessed, followed by a thorough review of the literature on the topic.

The second step is to determine the testing format that is most appropriate for the purposes of the test. With the large number of administration formats now available for psychological testing, this issue is fundamental to the test construction process. In addition to the traditional paper-and-pencil format, a conventional test (one in which all examinees are administered the same set of items) may also be administered by stand-alone computers or using the Internet. In contrast to fixed conventional tests, CATs select items to be appropriately difficult for each examinee. This format is increasing in importance, particularly for licensing and credentialing exams. Another choice is the setting for test administration. In contrast to the traditional proctored environment,

unproctored Internet testing allows unsupervised examinees to take the exam at a time and place of their convenience.

Each of these testing formats has implications for the type and number of items used. In a computerized format, novel item stimuli may be presented interactively as audio, video, pictures, or some combination of media. For CATs to operate effectively, a large pool of items is required to ensure accurate ability estimates and increase test security. Similar security issues are salient for unproctored tests. As a result, it is often advisable to administer both an unproctored selection test and, later, a proctored confirmation test to verify results. Here, the proctored confirmation test may be a parallel form of the unproctored exam.

The choice between administration methods may also affect the third step in test construction where test specifications are formulated. These guidelines should be used as a road map for item writers. For example, test specifications would detail the number of items assessing verbal, quantitative, and spatial abilities in a measure of cognitive ability. In addition, these plans may specify the item difficulty and discrimination levels required for accurate ability estimates. These criteria are particularly important for CATs, where the quality of ability estimates improves when the item pool contains items with a wide range of difficulties.

The test specifications give the appropriate number and content of items as well as the format for the test. The number of items should be chosen based on considerations for reliability, content coverage, and test security. However, workplace assessments must effectively balance content sampling with space and time limitations. Assessments with too few items may not adequately measure the entire domain of the trait (content validity) or provide consistent results (reliability). And assessments with too many items may result in test-taker fatigue or negative reactions and may not be appropriate for situations with strict time constraints.

The choice of the item format may mitigate some of the disadvantages traditionally associated with measurement. For example, forced-choice response formats, where respondents must choose between two or more items matched on social desirability, may reduce the prevalence of faking on personality items. Other novel stimuli may also be appropriate. Video- or computer-simulation

tests may provide an effective means for measuring context-based phenomena such as emotional intelligence or situational judgment.

The next step is to create an item pool. Ensuring content and construct validity through the generation of appropriate items is one of the most difficult and important tasks of the test developer. Content validity addresses the appropriateness of the content covered by the test, whereas construct validity examines whether the test assesses the trait it is designed to measure (see the section on validity below and Chapter Twenty for further discussion). Without these forms of validity, the interpretation of results may be difficult. Developing content-valid items will be easier if the domain is well defined and the test specifications ensure adequate content coverage. However, generating construct-valid items can be more complex. It is surprisingly difficult to develop items that assess a single construct; other traits may be substantially correlated with an item because of common underlying antecedents.

Issues with the validity of items in the pool illustrate the importance of the next step in test development: item pretesting. Few test developers would put a new item on an operational test without first evaluating its measurement properties in a pretest sample. Ideally a large and representative sample is used for pretesting. Perhaps the best situation is one in which new pretest items are embedded in operational test forms and administered to job applicants; this is what is done to pretest items for the Armed Services Vocational Aptitude Battery (ASVAB). Then items can be evaluated statistically with classical test theory (CTT) statistics such as the proportion right, $\hat{p}$, the item-total point-biserial correlation, and the item-total biserial correlation. These analyses are often used as a first step in the evaluation process.

Many testing programs also use IRT to conduct item-level analyses. Although IRT techniques are mathematically complex, there are several important benefits to using them in addition to the traditional CTT methods. First, IRT item parameters are invariant across samples of test takers. Thus, in contrast to CTT statistics that are affected by the ability distribution of the sample, IRT

parameters will be equivalent across groups. For example, items on a job knowledge test may appear difficult for a sample of novices (a low-ability sample) while simultaneously appearing easy for more experienced workers (a high-ability sample) when CTT statistics are used. This is especially important when the sample used to pretest items (say, current employees who are experienced) differs from the sample for which the test will be used (job applicants who are likely to be novices). Second, IRT methods can be used to ensure that a test adequately assesses ability or skill at key points on the latent trait continuum (for example, at important cut scores). Finally, ability estimates are invariant across items. Whereas the number-right score of CTT is affected by the difficulty of the items (it is harder to obtain a high score on a test with more difficult items), IRT ability estimates take into account item characteristics such as difficulty; this is the key reason IRT is needed for CAT. Given these important characteristics, we discuss IRT methodology as well as traditional methods throughout the rest of this chapter.

The basic building block of IRT is the item response function (IRF), which describes the relationship between the probability of correctly answering an item and an individual's latent trait level. Figure 2.2 shows the proportion correct on an item for respondents who answered different numbers of items correctly on a thirty-item test. Clearly the proportion correct increases for individuals with higher test scores. Replacing the number-correct score with θ, the latent trait of IRT, leads to the IRF illustrated in Figure 2.3. This IRF can be represented for item i by equation 1 in Table 2.1. Here $u_i = 1$ indicates a correct response was made to item i and $P(u_i = 1|\theta)$ is the probability of that positive response given an examinee trait level θ. In equation 1, a_i represents the item discrimination or the steepness of the IRF, b_i represents the item difficulty, and c_i represents the guessing parameter.

After items have been selected, the final step in the development process is to evaluate the quality of the test as a whole. The primary criteria for this evaluation are the reliability and validity of the assessment. In the following sections, we address each of these issues.

Figure 2.2. Proportion Correct on an Item by Individuals with Different Total Test Scores

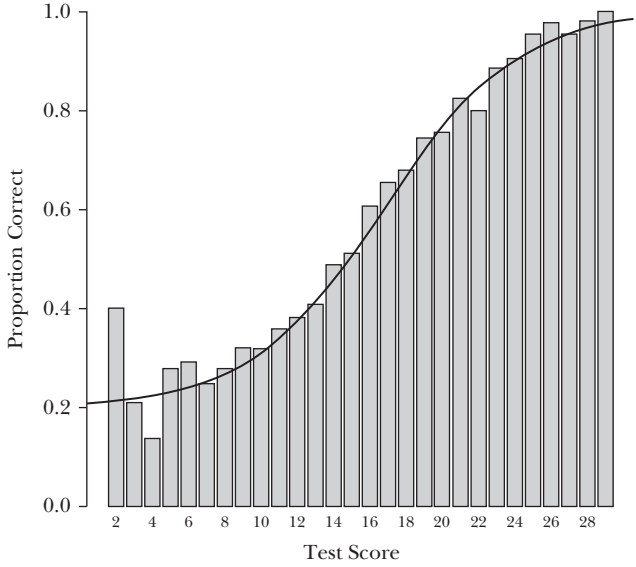


Figure 2.3. Three-Parameter Logistic Item Response Function for a Hypothetical Job Knowledge Test

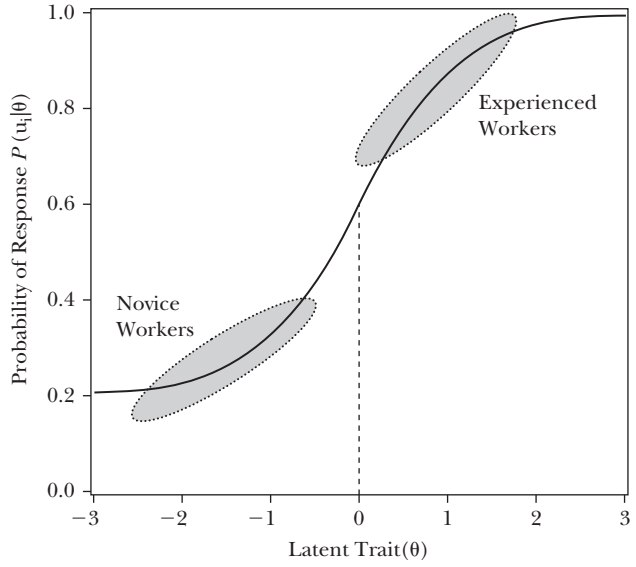


Table 2.1. IRT and CTT Equations for Evaluating Quality Assessments

Description	*Equations*	
Item response function	$P(u_i = 1 \mid \theta) = c_i + \dfrac{1 - c_i}{1 + \exp[-1.702a_i(\theta - b_i)]}$	Equation 1
Standard error of measurement in CTT	$SE(X) = \hat{\sigma}_x \sqrt{1 - r_{xx}}$	Equation 2
Number right score	$X = \sum_{i=1}^{n} u_i$	Equation 3
Relationship between τ and θ	$\tau = E(X) = \sum_{i=1}^{n} P_i(\theta)$	Equation 4
Conditional standard error of measurement in IRT	$SE(X \mid \tau) = SE(X \mid \theta_\tau) = \sqrt{\sum_{i=1}^{n} P_i(\theta_\tau)[1 - P_i(\theta_\tau)]}$	Equation 5

Reliability

Reliability refers to the extent to which test scores are consistent or free from random error. It is a crucial property because no test can be valid unless it is reliable. Just as a test can be valid for one purpose but not another, a test can be reliable in one context but not in another. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) state that test users have the responsibility of determining whether a measure of reliability is relevant to their intended use and interpretation. If the reliability estimate is not relevant, test users have the obligation to obtain the necessary reliability evidence.

Although it is difficult to draw a line in the sand, reliability should be in the neighborhood of .90 (or greater) for high-stakes decisions (such as hiring versus not hiring) based on one test score. If a measure is used in a selection composite or as one of several pieces of

information considered when making a high-stakes decision, reliability should be at least .80. A measure that does not reach an adequate level of reliability should be revised.

Traditional Forms of Reliability

In this section we review traditional measures of reliability and their limitations. These reliability indexes all range from 0 to 1, with 1 indicating perfect reliability.

Test-retest reliability is estimated by administering a test or scale to a sample at two points in time and then correlating the scores. It is an important index for characteristics that should be stable across time. For example, intelligence is a highly stable trait, and consequently a minimal requirement for an intelligence test is to have substantial test-retest reliability.

Internal consistency reliability includes split-half reliability, the Kuder-Richardson KR20 and KR21 reliabilities, and Cronbach's coefficient alpha. All of these measures are functions of the intercorrelations of the items constituting a test. Thus, for a fixed test length, internal consistency reliability is higher when the test's items are more strongly correlated.

Reliability coefficients can be manipulated and artificially inflated. Therefore, it is important to consider several factors when interpreting a reliability coefficient, including test content, interitem correlations, test length, and the sample used to estimate reliability.

By incorporating highly redundant items, it is possible to manipulate reliability (and particularly internal consistency reliability) to produce substantially inflated values. Therefore, before giving credibility to a measure of reliability, it is important to examine the content of the measure for substantive richness and breadth. A narrow and excessively redundant measure may have an internal consistency reliability in excess of .95 but nonetheless be lacking in regard to other important properties, such as construct validity, which would reduce its correlation with job performance and other important variables.

For many types of measures, the average interitem correlation should fall in the range of .15 to .50. Having several items that are highly correlated (for example, .80) indicates excessive

redundancy. For example, when assessing conscientiousness, two items might be, "I am careful in my work" and "I am meticulous in my work." Or in assessing math ability, two items could be restatements of the same problem but employ different numbers. Because variants of the same item will be answered by applicants in similar ways, such redundant items should be excluded because they ostensibly increase reliability but do not truly add new information.

Classical test theory shows that reliability can be increased by adding more items. Some high-stakes licensing exams, for example, consist of several hundred items. If a test has a long form and a short form, the reliability of the long form should be larger than the reliability of the short form, and it is important not to confuse the two. Unfortunately, high reliabilities of long tests are sometimes mistaken as indicating unidimensionality.

Reliability also depends on the characteristics of the sample. Range restriction, which occurs when the selection process has resulted in a sample that displays a truncated range of test scores, lowers inter-item correlations and results in lower reliability. Conversely, an artificially broad sample for example, using a sample of third-, fourth-, and fifth-grade students to estimate the reliability of a math achievement test designed for fourth graders, will inflate reliability. Because estimates of test reliability are sample dependent, it is important to ask whether the sample that was used to estimate test reliability is similar to the sample used for a specific organizational assessment purpose. If it is not, then the reliability estimate will be less informative for the organization.

Perhaps the greatest limitation on test-retest reliability results from the fact that reliability is sample dependent. Test-retest reliability is often estimated in a small, experimental study because it is difficult to administer the same test twice to a random sample under operational conditions. Thus, the question arises of whether results from the sample in the small research study can be generalized to other groups of test takers. Answering this question can be difficult or impossible.

Because reliability is subgroup dependent, it is inappropriate to say, "The reliability of test X is .92." Instead, a statement about reliability should include information about the group for which it was computed.

Additional concerns can be seen by looking at the technical definition of *reliability* as defined with classical test theory (the squared correlation between true scores, that is, the hypothetical scores people would receive if assessed with a perfect test, and observed scores). For example, the traditional reliability index is uninformative as to test precision at different score levels; one value of reliability is given for the test. Similarly, the standard error of measurement (the standard deviation of observed scores around the examinee's true score) of a test score X is given by equation 2 in Table 2.1, where $\hat{\sigma}_x$ is the standard deviation of test scores, and r_{xx} is the test's reliability; no differentiation is made between high, low, or moderate values of X. In many situations, it is critical to determine the test's precision at important cut scores where high-stakes decisions are made (AERA/APA/NCME, 1999).

Although CTT provides only a single standard error of measurement for a test, the standard error in IRT is conditional on the level of the latent trait. Thus, we denote the conditional standard error of measurement at a given true score τ by $SE(X|\tau)$ and the conditional standard error at a given latent trait score θ by $SE(X|\theta)$. These values, computed using IRT, allow test users to understand the magnitude of measurement error at critical score ranges.

Modern Forms of Reliability

A modern perspective on reliability is grounded in IRT, so the details are more complicated.

If number-right scoring is used on a test, the total test score is determined by counting the number of items answered correctly. Mathematically, the number-right score can be defined by equation 3 in Table 2.1, where X is the total score on the n item test and the score on item i is coded $u_i = 1$ if correct and 0 if incorrect. It can be shown that there is a one-to-one correspondence between the true score τ of classical test theory and the θ of IRT when the assumptions of IRT hold (see equation 4 in Table 2.1). Using θ_τ to indicate the value of θ corresponding to a particular true score τ, the conditional standard error of measurement is given in equation 5 in Table 2.1.

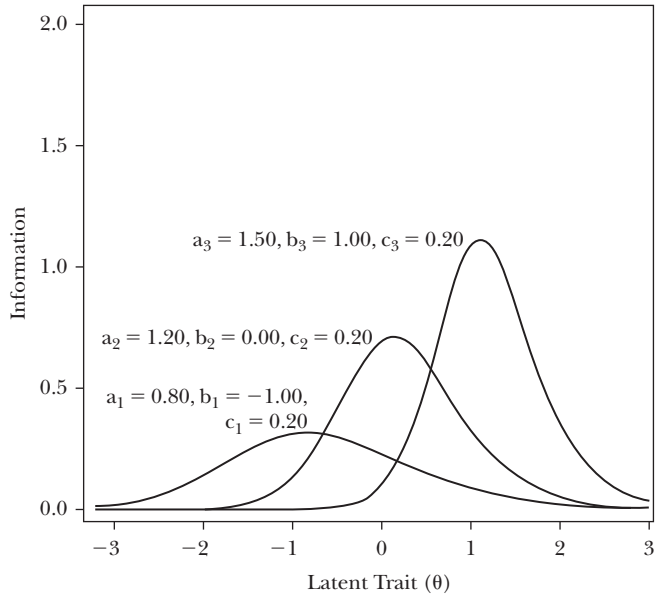
An alternative process can be used to compute the standard error of the estimate $\hat{\theta}$ of θ. This begins with the item information

curve, as shown in Figure 2.4, which can be constructed for each item by using its item parameters a_i, b_i, and c_i. Notice that the peak of each item information function is close to the difficulty (b_i) of the item. Moreover, items with greater discrimination (larger a_i values) yield more information.

The test information curve (TIC), denoted as $I(\theta)$, is the sum of the item information curves. The TIC is important because it is inversely related to the conditional standard error of $\hat{\theta}$, specifically, $SE(\hat{\theta} \mid \theta) = \left(1 / \sqrt{I(\theta)}\right)$. Note that IRT formalizes the intuition that items that discriminate at specific ability levels are most informative at those ability levels but much less informative at other levels.

Item information curves play a critical role in test development. One can examine the item information curves of all the items that have been pretested and then select the items that yield the most information at trait levels corresponding to important cut scores. In this way, test length can be minimized while providing highly precise measurement at the cut scores.

Figure 2.4. Example of Three-Item Information Curves for Items with Varying Levels of Difficulty and Discrimination



A much shorter test can be more informative than a longer test at a specific point on the trait continuum (for example, the cut score for determining passing on a licensing exam). Consider a shorter test formed by selecting items that are most informative at the specific trait level and a longer test constructed using a traditional approach where items are selected with varying levels of difficulty. An IRT analysis could show that the standard error of measurement at the cut score was smaller for the shorter test. If a test is built to be precise at only one cut score, it can be much shorter than a test built by traditional means, perhaps a third or a quarter as long.

Validity

When evaluating the quality of a test, it is important to assess both reliability and validity. Whereas reliability reflects the consistency with which the latent trait is measured, validity provides the justification for the inferences drawn from scores on the test. Although both factors play an active role in high-quality assessments, the 1985 *Test Standards* stated clearly and unambiguously that "validity is the most important consideration in test evaluation" (AERA/APA/NCME, 1985, p. 9).

Although organizational psychologists ordinarily use the term *validity coefficient* to refer to the correlation between a test *X* and a measure of job performance *Y*, *validity* is actually a much broader term. The 1999 *Test Standards* stated, "Validity refers to the degree to which evidence and theory support the interpretations of test scores" (AERA/APA/NCME, 1999, p. 9). There are a multitude of sources of evidence, including test content, the internal structure of the test, convergent and discriminant validity, test-criterion relationships, and validity generalization (see Chapter Twenty-One for a full description of these concepts).

Before purchasing a test, an organization should consider the evidence in relation to these aspects of validity. Alternatively, if an organization decides to create its own test, it should begin to accumulate these types of evidence to support its test use. In the remainder of this section, we review the various types of validity evidence and comment on challenges that may be encountered.

At the beginning of the test development process, the developer should carefully define the KSAs or other characteristics to be assessed by the test. Then the test blueprint should specify the content areas to be tested and how many items to include from each area. A first question is therefore, "Was the test actually built according to the original trait definition and test blueprint?" Yogi Berra is credited as saying, "You can see a lot by just looking," and the obvious implication is that test users should examine a test's content to ascertain whether it matches the test blueprint. For employee selection and development, the organization should ask whether the test assesses KSAs or some other characteristic that a job analysis indicated is critical for individuals in a particular job.

Factor analysis is frequently used to examine the internal structure of a test. If, for example, the trait definition states that a unitary characteristic is evaluated by the test, then a dominant first factor should appear. If a measure of emotional stability has facets of No Anxiety, Even Tempered, and Well-Being, a three-factor structure should be obtained.

Convergent and discriminant validity evidence is also important. Convergent validity is obtained when the test is positively and substantially correlated with another measure that it should correlate with. Discriminant validity is obtained when the test is not correlated with measures of theoretically distinct traits. Measures of emotional intelligence, for example, have been criticized as not exhibiting discriminant validity because they are excessively correlated with measures of other well-established traits (Davies, Stankov, & Roberts, 1998).

For good reasons, organizational psychologists have historically emphasized criterion-related validity, and consequently the most often used measure of the test-criterion relationship is the correlation coefficient. In fact, the massive quantity of such correlations enabled Schmidt and Hunter (1977) to create the validity generalization (VG) paradigm as a means of integrating findings across multiple studies. Central to VG are steps designed to overcome problems commonly encountered in criterion-related studies.

A first problem results from unreliability. If we could measure, say, mechanical aptitude and job performance perfectly for a sample

of mechanical maintenance workers, we would obtain a higher test-performance correlation than we ordinarily find when mechanical aptitude and job performance are measured with error. Let τ_X denote a true score on a test X and τ_Y denote the true score on a job performance measure Y. Then the correlation $r(\tau_X, \tau_Y)$ between true scores τ_X and τ_Y is related to the correlation $r(X,Y)$ between observed scores X and Y by the equation

$$r(X,Y) = r(\tau_X, \tau_Y)\sqrt{r_{XX} r_{YY}}$$

where r_{XX} and r_{YY} are the reliabilities of X and Y. Consequently, unreliability in X and Y attenuates the correlation. For example, if the correlation between true scores for a test and job performance was .50, the reliability of the test was .81, and the reliability of the job performance measure was .49, then the correlation between the two observed measures would be $.50 \times \sqrt{.49 \times .81} = .315$.

The VG analysis corrects for unreliability in the criterion measure, but not unreliability in the test because unreliability in the test degrades the quality of selection decisions. Therefore, Schmidt and Hunter (1977) did not correct for less-than-perfect r_{XX}. In evaluating a test, it is very important to remember that test reliability affects the quality of decisions, and consequently reliability should be of the magnitude previously described (and computed from a relevant sample).

Another problem encountered in criterion-related validity studies results from restriction of range. Organizations obviously prefer applicants who do well on selection tests to applicants who do poorly, and thus the full range of scores on the selection test is not typically observed for the group that is hired and therefore has criterion data available. The significance of this problem is illustrated by a study of U.S. Army Air Force pilot trainees conducted during World War II (Thorndike, 1949). A group of 1,036 completed pilot training because of the need for pilots during the war; only 136 would have been selected for training on the basis of a composite selection index under normal conditions. The correlation of the selection composite with performance in training was .64 for the total group, but only .18 for the 136 individuals with the highest composite scores.

Lord and Novick (1968, p. 143) provide the rather complicated formula that can be used to correct an observed correlation for range restriction. In evaluating the criterion-related validity evidence of a test, it is very important to know whether correlations have been corrected for range restriction and unreliability in the criterion. If such correlations have not been corrected, the validity evidence is biased (and often substantially biased) in the direction of no relationship.

As suggested in the section on test development, the use of small samples has a strong impact on the sampling error of validity coefficients. This variability led many to believe that the relationship between a test and job performance was context specific and that local validity studies were required to justify the use of predictors in each organization. VG research has shown that this belief is clearly false. Nonetheless, any observed correlation of a test with job performance based on a sample of less than several hundred is highly suspect: sample error has an inescapable effect. VG is the best solution to this problem, but an adequate number of studies needs to be included in the analysis. For example, in their original paper, Schmidt and Hunter (1977) used data from $k = 114$ studies to demonstrate that tests of mechanical principles were highly effective in predicting the job performance of mechanical repairmen. It is very important to note that results of a VG analysis of $k = 4$ or 5 studies should be given little credibility.

Another issue to consider when interpreting the results of any criterion-related validity study or VG study concerns the conditions under which the data were collected. Were they collected in an actual operational setting where applicants knew that their scores would affect the likelihood of their being hired? Or were they told that scores were collected "for research purposes only"? Results from studies that ask participants to participate for research purposes may not generalize to the operational context of an assessment. A particularly striking example was provided by White, Young, Hunter, and Rumsey (2008). These authors described the substantial differences between validity coefficients obtained from a large military concurrent validation study using job incumbents (people who were told their participation was for research

purposes only) and longitudinal research on applicant samples responding to operational exams. Although the concurrent validation study showed that socially desirable responding did not affect the validity of a personality measure, validity was severely attenuated in the operational sample. The lesson learned from White et al. is that generalizing research findings to operational contexts is difficult, particularly for measures where test takers can deliberately manipulate their scores by "faking good."

Although test-criterion relationships are typically reported in terms of correlations, other approaches are possible. Utility analysis, for example, links the validity of an assessment to its impact on performance. Taylor and Russell (1939) defined *utility* as the increased accuracy in the prediction of a dichotomous job performance measure (one that classifies people into "successful" and "unsuccessful" categories, for example) obtained from using a particular selection measure. Their conceptualization incorporates the correlation between a test and job performance, the selection ratio (the proportion of applicants who are hired), and the base rate of successful employees (the proportion of new employees who would be successful if the test were not used). The Taylor-Russell tables provide the improvement in success rate (the proportion of selected applicants who are subsequently determined to be successful) that can result from various combinations of these factors.

Cascio (1991) described several disadvantages of this conceptualization of utility. First, the Taylor-Russell tables assess utility only relative to the success rate rather than a monetary outcome. Second, this approach defines success as a dichotomous variable and therefore does not quantify the magnitude of success. In other words, the dichotomous success variable may underestimate the true utility of a measure.

Another popular method of assessing utility was developed by Naylor and Shine (1965). These authors conceptualized utility in terms of the increase in the average job performance score that can be obtained by using an assessment. The disadvantages of this method are that it does not account for the administration costs of the assessments and does not reflect the economic impact of using a particular predictor (Cascio, 1991).

Brogden (1949) and Cronbach and Gleser (1965) proposed a utility estimate that is assessed as the dollar value of work output

rather than the expected improvement in job performance. This method defines the net dollar gain, ΔU, as

$$\Delta U = N \times \bar{T} \times SD_Y \times r_{XY} \times \bar{Z}_X - C$$

where N is the number of people hired in a year, $\bar{T}$ is the average tenure for new employees, SD_Y is the standard deviation of job performance expressed in dollars, r_{XY} is the validity of the test, $\bar{Z}_X$ is the average standardized score of the selected applicants on the test, and C is the total cost of administrating the test.

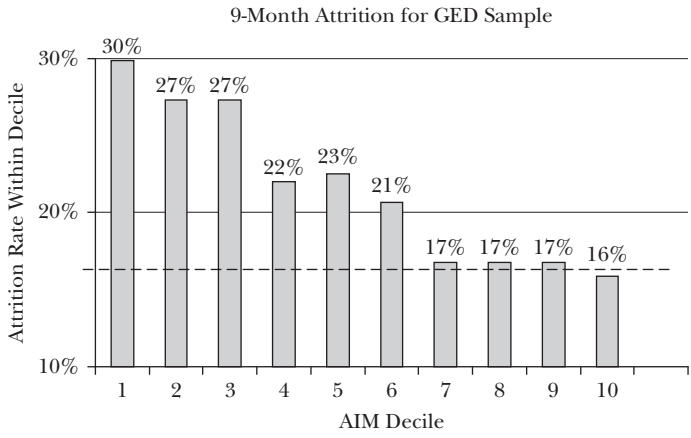
Although the above equation is widely known, it appears to have had limited impact. For example, in the late 1980s, military researchers (Automated Sciences Group & CACI, 1988) conducted this type of utility analysis in an attempt to justify the implementation of the computer-adaptive version of the ASVAB, known as CAT-ASVAB. Military leaders were not impressed with "utility dollars" and did not order implementation of CAT-ASVAB. A few years later, a financial analysis was conducted (Hogan, McBride, & Curran, 1995) that compared the total cost (in actual dollars, not utility dollars) of continuing to administer the paper-and-pencil ASVAB in the Military Enlistment Processing Stations (MEPS) versus the cost of buying computers and implementing CAT-ASVAB in the MEPS. It turned out that the Department of Defense could save millions of dollars per year by implementing CAT-ASVAB, and consequently military leaders decided to implement it. (The principal source of savings was reduced processing time, so fewer applicants needed hotel accommodations for an overnight stay.) In sum, actual dollars were compelling to the Department of Defense leadership, but utility dollars were not.

Finally, another method of characterizing test-criterion relationships has been figuratively called a return-on-investment (ROI) study. Here the organization assesses some important aspect of job performance and then compares this type of performance for people who are high, medium, and low on the selection measure. For example, the U.S. Army is very concerned about attrition during the first term of enlistment, particularly for individuals who do not have a high school diploma. Young, Heggestad, Rumsey, and White (2000) developed the Assessment of Individual Motivation (AIM) to identify military applicants who are likely to complete

their first term. The AIM was administered to a sample of 11,848 GED holders, scores on the AIM composite were stratified into ten decile groups, and attrition rates were computed after nine months in the Army. Figure 2.5 shows that accessions in the top 40 percent on the AIM composite had attrition rates very close to the rate of high school diploma holders (this latter group had a 16.3 percent attrition rate), but attrition rates were much higher for individuals with lower composite scores.

The ROI plot clearly shows the value of the AIM composite; this approach appears to be gaining traction with organizational leaders. Ironically, the correlation of the AIM composite with nine-month attrition was a seemingly trivial −.114. Although statistically significant (the sample size was over eleven thousand), a correlation of −.114 would be likely to elicit the reaction that the selection tool accounted for just 1 percent of the variance in attrition and was therefore virtually useless. Figure 2.5 demonstrates that individuals with as high as 30 percent attrition rates can be screened out, and individuals with a 16 or 17 percent attrition rate screened in. Obviously this is not a trivial difference, and it

Figure 2.5. ROI Plot Depicting Attrition Rates Across Levels of the Army's AIM Composite



Note: We thank Dr. Leonard White and Dr. Mark Young for this figure.

clearly demonstrates that the AIM can make an important contribution to enlistment screening.

Operational Models for Assessment

The proliferation of technology in the workplace has allowed organizations to increase efficiency at lower costs. Similarly, new test systems have also been developed to tap advantages of technological advances. Chief among them are computerized tests delivered over the Internet or an intranet. Such tests can be administered either adaptively (CAT) or nonadaptively (with computerized page turners); they may also be administered in proctored or unproctored settings.

To help organizations decide whether computerized testing adds value or whether unproctored testing should be implemented, we discuss the advantages and disadvantages of these delivery options. Next, we examine the related issue of test security. In the final subsection, we review best practices for test score reporting.

Computer-Administered Tests

Compared to paper-and-pencil tests, computerized tests can be easier to administer and score, and potentially they have lower costs. Because computerized tests usually provide several screens of instructions and are self-timed, they may require fewer proctors, which lowers cost. Furthermore, examinee responses are recorded automatically, allowing instantaneous scoring with minimal error. Tests can also be updated much more easily. For example, the system stores the operational test solely on a central server; when an examinee begins a test session, the exam is downloaded to his or her personal computer. In this situation, any change in the test can be made easily and implemented instantly. In contrast, months may be required to print and distribute a revised paper-and-pencil form. For all of these reasons, computerized tests are increasing in popularity.

Another advantage of computer-administered tests is that multimedia capabilities can be used. For example, a situational judgment test may use video clips depicting common workplace situations instead of text-based descriptions. The value of multimedia

cannot be overstated because research shows that video-based situational judgment tests are less correlated with cognitive ability than paper-and-pencil versions of the same test and therefore yield incremental validity above cognitive ability and less adverse impact (Olson-Buchanan et al., 1998). Multimedia is also particularly important when one attempts to assess other skills, such as negotiation or conflict resolution, not easily tested with a paper-and-pencil format. Moreover, well-constructed, multimedia assessment usually has a more positive response from test takers. Richman, Olson-Buchanan, and Drasgow (2000) found that managers completing a multimedia assessment perceived the assessment as more face valid and had more positive attitudes as compared to managers who were administered nonmultimedia assessments with equivalent content.

Adaptivity, which is conveniently implemented in a computerized setting, has important advantages. Specifically, CAT results in shorter tests with higher measurement precision. The general idea behind CAT is that items are selected adaptively so that their difficulty matches a test taker's ability. The prototypical algorithm involves computing a provisional estimate of ability ($\hat{\theta}$) based on the responses to items previously administered, selecting the next item from the item pool that is maximally informative, and administering this item to the examinee. This process is repeated until the ability estimate is sufficiently precise or a fixed test length is reached. Thus, each test is tailored to the individual test taker. This is unlike conventional tests where all test takers answer the same items regardless of whether they are too hard or too easy, which may result in boredom, random responding, or guessing. Most important, time savings associated with CAT can be converted into cost savings.

There are challenges that organizations must face if they wish to implement a CAT. First, it is costly to implement CAT because developing and pretesting a large item pool is time consuming and expensive. Furthermore, continued technical and psychometric expertise is necessary to support a CAT testing program.

Internet Testing: Unproctored and Proctored Testing

With its exponential growth, companies are turning to the Internet as a medium for testing. In fact, the United States alone has about 223 million Internet users, with another 795 million users