

Technology-Enhanced Assessment of Talent

Join Us at
Josseybass.com



JOSSEY-BASS™
An Imprint of
 **WILEY**

Register at **www.josseybass.com/email**
for more information on our publications,
authors, and to receive special offers.

Technology- Enhanced Assessment of Talent

Nancy T. Tippins
Seymour Adler
Editors

Foreword by
Allen I. Kraut

 **JOSSEY-BASS**
A Wiley Imprint
www.josseybass.com

Copyright © 2011 by John Wiley & Sons, Inc. All rights reserved.

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, or on the Web at www.copyright.com. Requests to the publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, 201-748-6011, fax 201-748-6008, or online at www.wiley.com/go/permissions.

Readers should be aware that Internet Web sites offered as citations and/or sources for further information may have changed or disappeared between the time this was written and when it is read.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Jossey-Bass books and products are available through most bookstores. To contact Jossey-Bass directly call our Customer Care Department within the U.S. at 800-956-7739, outside the U.S. at 317-572-3986, or fax 317-572-4002.

Jossey-Bass also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Technology-enhanced assessment of talent / Nancy T. Tippins, Seymour Adler, editors ; foreword by Allen Kraut.

p. cm. — (J-b siop professional practice series ; 30)

Includes bibliographical references and index.

ISBN 978-0-470-59158-1 (hardback)

1. Personnel management—Technological innovations. 2. Employee selection.
3. Management information systems. 4. Personnel management—Technological innovations—Case studies. 5. Employee selection—Case studies. 6. Management information systems. I. Tippins, Nancy Thomas, 1950- II. Adler, Seymour, 1948-
HF5549.5.T33T43 2011
658.3'125—dc22

2010052163

Printed in the United States of America

FIRST EDITION

HB Printing 10 9 8 7 6 5 4 3 2 1

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

Advancing Executive Coaching: Setting the Course for Successful Leadership Coaching

Gina Hernez-Broome and Lisa A. Boyce, Editors

Handbook of Workplace Assessment: Evidence-Based Practices for Selecting and Developing Organizational Talent

John C. Scott and Douglas H. Heynolds, Editors

Going Global: Practical Applications and Recommendations for HR and OD Professionals in the Global Workplace

Kyle Lundby, Editor

Strategy-Driven Talent Management: A Leadership Imperative

Rob Silzer and Ben E. Dowell, Editors

Performance Management

James W. Smither and Manuel London, Editors

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

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

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 Gilford 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

*For WMT and HHT
To Rivki, for a lifetime of love and support*

Technology-Enhanced Assessment of Talent

Contents

Foreword by Allen I. Kraut, Baruch College/Kraut Associates	xiii
The Editors	xvii
The Contributors	xix
Preface	xxxv
Acknowledgments	xxxvii
1. Overview of Technology-Enhanced Assessments	1
Nancy T. Tippins	
Section One: Measurement and Implementation Issues in Technology-Enhanced Assessments	19
2. Foundations for Measurement	21
John C. Scott and Alan D. Mead	
3. Implementing Assessment Technologies	66
Douglas H. Reynolds	
4. Cheating and Response Distortion on Remotely Delivered Assessments	99
Winfred Arthur, Jr., and Ryan M. Glaze	
5. Computerized Adaptive Testing	153
Rodney A. McCloy and Robert E. Gibby	
6. Applicant Reactions to Technology-Based Selection: What We Know So Far	190
Talya N. Bauer, Donald M. Truxillo, Kyle Mack, and Ana B. Costa	
7. International Issues, Standards, and Guidelines	224
Dave Bartram	
Section Two: Case Studies of Technology-Enhanced Assessments	251
8. Web-Based Management Simulations: Technology- Enhanced Assessment for Executive-Level Selection and Development	253
Terri McNelly, Brian J. Ruggenberg, and Carrol Ray Hall, Jr.	

9.	Bridging the Digital Divide Across a Global Business: Development of a Technology-Enabled Selection System for Low-Literacy Applicants	267
	Adam Malamut, David L. Van Rooy, and Victoria A. Davis	
10.	Promotional Assessment at the FBI: How the Search for a High-Tech Solution Led to a High-Fidelity Low-Tech Simulation	293
	Amy D. Grubb	
11.	Innovation in Senior-Level Assessment and Development: Grab 'Em When and Where You Can	307
	Sandra B. Hartog	
12.	Case Study of Technology-Enhanced Assessment Centers	324
	Rick Hense and Jay Janovics	
13.	Video-Based Testing at U.S. Customs and Border Protection	338
	Jeffrey M. Cucina, Henry H. Busciglio, Patricia Harris Thomas, Norma F. Callen, DeLisa D. Walker, and Rebecca J. Goldenberg Schoepfer	
14.	Going Online with Assessment: Putting the Science of Assessment to the Test of Client Need and 21st Century Technologies	355
	Eugene Burke, John Mahoney-Phillips, Wendy Bowler, and Kate Downey	
15.	Implementing Computer Adaptive Tests: Successes and Lessons Learned	380
	Mike Fetzer and Tracy Kantrowitz	
16.	Practice Agenda: Innovative Uses of Technology-Enhanced Assessment	394
	Michael J. Zickar and Christopher J. Lake	
17.	Concluding Comments: Open Questions	418
	Seymour Adler	
Indexes		
	Name Index	437
	Subject Index	445

Foreword

Who would have thought that one of the most important “laws” affecting personnel assessment would be Moore’s Law? About forty-five years ago, Gordon Moore, a co-founder of Intel, described his expectation that the cost of computer chips would be cut in half about every two years for the foreseeable future. This proposition, often described as Moore’s Law, turned out to be largely true, cutting the cost of computing and related processes by half every two years since then. These drops in cost have made computing power cheap and easily available in many devices such as personal computers, laptops, and cell phones.

Together with other advances in technology, such as fiber optics and space satellites, a vast array of devices are now available for personnel assessment and selection purposes that could not have been imagined just two decades earlier. Younger practitioners in human resources and industrial/organizational psychology must be forgiven if they think that such devices and applications were always with us. In fact, the computing power of a BlackBerry cell phone today exceeds the power of the computer used to guide the first spaceship to land on the moon.

Technology and what it allows us to do in the field of assessment have unfolded incredibly fast. The classic *Personnel Testing* by Robert M. Guion (1965) does not even contain the word “computer” in its index. Catching up, Guion and Highhouse’s 2006 book on the same subject does have “computerized testing” in the subject index, showing it was mentioned on four pages.

One of the earliest recognitions of the potential use of computers and allied devices in psychological assessment appeared in 1998 in a book chapter by Rob Silzer and Richard Jeanneret, two highly experienced and creative practitioners. They foresaw “the broader use of computers and information technology in the assessment process. . . . it will also be extended to simulations and exercises and to multi-rater questionnaires.” Hesitantly, they noted “one

exception might be the assessment interview.” Confidently though, they concluded, “This is the future of assessment design.” (1998, p. 455). Even their optimistic predictions have been far exceeded, as this volume demonstrates so well.

Nancy Tippins and Seymour Adler have gathered a stellar group of practitioners, showing how technology has enhanced assessment in many different applications. The technologies we see used now range from personal computers to cell phone and landline telephones, to DVDs and other video formats. The applications include hiring and selection interviews, assessments for promotion, professional skills, and management training, and even certification examinations. Current assessments are a lot more realistic, diverse in content, and varied in application than ever before.

Of course, the most basic questions of test reliability and validity call out. Are the scores from traditional paper-and-pencil assessments comparable to those taken with the use of more sophisticated technology? Will people taking unproctored exams do better than people in proctored settings? In fact, what is the likelihood for greater or lesser cheating? Will an individual’s performance be impacted by nervousness with a new technology? These and many other issues are dealt with in various chapters of this book by extremely knowledgeable practitioners and academics.

The chapter authors come from a wide variety of work settings, although many share the fact of being consultants and in-house practitioners. Most of these are associated with large, often global, firms. It is interesting to note that current technology permits, and may even encourage, a global reach. And, of course, large companies, especially if they are successful and forward-thinking, are the most likely to apply technology to their assessment and development needs.

This dynamic means that the contributors in this volume are at the forefront of assessment practice using the latest technology. These are the professionals who are developing and researching the most significant advances in this field. What they say in their chapters shows that they are among the most farsighted professionals in terms of anticipating and solving possible problems in the use of technology for assessment. While they see the prospective enhancements of technology, they also see the potential problems, and in many cases have developed solutions that others will follow.

But the use of these applications by large, successful, and global companies raises other problems. It challenges us to think about the cultural differences with which such new applications are greeted or interpreted. Do they favor candidates from more developed countries who are used to technology? Do they permit comparability of norms from country to country? Will they make a positive or negative impression on candidates, especially those who are not selected?

Even within any one country, is performance with technology-enhanced assessment influenced by the individuals' socioeconomic status and familiarity with technology? Will it unfairly favor candidates who are well off? Will it have a disparate impact on groups protected from discrimination by law?

Luckily for us, the contributors to this volume take on all these questions and more. Their answers are based on solid experience, thoughtfulness, and considerable research. Future developments will continue to be driven by Moore's Law. Applications of sophisticated technology will continue to become cheaper. Right now, large organizations can most afford and most benefit from such applications. They have the economies of scale that make up-front investments possible and attractive.

But we can expect technology-enhanced assessment to spread far and wide as costs come down even further. Just as important as lower costs is the potential for imaginative and more sophisticated applications of such technology. Increasing evidence of validity and improved organizational performance will fuel the demand for more such applications.

And because we can expect many more applications of technology-enhanced assessment, we are especially fortunate to have such a wonderful set of wise and experienced practitioners to share their knowledge with us. Anyone who wishes to use and really understand technology-enhanced assessment will be enriched by this terrific book. As series editor, I very much appreciate the work of Nancy Tippins, Seymour Adler, and all of their gifted colleagues who have contributed to this latest Professional Practices volume. We are all indebted to them.

Allen I. Kraut
Rye, New York
January 2011

References

- Guion, R. M. (1965). *Personnel testing*. New York, McGraw-Hill.
- Guion, R. M., & Highhouse, S. (2006). *Essentials of personnel assessment and selection*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Silzer, R., & Jeanneret, R. (1998). Anticipating the future: Assessment strategies for tomorrow. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment: Predicting behavior in organizational settings*. San Francisco: Jossey-Bass.

The Editors

Nancy T. Tippins, Ph.D., 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, Dr. Tippins 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 SIOP and a Fellow of SIOP, the American Psychological Association, and the Association for Psychological Science.

She served on the Ad Hoc Committee on the Revision of the Principles for the Validation and Use of Personnel Selection Procedures (2003) and currently sits on the committee to revise the Standards for Educational and Psychological Testing. She has served as the associate editor for the Scientist-Practitioner Forum of *Personnel Psychology* and a member of the editorial boards for *Personnel Psychology* and the *Journal of Applied Psychology*. She has published numerous papers on tests and assessments and unproctored Internet testing. Most recently, she co-edited the *Handbook of Employee Selection* and co-authored *Designing and Implementing Global Selection Systems*.

Dr. Tippins received M.S. and Ph.D. degrees in industrial and organizational psychology from the Georgia Institute of Technology.

Seymour Adler, Ph.D. is senior vice president in Aon Hewitt's Talent Consulting practice. He is based in New York and consults with client organizations throughout the country and with Global 500 organizations. He has served as co-practice leader for the Talent Solutions practice at Aon as well as head of the Human Capital Global Practice Council.

Dr. Adler directs the development and implementation of talent assessment, talent management, and leadership development programs, with an emphasis on sales, customer service, and management positions in the financial services, telecommunications, and high technology industries, as well as in the public sector.

He was a founder and principal of Assessment Solutions Incorporated, a firm he helped take public in 1997, which was acquired by Aon in 2001. In addition to a thirty-five-year career as a practitioner, Dr. Adler has taught in graduate programs at Tel Aviv University, Purdue University, Stevens Institute of Technology, New York University, and currently is an adjunct professor at Hofstra University's doctoral program in applied organizational psychology.

A graduate of the doctoral program in industrial/organizational psychology at New York University, he is a Fellow of the Society of Industrial/Organizational Psychology, has served as president of the Metropolitan New York Association of Applied Psychology, and has contributed to both the scientific and practitioner literatures in industrial/organizational psychology.

The Contributors

Winfred Arthur, Jr., Ph.D., is a full professor of psychology and management at Texas A&M University. He is a Fellow of the Society for Industrial and Organizational Psychology, the Association of Psychological Science, and the American Psychological Association. His research interests are in human performance; training development, design, implementation, and evaluation; team selection and training; acquisition and retention of complex skills; testing, selection, and validation; models of job performance; personnel psychology; and meta-analysis. He received his Ph.D. in industrial/organizational psychology from the University of Akron in 1988.

Dave Bartram, Ph.D., is chief psychologist for SHL Group Ltd. Prior to that he was dean of the faculty of science and the environment and professor of psychology in the Department of Psychology at the University of Hull. He is a Fellow of the British Psychological Society and a Fellow of the International Association of Applied Psychology (IAAP) and past president of IAAP Division 2 (Assessment and Evaluation). He is past president and current secretary of the ITC. He is the convener of the European Federation of Psychologists' Association's Standing Committee on Tests and Testing and a member and a past chair of the BPS's Steering Committee on Test Standards. He was appointed Special Professor in Occupational Psychology and Measurement at the University of Nottingham, UK, in 2007 and Professor Extraordinarius in the Department of Human Resource Management at the University of Pretoria, South Africa in 2010. Within SHL he led such developments as the Universal Competency Framework, the SHL Corporate Leadership Model, and the IRT-based OPQ32r.

Talya N. Bauer, Ph.D., received her doctorate from Purdue University and is the Cameron Professor of Management at Portland State University. Dr. Bauer is an award-winning teacher and researcher. She conducts research about relationships at work. More specifically, she works in the areas of new hire on-boarding, recruitment, selection, over-qualification, mentoring, and leadership, which have resulted in numerous journal publications published in outlets such as the *Academy of Management Journal*, *Academy of Learning and Education Journal*, *Journal of Applied Psychology*, *Journal of Management*, and *Personnel Psychology*. She has acted as a consultant for dozens of government, Fortune 1000, and start-up organizations. Dr. Bauer is involved in professional organizations and conferences at the national level, such as serving on elected positions such as the Human Resource Management Executive Committee of the Academy of Management and member at large for SIOP. Dr. Bauer is currently editor of the *Journal of Management*. In addition, she has also served on the editorial boards for the *Journal of Applied Psychology*, *Personnel Psychology*, and *Journal of Management*. In 2009, she published three textbooks: *Organizational Behavior* (with Berrin Erdogan), *Principles of Management* (with Mason Carpenter and Berrin Erdogan), and a management graphic novel entitled *Atlas Black: Managing to Succeed* (with Jeremy Short, Dave Ketchen, and illustrator Len Simon). In 2010, the team published *Atlas Black: Management Guru*. Her work has been discussed in several media outlets including *The New York Times*, *BusinessWeek*, *The Wall Street Journal*, *Oregonian*, *Portland Business Journal*, NPR's *All Things Considered*, and KGW News.

Wendy Bowler is a senior talent and resourcing manager at HSBC, where she has managed the deployment of online assessment solutions for the recruitment and selection of customer service agents as well as the alignment of competency models and associated assessments with industry best practices for the provision of banking services. She has held roles in the UK and Hong Kong and is a member of the British Psychological Society's Division of Occupational Psychology.

Eugene Burke is chief scientist at SHL Group Ltd. His experience in computer-based testing (CBT) extends back to the 1980s and the development of solutions combining models from experimental and cognitive psychology and psychometrics for the selection of pilots and aircrew for the Royal Air Force and United States Air Force. More recently, he has led on the verify solution to unproctored Internet testing (UIT) as well as several initiatives for secure delivery of CBT and Internet assessments. He is a past chair of the British Psychological Society's Steering Committee on Test Standards, where he led on guidelines for computer-based assessments, and has held positions as chair of the BPS Division of Occupational Psychology, council member of the International Test Commission (for which he is currently collaborating to develop an ITC guideline for test security), chair of the European Association of Test Publishers, and a member of the International Standards Organization (ISO) working group for an international standard for the use of assessment data in personnel decisions. He has authored several journal articles, book chapters, and best practice white papers in the areas of CBT, UIT, and secure assessment and has been a regular presenter at SIOP conferences on UIT and innovations in the use of technology in constructing and deploying assessments.

Henry Busciglio, Ph.D., is a senior personnel research psychologist with the U.S. Bureau of Customs and Border Protection (CBP), Department of Homeland Security. Henry received his Ph.D. in industrial/organizational psychology from the University of South Florida. Before coming to CBP, Henry's federal service began at the Army Research Institute and continued at the Office of Personnel Management. His research efforts at CBP have included selection and promotional assessments for CBP officers, Assessment Centers for Management and Senior Executive Development, and a long-term, multi-phased evaluation of the Quality Recruitment Program.

Norma F. Callen is a senior research psychologist with the Personnel Research and Assessment Division in the Office of Human Resources Management at U.S. Customs and Border Protection

(CBP). Her primary responsibility is the development and implementation of a variety of assessments for selection and promotion into CBP and U.S. Immigration and Custom Enforcement (ICE) mission-critical occupations. Her specialties lie in the area of structured interview development, structured interview administration training, in-basket assessments, online testing, and video-based testing. Ms. Callen has a master's degree in industrial/organizational psychology from George Mason University.

Ana B. Costa is a doctoral student in I/O psychology at Portland State University. She received a master's degree in I/O psychology from California State University, Sacramento, and consulted for several years prior to returning to school to obtain her doctorate. Her research interests include applicant reactions, training development, and evaluation, as well as wellness, prevention, and safety for both businesses and their employees.

Jeffrey M. Cucina, Ph.D., is a senior personnel research psychologist at U.S. Customs and Border Protection (CBP), Department of Homeland Security (DHS), where he works on the development and validation of entry-level and promotional assessments. He serves as the project director for CBP's video-based test and is actively involved in its development and implementation. Recently, Dr. Cucina worked on the development of four VBT versions for agriculture specialists at CBP. His VBT implementation responsibilities include maintaining the day-to-day aspects of the VBT program, providing training to test administrators and raters, and providing guidance to test administrators and raters in the field. In 2007, he was a co-recipient of the IPMAAC Innovations in Assessment Award for CBP's VBT. At CBP, he has also developed job knowledge tests, conducted survey research, performed job analyses, and conducted empirical research in support of CBP's testing program. He received a Ph.D. in industrial/organizational psychology from the George Washington University.

Victoria A. Davis, Ph.D., is manager, Talent Management Analytics and Solutions, at Marriott International. She serves as program

and research manager for the development, validation, and implementation of global selection and performance management systems. In addition, her role includes promoting program sustainability and conducting ongoing job analytics. In support of the global hourly selection initiative, she managed the day-to-day activities and collaborated on the design and implementation of the program. Prior to joining Marriott International, she partnered with The Home Depot's Organizational Effectiveness/Talent Management team to streamline hiring processes by designing and validating selection tools, conducting executive and organizational assessments, as well as facilitating leadership development and program evaluations. She was previously employed by Interactive, Inc., where her focus was on conducting qualitative and quantitative analyses (surveys, interviews, and achievement data) and evaluating the return on investment and achievement gains for educational technology programs. Dr. Davis received her doctorate degree in industrial and organizational psychology from Alliant International University in San Diego, California.

Kate Downey is an occupational psychologist at HSBC UK, where she is program director for two strategic leadership development initiatives aimed at developing senior management and future talent. Kate is also involved in designing global learning programs for the organization. Kate is a graduate of the University of Birmingham and Manchester Business School and a Chartered Occupational Psychologist with the British Psychological Society. Prior to joining HSBC, Kate was a senior consultant in the SHL Science and Innovation Group, where she designed online product and client solutions.

Mike Fetzer, Ph.D., is the global director of Advanced Assessment Technologies at SHLPreVisor, a leading provider of pre-employment assessments and employee selection solutions. In this role, he is responsible for the development of leading-edge assessments that are implemented on a global scale. SHLPreVisor's comprehensive library of more than one thousand assessments include a broad spectrum of personality, cognitive ability, skills, biodata, situational judgment, computer adaptive (CAT), and

simulation- and multimedia-based assessments used to measure talent in all jobs for all industries. Before joining SHLPreVisor, he championed the global testing initiatives at Development Dimensions International (DDI). Dr. Fetzner holds a doctoral degree in industrial/organizational psychology and is a member of the Society for Industrial and Organizational Psychology, the International Test Commission, the International Association of Applied Psychology, and the American Psychological Association.

Robert E. Gibby, Ph.D., is senior manager of human resources research and analytics for Procter & Gamble, headquartered in Cincinnati, Ohio. In this role, Dr. Gibby leads a team of industrial/organizational psychologists and HR professionals to deliver HR analytics, external selection and assessment, and the annual engagement survey for the company. He also has responsibility for developing and managing relationships with industrial/organizational partners in industry and academia and for consulting internally on HR systems and data needs. Dr. Gibby joined P&G in 2004 and completed his Ph.D. in industrial/organizational psychology from Bowling Green State University the same year. Outside P&G, he serves as a board member for Northern Kentucky University's Master's of Industrial/Organizational Psychology Program, where he has taught undergraduate and graduate courses. He is a member of the Society for Industrial and Organizational Psychology, an editorial board member of the *Journal of Personnel Psychology*, and actively contributes to the I/O field through publications, presentations, and speaking engagements.

Ryan M. Glaze is a graduate student at Texas A&M University. His research interests include personnel selection, testing and validation, human performance, team selection and training, complex skill acquisition, retention, and pre-training factors. He received his master's in industrial/organizational psychology from Texas A&M University in 2009.

Amy D. Grubb, Ph.D., is the senior industrial/organizational psychologist at the FBI, reporting directly to the executive assistant

director (EAD) of the Human Resources Branch (HRB). In her eleven years with the FBI, she has been responsible for the development and validation of selection and promotion systems at the FBI, as well as multiple organization development initiatives. In addition to the work with promotion assessments and processes at the FBI, she instituted the FBI Annual Employee Survey, including its integration with organizational processes and programs, and serves as an internal executive coach. She has worked extensively with nearly all populations within the FBI, from special agents to pilots to Hostage Rescue Team members to intelligence analysts to mid-level and executive populations within both the special agent and professional staff ranks. In addition, she serves on the FBI's Institutional Review Board (Human Subjects) and advises the executive leaders on risk from the human capital perspective regarding organizational performance, policy decisions, and change initiatives. She has presented numerous papers at professional conferences, has served on advisory boards both internal and external to the FBI, and liaisons extensively within the government and the public sector communities. Prior to the FBI, she worked as a consultant with the Vandaveer Group in Houston, Texas. An FBI Director's Award for Excellence recipient, Dr. Grubb earned her Ph.D. and M.A. in industrial/organizational psychology from the University of Houston, earning her bachelor's degree from Villanova University.

Carrol Ray “Buddy” Hall, Jr., is the vice president of human resources for Darden's Specialty Restaurant Group. He has twenty-three years of human resources experience across a broad spectrum of manufacturing and service environments. Prior to his current role, Hall was the director of talent assessment for Darden. In that capacity, he led the architecture of a behaviorally based assessment system used in critical jobs across the organization. Hall is an accomplished public speaker and classically trained pianist. He lives in Orlando with his wife and two children. He received his undergraduate degree from the University of South Alabama and is a certified professional in human resources (PHR).

Sandra B. Hartog, Ph.D., is the president and founder of Sandra Hartog & Associates, a boutique talent management consulting

firm with a global client base. She is also the CEO and founder of Fenestra, Inc., which houses SH&A's innovative virtual assessment center platform. Dr. Hartog has more than twenty-five years of experience as a consultant to Fortune 500 companies and other organizations. She has extensive experience in the design and implementation of virtual assessment centers for selection and development purposes. She also does work in succession management, leadership competency studies, enhanced 360-degree feedback design, assessment and feedback delivery, and executive coaching. Dr. Hartog holds a Ph.D. in industrial/organizational psychology from the City University of New York Graduate Center. She teaches at the graduate level in the U.S. and internationally.

Rick Hense, Ph.D., is VP of talent selection and assessment at Bank of America, where he leads initiatives to develop, validate, and implement cost-effective systems for selecting high-quality associates. Since joining Bank of America in 2005, Dr. Hense is responsible for selection consulting, conducting job/competency analysis, developing assessment solutions for selection and leadership development, and measuring return on investment. Prior to joining Bank of America, he was an organizational effectiveness manager at Raymond James and a selection manager at Capital One. Dr. Hense earned his Ph.D. in industrial/organizational psychology from the University of South Florida. He has presented papers and symposia at multiple conferences and published several journal articles.

Jay Janovics, Ph.D., is the director of optimization services on the Professional Services team at SHLPreVisor. In this role he leads a team of industrial/organizational psychologists dedicated to conducting validation and business outcome studies to demonstrate the business impact of SHLPreVisor assessment solutions. Since joining SHLPreVisor in 2004, he has worked with a number of Fortune 500 clients in post-implementation analyses of SHLPreVisor's assessment solutions. In this role his focus is on demonstrating and enhancing the assessments' usefulness in predicting critical business outcomes. Dr. Janovics has developed and validated a variety of assessment instruments including measures of cognitive

ability, personality, and biodata, work simulations and video-based situational judgment tests. Prior to joining SHLPreVisor, he was employed as the research director at a small survey research firm and as an internal staffing consultant at a Fortune 100 manufacturing organization. He holds a Ph.D. in industrial/organizational psychology from Central Michigan University.

Tracy Kantrowitz, Ph.D., is the director of research and development at SHLPreVisor. In this role, she is responsible for the development of assessment content and research related to employee selection. Dr. Kantrowitz has published in leading journals and presented at national conferences on topics such as predictors of job performance, computer adaptive testing (CAT), and unproctored Internet testing (UIT). In 2010, she was a member of the team awarded the M. Scott Myers Award for applied research in the workplace from the Society for Industrial and Organizational Psychology for work on computer adaptive personality testing. Dr. Kantrowitz holds a Ph.D. in industrial/organizational psychology from the Georgia Institute of Technology. She is a member of the American Psychological Association and Society for Industrial and Organizational Psychology.

Allen I. Kraut, series editor for the SIOP Professional Practices Series, is Professor Emeritus of Management at Baruch College, City University of New York, which he joined in 1989. For much of his professional career, he worked at the IBM Corporation, where he held managerial posts in personnel research and management development before leaving in 1989. In 1995, he received the SIOP's Distinguished Professional Contributions Award, recognizing his work in advancing the usefulness of organizational surveys. In 1996, Jossey-Bass published his book, *Organizational Surveys: Tools for Assessment and Change*. His latest book, *Getting Action from Organizational Surveys: New Concepts, Technologies, and Applications*, was published by Jossey-Bass in 2006.

Christopher J. Lake is currently a doctoral student in industrial/organizational psychology at Bowling Green State University. His research focuses on work-related testing, measurement, and

methodological issues. He is particularly interested in using psychometric techniques to gain an increased understanding of workplace attitudes and organizational turnover.

Kyle Mack is a doctoral student in I/O psychology at Portland State University. He holds masters' degrees in literature from the University of California, Santa Cruz, and I/O psychology from Portland State University. His research interests include applicant reactions and motivation, person-environment fit, and attributions in performance situations. He believes that research and practice should inform and complement each other and has a wide range of practical business experience, including product management, business development, and consulting experience.

Dr. John Mahoney-Phillips is global head of Human Capital at UBS AG, where he leads the team responsible for performance appraisals across the whole company and for the most senior teams, employee surveys with a special focus and expertise on engagement, and for selection and assessment testing and HR metrics. He joined UBS AG in the Private Banking Division as head of Leadership and Management Development and then moved into a corporate role, setting up and leading the UBS human capital function. Prior to joining UBS, John was global head of international projects for SHL, leading major client talent acquisition, development, and succession projects. John's interests inside work are in leadership succession and assessment, performance management and culture, employee engagement, and talent identification. He is a Chartered Occupational Psychologist, Chartered Scientist, Associate Fellow of the British Psychological Society and honorary researcher at the School of Psychology of the University of East London.

Adam Malamut, Ph.D., serves in two leadership roles at Marriott International. As global HR officer—Information Resources, Dr. Malamut serves as the senior human resources leader for the Information Resources (Technology) division. In this role, he is responsible for establishing the global human resources strategy for Information Resources, including design, development, and

deployment for talent acquisition, leadership and career development, performance management, and workforce analytics. He is also responsible for strategy and program integration across multiple areas of talent management for all global brands and businesses: assessment and selection systems, succession planning, leadership development, performance management, workforce analytics and program evaluation, and engagement surveys and research. His work on organizational climate, diversity and inclusion, and employee selection and assessment systems has been presented at numerous professional conferences and published in peer-reviewed publications. Dr. Malamut's applied research on workforce diversity and inclusion was recognized and supported by a multi-year grant from the National Science Foundation. He has a doctorate in industrial/organizational psychology from The George Washington University and a B.S. in psychology from Penn State University. He is also a member of the American Psychological Association and the Society for Industrial and Organizational Psychology.

Rodney A. McCloy, Ph.D., 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 (structural equation modeling, event history analysis, hierarchical linear modeling) and has applied them to numerous research questions, particularly those involving personnel selection and classification, job performance measurement and modeling, and attrition/turnover. His assessment and testing experience has spanned both cognitive and non-cognitive domains and has involved several large-scale assessment programs (Armed Services Vocational Aptitude Battery, National Assessment of Educational Progress, General Aptitude Test Battery). His recent research includes the development of computerized adaptive tests of cognitive ability for use as selection screens in unsupervised settings. He is a Fellow of the Society for Industrial and Organizational Psychology and the American Psychological Association. He received his B.S. in psychology from Duke University and his Ph.D. in industrial/organizational psychology from the University of Minnesota.

Terri L. McNelly is a senior manager for Aon Hewitt. She consults in the areas of competency modeling and development, employee and managerial selection and assessment, training and instructional design, and leadership assessment and development. Her work is focused on improving the performance of organizations through selection, training, and development of people. She has experience working in a variety of private- and public-sector organizations, including textiles, chemical processing, petroleum, manufacturing, automotive finance, quick service and casual dining restaurants, and telecommunications. McNelly holds a B.S. and M.S. and is ABD for her Ph.D., all from Texas A&M University. She has presented at several professional conferences, published in professional journals, and is active in the American Psychological Association and Society for Industrial/Organizational Psychology.

Alan D. Mead, Ph.D., is an assistant professor of psychology at the Illinois Institute of Technology. He received his Ph.D. from the University of Illinois-Urbana. Prior to becoming an academic, Dr. Mead was a practitioner for over a dozen years, working as a consultant, research scientist, and psychometrician. Dr. Mead teaches psychometric theory, individual differences, compensation, and quantitative topics such as validity generalization, utility theory, and synthetic validity. Dr. Mead's research interests include psychometric applications (such as differential item functioning), computerized testing, and applications of personality theory. He has been published in journals such as *Psychological Bulletin*, *Personnel Psychology*, and *Applied Psychological Measurement* and he is a frequent participant at the Society for Industrial and Organizational Psychology annual meeting.

Douglas H. Reynolds, Ph.D., is vice president of assessment technology at Development Dimensions International, where his department develops and implements new assessment and testing products in Fortune 500 companies. Dr. Reynolds has focused his research on the use of Internet technologies for the delivery of assessments in the workplace, and his products range from large-scale hiring systems to computer-delivered simulations for 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 Dr. Reynolds co-edited the *Handbook of Workplace Assessment*, a volume in SIOP's Professional Practice Series, and co-authored *Online Recruiting and Selection*, a book on the integration of technology with personnel selection practices. Dr. Reynolds is active in SIOP leadership, serving on the board as communications officer and president for the 2012–2013 term. He earned his Ph.D. in industrial/organizational psychology from Colorado State University.

Brian J. Ruggeberg, Ph.D., is currently senior vice president with Aon Hewitt and serves as co-leader of the National Staffing and Development function for the overall Talent Practice. Prior to the Aon Hewitt merger, Dr. Ruggeberg served as the Northeast Regional Practice leader for the overall Human Capital practice of Aon Consulting. Dr. Ruggeberg has over twenty years of consulting experience and has been involved in the development and implementation of various assessment and selection, leadership assessment and development, performance management, and training programs. He has experience in several different industry sectors, including healthcare, pharmaceutical, government, technology, telecom, finance, insurance, manufacturing, hospitality/restaurant, education, and utilities. Prior to joining Aon in 1996, Dr. Ruggeberg served as an independent consultant and worked with organizations such as Bell Atlantic; BellSouth; Canon Virginia, Inc.; Norfolk Southern Corp.; and the Office of Personnel Management. Dr. Ruggeberg holds a B.A. in psychology from Cornell College and an M.S. in psychology and a Ph.D. in industrial/organizational psychology from Old Dominion University. Dr. Ruggeberg is an active member of the American Psychological Association and the Society for Industrial and Organization Psychology and was the 2002–2003 president of the Metropolitan New York Association of Applied Psychology.

Rebecca J. Goldenberg Schoepfer, Ph.D., is currently the director of Talent Management Operations for Novo Nordisk, Inc., the U.S. affiliate of a global diabetes pharmaceutical company. In this role, she is responsible for operational components of talent management, including the functions that support training excellence.

These include the measures, metrics, and design, learning technology, training logistics, and talent management budget teams. Prior to working at Novo Nordisk, Inc., Dr. Schoepfer served as a project manager with APT, Inc., a human resources consulting firm, where she provided consulting services in the design and validation of employee selection procedures, training design and delivery, performance management, downsizing, and competency modeling. Earlier in her career, she worked for U.S. Customs and Border Protection as a personnel research psychologist. She designed, implemented, and administered paper-based and multimedia assessment tools and programs for specific mission-critical occupations. She also conducted research studies and statistical analyses to evaluate organizational assessments for compliance with organization, professional, and legal standards. Dr. Schoepfer received her Ph.D. in industrial/organizational psychology from The George Washington University and her B.A. in psychology from the University of Michigan. She has also received her SPHR certification.

John C. Scott, Ph.D., is chief operating officer and cofounder of *APTMetrics, 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 APTMetric'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 two books, the *Handbook of Workplace Assessment: Evidence-Based Practices for Selecting and Developing Organizational Talent* and *The Human Resources Program-Evaluation Handbook*. He is 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 and SIOP's Organizational Frontiers Series. Dr. Scott is a Fellow of the Society for Industrial and Organizational

Psychology 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.

Patricia Harris Thomas is director of the Personnel Research and Assessment Division within the Office of Human Resources Management at U.S. Customs and Border Protection (CBP). She manages an experienced staff of industrial/organizational psychologists who are responsible for the development and implementation of assessments for selection and promotion into mission-critical CBP occupations and for research that supports the evaluation of organizational effectiveness within the agency. Through her leadership, her staff was winner of the 2007 IPMAAC Innovations in Assessment Award for its video-based testing program. She received the CBP commissioner's 2009 Top Unit Performance Award and the 2009 Quiet Leadership Award. Ms. Harris-Thomas has thirty-eight years of federal service as an I/O psychologist and is a former manager of test development and validation research at the U.S. Office of Personnel Management. She also has served as lecturer on a variety of HR topics at universities, federal, state, and local organizations, and professional psychological conferences.

Donald M. Truxillo, Ph.D., is a professor of psychology at Portland State University. His research examines the methods employers use to hire workers and the way that applicants perceive their potential employer during the hiring process. In addition, Dr. Truxillo examines issues associated with older workers, including older worker stereotypes, perceptions of aging, and older worker motivation. His work has been published in outlets such as *Journal of Applied Psychology*, *Personnel Psychology*, and *Journal of Management*. He currently is an associate editor at the *Journal of Management*, and he is a member of editorial boards for *Journal of Applied Psychology* and *Personnel Psychology*. He served as both program chair and conference chair for the SIOP Conference. He is a Fellow of SIOP, the American Psychological Association, and the Association for Psychological Science, and has served as member-at-large on the SIOP Executive Board.

David L. Van Rooy, Ph.D., is senior director, HR Business Process Management, at Marriott International. In this role he is responsible for Global HR Operations and Systems for several centers of expertise (COE) including learning and development, compensation, benefits, workforce planning, performance management, and associate engagement. Prior to this he was director, Talent Management Solutions, at Marriott International, where he served as the program lead for the global hourly selection initiative and also led the associate engagement survey and performance management programs. He was previously employed in the Talent and Organization Development Department of Burger King Corporation, where his role spanned leadership development, performance management, competency analysis, employee selection, and analysis, among other areas. Dr. Van Rooy received his doctoral degree in the I/O psychology from Florida International University.

Delisa D. Walker currently serves as the chief of the HR Research and Assessment Office for the U.S. Secret Service. Ms. Walker has extensive experience with the development and validation of competency-based assessments for various occupations within the federal sector, including video-based assessments. Her highlights include the design and implementation of the Secret Service's Special Agent promotion assessments, Customs and Border Protection's Video-Based Test for the CBP officer and agriculture specialist positions, and the design and implementation of the FBI's Police Officer Selection System, which included a video-based assessment component. Ms. Walker received an M.S. degree in Industrial/Organizational Psychology from Radford University in 1998 and is a doctoral candidate in the George Washington University's Human and Organizational Learning Doctoral Program.

Michael J. Zickar, Ph.D., is an associate professor of psychology and department chair at Bowling Green State University, has published extensively in the areas of applied psychometrics as well as the history of industrial-organizational psychology. He is a former historian of the Society of Industrial/Organizational Psychology and has served as an expert witness for employment testing.

Preface

For more than one hundred years, tests and assessments have been tools used by organizations to identify capable individuals for selection and promotion into a wide array of positions. Over these years, the science of accurate and fair assessments has evolved, built on accumulating research on both assessment tools and criterion measures as well as the relationship between them. However, until the introduction of technology and its subsequent widespread use in most areas of business and industry, little changed in administration practices. Paper-and-pencil administration of tests was the primary method of delivering assessment procedures in high volume for most of those years. Technology has increased the pace of change in assessment practices as well as enabled sophisticated assessment techniques that would not have been possible without computers. Technology—and particularly the Internet—has also opened up assessment to a wider and more global marketplace.

This book is focused on the recent changes in assessment that are due to technology. It covers both the effects of deploying technology in the design and delivery of assessments on the psychometric properties of the instruments and procedures as well as the implications for practice. The foundation chapters summarize current concerns about technology-enhanced assessments and the research findings to date. The case studies offer detailed examples of assessment programs that have been carefully designed, validated, and implanted and provide advice to the practitioner who is contemplating technology-enhanced assessment.

We hope that this volume informs you and challenges you. The book is intended to give you an overview of the state of the art and science of technology-enabled assessment; highlight effective practice in the development, validation, and

implementation of a variety of assessment procedures; and caution you about the potential pitfalls when technology is used in assessment. In addition to guiding the industrial and organizational psychologist and the human resources professional who undertake technology-enabled assessment, we also hope this book will stimulate new ways of using technology effectively in evaluating the knowledge, skills, abilities, and other characteristics of applicants and employees. As technology evolves, so should technology-enhanced assessment.

Nancy Tippins
Seymour Adler
March 2011

Acknowledgements

Edited volumes require the time and attention of many people. We appreciate the efforts of the highly qualified authors who contributed chapters to this book; Allen Kraut, the editor of the Professional Practice Series; and the editorial staff of Jossey-Bass. Without any one of them, this book would not exist. We are particularly grateful to the employers of the many practitioners who contributed their experiences for allowing them to share their work and further the field of assessment in organizations. Finally, we must acknowledge our own employers and colleagues for their forbearance during this process.

Chapter One

OVERVIEW OF TECHNOLOGY- ENHANCED ASSESSMENTS

Nancy T. Tippins

The availability of affordable, easy-to-use technology-based tools and their interconnectivity via the Internet have hastened the spread of technology into virtually all aspects of our lives. The Internet provides quick access to huge amounts of information and facilitates all kinds of relationships among all kinds of people. In the last fifty years, technology has changed how we do our work, how we spend our leisure time, and how we interact with others. The next fifty years promises more of the same.

Technology has permeated work in the 21st century, and the field of talent assessment is no exception. Technology has influenced what kinds of assessment tools are used as well as how they are developed and administered, enhancing some traditional practices and fundamentally changing others. For example, a multiple-choice test may now be administered via a computer that displays items, scores responses, and stores test results. Alternatively, realistic work samples administered via a computer that might have been too labor intensive or too inconsistent in the past can replace a more abstract form of standardized testing. Unproctored prescreens administered via the telephone or a computer narrow

the applicant pool to a more manageable size. In the development phase of a test, large numbers of items are developed and their item parameters are defined so that many equivalent forms can be constructed automatically for computer adaptive testing.

The objective of this volume is to enable practitioners to make better decisions about using technology-enhanced assessments in the workplace that are based on current scientific knowledge and best professional practices. This volume explores the methodological underpinnings of technology-enhanced assessment as well as the measurement concerns its use raises, and then provides examples of how technology has been employed in assessment procedures in real-world applications. The purpose of this first chapter is to set the stage by defining the scope of what will be covered in the volume and then providing a brief discussion of the opportunities and the challenges the use of technology-enhanced assessments presents. Brief sections on the future of technology-enhanced assessment are presented before the chapter concludes with an overview of the entire volume.

What Is Technology-Enhanced Assessment?

In this book, we refer to technology-enhanced assessment as the use of any form of technology in any aspect of testing or assessment. Technology-enhanced assessments can include various technologies used for presenting or scoring items or other assessment materials. For example, computers, personal digital assistants (PDAs), telephones, interactive voice response (IVR) equipment, or video-teleconferencing equipment may be used to present testing materials. Perhaps most simplistically, a computer serves as a page turner that presents items and response alternatives on a screen and collects responses, or an IVR presents items orally and records responses. At the other extreme, complex in-baskets that involve emails, voice messages, memos, telephone calls, and appointments simulate actual work and require test-takers to behave as they would in a realistic setting.

The range of testing formats used when technology is introduced is broad. High-volume testing programs, particularly those focused on screening candidates, continue to use multiple-choice

formats administered on computers and IVRs. Yet, other item formats are increasingly used in conjunction with technological tools. Structured interviews have been adapted for delivery over computers. The work samples and simulations that are components of assessment centers or sophisticated selection batteries are often delivered via computer and responses to them are captured electronically via a computer or video or audio recording equipment. Some assessments make use of video teleconferencing equipment that enables assessment center participants and assessors to work from different locations.

Computers can be programmed to deliver fixed forms of a test or determine which items or tests to present depending on answers to questions on an application blank (for example, For which jobs are you applying?) when the assessment system is integrated with an applicant tracking system (ATS), or on responses to previous items as in computer adaptive testing. Similarly, sophisticated, electronic test data bases can automatically determine who is eligible to test and when they are eligible.

Computers are often used to score test items by determining which responses are correct and incorrect as well as aggregating responses to items into test scores and test scores into battery scores, sometimes based on complex algorithms. Traditionally, computers have simply been used to execute programs that specified exactly what was right and what was wrong. Increasingly, computers can take into account patterns of responses to items in more complex scoring procedures. An emerging technology that is beginning to be used more often involves data mining techniques that evaluate complex written responses. Although many constructed responses must still be evaluated by human evaluators, video technology that records the responses allows checks of the scoring process that increase accuracy.

Similarly, computers can be used to store responses as simple as the number or letter of a response alternative or as complex as the summaries and spreadsheets associated with a business case or a videotape of an interactive role play simulation or the written responses to a structured interview that has been presented online. Typically, responses to test items that are presented electronically are also stored electronically. Even when tests are not delivered via a computer, test results may be

entered into an electronic database. Increasingly, technology is used to distribute test responses and test scores. For example, work sample products are distributed to assessors electronically; test qualification status is sent to hiring managers; or test feedback is sent to candidates. Because of security concerns, many test users avoid distributing confidential information such as test results via email; instead, these are stored in “eRooms” where authorized users may access confidential data.

What Are the Advantages and Disadvantages of Technology-Enhanced Assessments?

As technology has become increasingly easy to use, affordable, and widely available, industrial and organizational psychologists have learned that there are many factors to be considered when making decisions about how to use technology in assessment and that few factors can be considered solely an advantage or a distinct disadvantage. Instead, the thoughtful industrial and organizational psychologist must consider the entire set of benefits and liabilities of a specific technology-based approach in his or her specific situation and compare them to the pros and cons associated with each of the alternatives. The next section highlights the most important factors.

Cost

An overall assessment of cost is particularly difficult to obtain because there are typically many sources of costs in a technology-enhanced assessment program. For example, there is the cost of administration, and there is the cost of developing items. Moreover, there are tradeoffs between costs and the anticipated benefits. For example, an organization may spend the money to develop a computerized work sample not because it is cheaper but because the realistic assessment results in a better estimate of an individual’s skills, attracts better-qualified candidates, or provides a realistic job preview. Another organization may computerize its executive assessment process in order to standardize the process globally, even in locations where face-to-face assessment is practical.

In many respects, the use of technology has lowered the cost of assessments. As technology has replaced live administrators, proctors, scorers, and data-entry personnel, labor costs have undoubtedly decreased dramatically. Even when an organization considers the cost of outsourced test administration services, the costs are typically reduced whenever personnel have been replaced by computers.

When computer-administered tests began to be used in private industry for large scale selection programs, many industrial and organizational psychologists were concerned about the high cost of equipment. However, two important things have happened since that time to alleviate that concern. First, the cost of equipment has dropped substantially. At this point in time, it is reasonable to assume that most equipment costs are more than compensated for by reductions in labor costs. Second, many testing programs have shifted the obligation to provide equipment from the employer to the candidate through unproctored testing programs.

The equipment on which a test or assessment is administered is only one type of equipment usually required. Large scale testing programs often require servers that contain the administration programs and executable modules to be downloaded to the user's computer as well as data bases to store results, including data at the item, test, and battery levels. Increasingly, demands for reliable accessibility require redundant servers, and security concerns necessitate highly technical barriers to these servers. Some users of assessments that transmit real-time video may find that the bandwidth required is not available in some countries at any price.

At the same time that costs of labor associated with administration and scoring and equipment have diminished, other sources of costs may have increased or new sources of costs may have been introduced. For example, the number of items that are required for computerized tests often increases substantially because of security concerns, particularly when unproctored Internet testing (UIT) is used. Thus, more labor is required to develop and maintain a larger pool of items. Similarly, test administration procedures involving computer adaptive testing were not feasible in most situations without a computer; yet, the programs for such administration have to be

written and maintained. Similarly, complex work simulations like in-baskets may require substantial programming expense.

In addition to requiring large number of items with accurately defined item parameters, UIT used for selection purposes can introduce other costs. For example, companies that use verification should account for the costs of the UIT plus those associated with later verification testing. Moreover, the employer must also consider what effect UIT has on its applicant pool and determine if the UIT has an effect that has implications for costs such as broadening the applicant pool or reducing the number of qualified people who remain in the recruitment and selection process or who are likely to accept a job offer. A technology-enhanced testing program that is off-putting to qualified candidates may reduce testing costs while increasing recruiting costs.

Whether the reductions in costs exceed the increases in costs is obviously dependent upon many factors including the choice of instruments, the organization, its staffing context, the resources available, and the expectations of its applicant pool. Direct comparisons of total costs for various approaches are difficult if not impossible to make. Each assessment user is advised to carefully consider all the sources of expense as well as the tradeoffs among various elements of the staffing process, and plan accordingly.

Effect on the Quality and Quantity of Candidate Pool

A critical concern for organizations that use any sort of assessment for selection purposes is the impact on the quality and quantity of the candidate pool. Many assessment programs that have incorporated technology into their delivery are still administered in controlled settings with proctors. In theory at least, these technology-enhanced assessments should have no effect on the size of the candidate pool when compared to a proctored paper-and-pencil version of the same test. However, apple-to-apple comparisons are often not made. When apples are compared to oranges, some might argue that a realistic, technology-enabled assessment program used for selection may be more engaging and may help keep some candidates in the applicant pool longer than a less realistic form of evaluation that does not require technology.

Many technology-enhanced assessments are administered in unproctored conditions at times and places of the candidate's convenience; yet, there is little consensus on the effect of this flexibility on applicant behavior during the recruiting, selection, and hiring processes. Many staffing professionals argue that the freedom to take a pre-employment assessment any time or any place greatly expands the number of people who actually take the test. The lack of constraints on the actual testing event may also improve the quality of applicants because the employed are able to look for other employment without taking time off from their current jobs. There are contrasting arguments, however, that suggest the number and quality of applicants may be limited when UIT is used and the applicant must supply the equipment necessary to take the test. If a digital divide exists, UIT may have no effect on applicants from higher socio-economic status brackets but severely limit representation from lower brackets. Based on anecdotal evidence, recruiters often argue that many applicants have a low tolerance for completing lengthy applications and tests on the Internet and only the most desperate candidates will pursue lengthy and rigorous online selection procedures. Others postulate that highly qualified candidates have higher expectations regarding their treatment as applicants and drop out of the recruiting process when the selection procedures do not acknowledge their special qualities. Simultaneously, one could hypothesize that some applicants appreciate the respect for their time and the recognition that some assessments do not need to be administered in a face-to-face setting. Some employers fear the use of UIT will dissuade the honest applicant from pursuing employment because of the company's assumed acceptance of malfeasant behavior. Further, UIT may increase the amount of cheating that occurs on some types of tests and consequently result in a less qualified pool of candidates for the next step in the hiring process.

Perhaps, the most obvious effect of increasing or decreasing the quality or quantity of the applicant pool is on recruiting costs. If UIT increases the number of candidates who apply and remain in the selection process when recruiting costs are held constant, the per hire recruiting expense decreases. A larger proportion of more qualified candidates reduces the number of

people who must be attracted to the hiring process and evaluated. Equally important but sometimes overlooked is the effect of a larger applicant pool on the capabilities of new employees. Many employers want the best of the applicant pool and not merely the acceptable. As the applicant pool increases relative to the need for new employees, an organization may raise its standards and select individuals with higher abilities. The user of a technology-enhanced assessment for selection purposes should anticipate its effect on the quality and quantity of the candidate pool and take into account the implications for recruiting costs and for the capability of the workforce.

Candidate Expectations and Reactions

Closely related to concerns about the number of candidates and their capabilities are issues regarding candidate reactions to technology-enhanced assessments. Generally, organizations want positive candidate reactions because they are typically associated with candidates who stay in the employment process rather than drop out. In addition, many employers want to maintain positive relationships with applicants because they are also customers of the firm's products and services.

It is, of course, impossible to answer the question, "Do technology-enhanced assessments increase positive candidate reactions?" for all situations. The answer depends on what kind of assessment is used and what the candidate's expectations regarding assessment are. Some candidates will expect to see technology embedded in a testing program for some jobs in some types of companies, while others will expect a high-touch evaluation without technological intervention. For example, applicants to a manufacturing technician position in a high-tech firm might expect a highly mechanized selection process, but applicants to executive level positions in a service-oriented business might be disappointed in the selection system unless face-to-face interviews with the firm's management were used.

Different types of test and technologies also generate different reactions. One candidate may find a computer adaptive multiple-choice test somewhat irritating because everyone seems to get different numbers of items on a test. At the same time, this

candidate may enjoy a work sample test that measures arithmetic skills used in a teller job in a realistic setting. Others taking the computer adaptive multiple-choice test may prefer its efficiency to a lengthy test that presents a large number of items that are not particularly challenging. Candidates completing a realistic, technology-enhanced in-basket in the context of a leadership development and selection program may have negative reactions because the technology is different from the tools they use every day (for example, email programs, word processing programs). Other candidates' less than positive reactions to the in-basket may have more to do with the quality of the in-basket items and less to do with the technology.

It merits noting again that a candidate's reaction to the testing situation may affect his or her propensity to remain in the employment process, but once again, contradictory outcomes are possible. Some believe that capable candidates will exit the process because of concerns about an organization that appears to tolerate cheating in unproctored testing environments. At the same time, it is possible that more qualified candidates remain in a selection process when the selection process is efficient or realistic, as is the case with UIT or some work samples, respectively.

In addition to keeping well-qualified candidates in the applicant pool, another concern of many employers is the perceptions candidates have regarding the fairness of the selection system. The realism of the assessment practice often determines the candidate's perceptions of fairness regarding the testing system. The more like the job the assessment is, the less likely candidates are to claim the testing is unfair. Technology can make assessments more job-relevant, particularly when there is a heavy technological component on the job. At the same time, complex test administration systems (for example, computer adaptive testing) that use abstract items may be particularly subject to feelings that the test is not relevant.

Consistency of Administration and Scoring

Standardized testing conditions have been emphasized in industrial and organizational psychology because they increase the reliability of the test score and its validity and support a

common interpretation of the test score. One of the significant advantages of computer-based administration is the consistency of test administration and scoring. The computer does both tasks as programmed, and no test administrator forgets to time the test correctly or uses a key incorrectly or makes an error in scoring. Yet, because technology enables the distribution of assessments and many programs allow test-takers to take the tests and exercises any place and any time, variation in the testing environment is introduced. In contrast to a proctored test environment in which seating, lighting, temperature, etc., are often specified, an unproctored test may be taken under conditions that are rife with distractions and result in the test-taker's performance being less than maximal.

Security of Test Materials

Test development is expensive, and test materials (for example, test items, scoring keys) that have been compromised by UIT may result in test scores that are not interpretable because some people have had assistance that is not available to all candidates taking the test. Thus, employers usually take great care in protecting the materials. Traditional, paper-based testing programs have often emphasized procedures for accounting for testing materials such as serializing tests, accounting for all forms before and after use, storing them in secure areas, etc. Industrial and organizational psychologists who have managed test administration in such programs are all too familiar with security breaches ranging from candidates stealing tests and administrators leaving tests on copy machines to printers losing entire shipments of tests and unknown persons breaking into employment offices.

In many respects, simply placing a test that is proctored on a computer increases test security. With sufficient password protection, unauthorized people have a difficult time getting to the test. In monitored conditions, stealing a computer is more difficult than swiping a piece of paper. Yet, in unproctored conditions, candidates can capture questions from computers as easily as from paper documents by simply writing them down. Although it is difficult to leave a copy of a computer on a photocopier, there are other methods of acquiring test content through electronic

means, particularly if the Internet-based test is unproctored. Ironically, some of these methods of theft may be much more difficult and expensive to detect than a missing serial number.

Administrative Ease and Flexibility

The administrative ease and flexibility of a test program is a prerequisite to speedy and accurate selection decisions, and that administrative ease and flexibility depend on a number of things, including the personnel required to administer the test, the training required of those personnel, the equipment necessary for administration and its mobility, the ability to update the testing materials easily, and the dependability of the testing process.

Most employers pay close attention to the ability of their staffing personnel to administer and score an assessment in an efficient manner. Assessment programs that require large amounts of administrator time (for example, face-to-face assessment centers, structured interviews) usually cost more than those that require less time. Once developed, many computer-based administration programs are simple to use and require little administrator time. Once the administrator initiates the test, the computer often presents the instructions and items, times the test, scores responses, etc. Most computer-based tests can be easily updated by automatic downloads of which a user may not even be aware. The barrier to administrative ease and flexibility, however, is the initial programming that makes some of these functions possible. The costs of developing and maintaining the software that administers tests and upgrades programs is usually not trivial. Although a computer-based test may be easy to administer by personnel with minimal training, not all technology-enhanced assessments are necessarily used without extensive personnel training. For example, a computer-based in-basket may require extensive training for professionals who are already well-schooled in assessment procedures, and any changes to the scoring process may require retraining.

The flexibility of where an assessment is administered is one area in which there is no clear advantage of technology-enhanced assessments. On one hand, UIT facilitates administration virtually anywhere. On the other hand, the requirement for a computer

and an Internet connection may severely limit the places where a test can be administered. For example, job fairs may not be conducive to on-the-spot testing because of Internet connectivity issues. The flexibility of administration location for other forms of technology-enhanced assessments is also mixed. For example, physical ability work sample tests can be very difficult to move around. A pole climbing test can only be conducted where poles are planted. However, a strength test using an electronic load cell may be highly portable in contrast to a series of weights that are consistent with materials lifted on a job. Telephones for IVR prescreens are ubiquitous. Reliable audio and video equipment may be less widely accessible, particularly when used internationally.

Because it may be difficult to get applicants to a testing event, most employers want reliable assessment procedures that are ready to be used. Many perceive the IVR-delivered test to provide maximum flexibility in both when and where the assessment is administered and how available it is. Computer-based tests are also flexible in terms of time and place, and many are consistently accessible by the test-taker. Yet, again, there is no clear advantage of technology-based assessments over more traditional forms. Telephones, computers, and Internet connections have all been known to fail. Although telephone land lines represent one of the most reliable technologies used today, mobile telephones certainly drop calls. Many test delivery programs can function without the Internet connection, which must be re-established for scoring and storage of test scores. In contrast, a paper-and-pencil test is highly dependable; nevertheless, such a test requires qualified personnel to administer it and score it.

Cheating

Almost any mention of UIT, which assumes technology, raises questions about all kinds of cheating. How much cheating takes place? What kinds of cheating occur? Who cheats? What actions can prevent cheating? Etc. Although cheating is a major concern in UIT and the IVR prescreens, it is naïve to believe that cheating does not occur in proctored settings that involve no technology. Cheating can occur (and probably will occur) whether technology or a proctor is involved or not. Although UIT opens the door

to cheating, computerized administration also offers novel means of detection, albeit after the fact.

Some forms of technology-enhanced assessment may actually lessen the amount of cheating that occurs. For example, interactive assessment centers that occur via the Internet and are taped may actually enhance security by ensuring the person taking the assessment is actually the correct individual.

What Does the Future Hold?

A certain amount of hubris is always involved in predicting the future. Nevertheless, it seems safe to proffer the notions that (1) technological changes will continue at a rapid pace and (2) these new technologies will affect talent assessment as well as the workplace in general. If these speculations are true, the time of industrial and organizational psychologists must be spent learning about current technologies, staying abreast of emerging technologies, developing creative applications that use these technologies, and dealing with the problems inherent in them.

In addition to the need to understand technology as it relates to tests and assessments, however, the industrial and organizational psychologist must continue to explore the impact technology has on the evaluation of individuals and the conclusions the organization may draw from test scores. Many questions remain unanswered and, at times, the number of questions appears to be growing faster than the repository of research and answers. Although there are myriad questions that must be addressed for each type of technology-enhanced assessments, they can be grouped into two general categories: (1) effect on candidate behavior and reactions and (2) effect on the organization.

A fundamental question for all technology-enhanced assessments is the effect their use has on the test-taker, particularly those who are taking an assessment for selection purposes. Does the incorporation of technology make an individual more or less likely to apply for a job and take a test? To exhibit some form of malfeasant behavior? To remain in the hiring process? To accept a job that is offered? The answers to such questions are further complicated because the answers are contingent upon other factors, including the type of test, the kind of technology used, the

purpose of the assessment, and individual differences across test-takers. For example, the recent college graduate taking a cognitive ability test in an unproctored setting to obtain a high paying job may feel and act differently than a middle-aged employee completing a biodata form as part of a developmental assessment.

Just as technology-enhanced assessment affects individuals, it also can affect organizations. Organizations need to be able to accurately interpret test scores and draw appropriate inferences. Consequently, the effect of technology on validity and reliability may color how the organization uses the test score as well as the test policies that are set. Most organizations are cost-conscious and will be vitally interested in the costs and benefits of technology directly on testing and indirectly on recruiting. Ultimately, the organization will want to know the level of employee capability technology-enhanced assessment produces. Just as there are numerous factors that must be considered when evaluating effects of technology-enhanced assessments on the individual, there are other factors to be considered when answering organizational questions. What kind of test? What kind of technology? For what purpose is the test used? What kinds of individual differences?

Cheating on UITs has important implications for individual behavior as well as organizational behavior, and for most test users, the topic of cheating is critical. Researchers must continue the research on topics such as when cheating occurs, the extent of cheating, and the impact on validity, and practitioners must advise organizations in their use of tests administered in unproctored conditions, developing the appropriate guidelines and policies for use UITs and interpretation of their scores.

Finally, testing is not without ethical implications and legal consequences, at least in the United States. In the future, researchers and practitioners must provide guidance on the ethical use of technology-enhanced assessments, especially UIT, and maintain their understanding of the legal limits to their use.

Overview of the Chapters

To achieve its goal of aiding industrial and organizational practitioners in making wise decisions based on current science and best practices, the first half of this book contains a set of foundation

chapters that address critical measurement issues that face the practitioner considering a technology-enhanced assessment. The second half provides examples of innovative uses of technology that illustrate how technology has been used in employee selection.

The first section of the book begins with Chapter 2, which provides an overview of measurement issues by John Scott and Alan Mead. The authors emphasize traditional criteria for effective testing and provide a thorough description of the steps that must be undertaken to develop a test that is both reliable and valid. The authors explore the importance of standardization and measurement equivalence and discuss how cheating can affect the psychometric properties of a test.

Aiming to assist the practitioner in successful implementation of technology-enhanced assessments, in Chapter 3 Doug Reynolds introduces a framework for implementation to achieve four outcomes: (1) awareness building, (2) alignment and support, (3) planned flexibility, and (4) sustainability. Doug underlines the importance of understanding the environment and the organizational context as well as the assessment itself.

Winfred Arthur and Ryan Glaze take on one of the most critical issues in many technology-enhanced assessments—cheating on unproctored tests—in Chapter 4. The authors define cheating or malfeasant behaviors in both cognitive and non-cognitive tests. The chapter is organized around five questions: (1) What are cheating and response distortion? (2) What is the extent of cheating and response distortion? (3) How can they be detected and how effective are those methods? (4) What should an organization do with information about cheating or response distortion? (5) How can these behaviors be deterred?

Robert Gibby and Rod McCloy discuss computer-adaptive testing in Chapter 5. After a brief description of CAT and explanations of general principles, they provide an example of an unproctored, cognitive CAT that was developed in Procter & Gamble for employee selection.

In Chapter 6, Talya Bauer, Donald Truxillo, Kyle Mack, and Ana Costa consider the special problems and opportunities associated with candidate reactions that may be raised by technology-enhanced assessments. Using Gilliland's model of applicant reactions, they discuss the effects of technology on test-taker reactions and provide recommendations for practice.

Dave Bartram presents the special international issues that are relevant when technology is used to deploy testing globally in Chapter 7 and emphasizes the international guidelines that shape professional testing practice.

The second set of chapters provides case studies presented by practitioners who have used technology in their assessment programs.

In Chapter 8, Terri McNelly, Brian Rugeberg, and Carrol Ray Hall describe an executive assessment program at Darden, a large restaurant company. Using a virtual assessment center, this team delivered realistic simulations that were aligned with the organization's competencies, cost-effective, realistic to be used for external selection, internal promotion, and internal development.

In Chapter 9, Adam Malamut, David Van Rooy, and Victoria Davis share their experiences with a web-based screening program, "Hourly eHiring" at Marriott, which has three components: an online applicant tracking system (ATS), web-based assessments, and fully integrated HR systems. Of particular interest here is the web-based assessments for "heart of the house" jobs such as housekeepers, kitchen helpers, and groundskeepers that assess job relevant knowledge, skills, abilities, and other characteristics while keeping literacy requirements to a minimum.

Amy Grubb has described her work with high-fidelity simulations that are used for promotions to mid-level managerial positions in the Federal Bureau of Investigation (FBI) in Chapter 10. In response to a consent decree and the challenges brought about by the events of 9/11, the FBI overhauled its promotion process and designed and implemented a realistic, "day-in-the-life" simulation that is remotely administered. Amy shares the processes she used for development, validation, and implementation.

Sandra Hartog covers technology-based assessment centers and coaching in Chapter 11. In partnership with The Interpublic Group of Companies, Inc., Sandra and her colleagues developed MyLead, an experiential leadership development program for individuals in mid- to senior-level leadership roles located around the world to provide information for a succession management program.

Chapter 12 presents another approach to technology-based assessment centers for bank branch managers created by Rick Hense and Jay Janovics. Their assessment process uses multimedia (video-based) and psychometric (computer adaptive testing) technology to provide realistic assessment exercises in an unproctored setting.

The work of Jeff Cucina, Henry H. Busciglio, Patricia Harris Thomas, Norma Callen, DeLisa D. Walker, and Rebecca J. Goldenberg Schoepfer in the area of video-based tests (VBT) to evaluate applicants for law enforcement officer positions at the U.S. Customs and Border Protection is presented in Chapter 13. The VBT uses video technology to evaluate judgment and interactional skills in a realistic setting. The chapter shares the psychometric properties of the VBT and outlines the development process used.

In Chapter 14, Eugene Burke, John Mahoney-Philips, Wendy Bowler, and Kate Downey share their experiences in two international companies with using UIT for campus recruitment of college hires and customer service employees in call centers.

In Chapter 15, Mike Fetzer and Tracy Kantrowitz provide an example of computer adaptive testing in the public sector, the Human Resources Department of Riverside County, California. These authors provide the organizational context as well as information about the test and its implementation.

The final two chapters present agendas for future research and practice. Mike Zickar and Christopher Lake present a practice agenda in Chapter 16. In this entertaining chapter, Zickar and Lake emphasize ethical, scientific, and practical issues in three examples of technology that may be used for assessment in the future: use of personal information from the Internet (digging for dirt), brain scanning and imaging, and virtual reality. They conclude with advice to the practitioner for staying current with new technologies.

In the concluding chapter, Seymour Adler highlights the unanswered questions the profession still has. Seymour organizes these questions around four categories: The Assessment; The Candidate; The Organization, and Society.

Final Word

We hope this volume will stimulate your thinking about technology-enhanced assessments, guide your practice, and shape your research; however, as a final word of caution, we point out the obvious. Technology changes rapidly, so much of this book will soon be out-of-date. Although the foundational chapters, especially the criteria for good testing, will remain relatively constant, the innovative uses of technology described in the case studies today may be somewhat stale in the future. New questions will undoubtedly emerge and new research will inform our understanding. Consequently, the industrial and organizational psychologists who work in this area must continually update their knowledge and skills.

Section One

Measurement and Implementation Issues in Technology- Enhanced Assessments

Chapter Two

FOUNDATIONS FOR MEASUREMENT

John C. Scott and Alan D. Mead

Technology-enhanced assessment offers tremendous opportunities and unique challenges in the measurement and prediction of human behavior. By harnessing emerging technologies, organizations can reach across the boundaries of language and geography to accurately assess an almost limitless array of candidate attributes. Test users can now leverage sophisticated, web-based, assessment platforms to simulate any number of work environments and situations—effectively capturing candidates' ability to respond under real-life conditions. These advances in technology have both demanded and facilitated the development of new measurement practices and theories (for example, adaptive testing, item response theory) that have resulted in significant enhancements in assessment precision and efficiency. When used properly, automated assessments have the potential to provide a much more reliable, accurate and efficient means of measuring human characteristics than their erstwhile (paper-and-pencil) counterparts.

Despite the clear benefits and advances that technology-enhanced assessments bring to the table, there remain some key challenges that must be addressed to ensure alignment with sound measurement principles and practices. Increasingly, pressure has been mounting by a variety of test

users to reexamine certain testing principles that they believe are limiting the full potential of automated assessments. One notable example relates to the use of un-proctored Internet tests (UIT). Not so long ago, good testing practice would require a group of test takers (such as candidates for a job) to be assembled in a well-lit, distraction-free room with trained proctors who would verify each test taker's identification, distribute the tests, read aloud the instructions, answer any questions, monitor the time limits, ensure test security and collect and log the tests and all associated materials. This standardized mode of administration was established to duplicate the procedures used in validating the test so that the results could be confidently interpreted for making sound decisions. For larger organizations that may test thousands or even tens of thousands of candidates a month, the logistics, time and costs associated with these sorts of standardized testing practices have led to questions regarding their real value and whether the rewards associated with violating a few established practices might in fact outweigh the risks.

There is no question that a clear business case can be made for the use of technology-enhanced assessments. In fact, as organizations begin to recognize the potential of automated assessments, their use will increase significantly and continue to expand on a global scale. The question then becomes how to achieve the right balance between a business's return-on-investment priorities with that of sound measurement practices so that critical assessment decisions can be made with efficiency, accuracy, and integrity.

The purpose of this chapter is to address the measurement challenges—and highlight the opportunities—that technological advances bring to the assessment field. We begin by laying the foundation for sound measurement practice that will provide solid support for building and implementing high-quality assessments. We then explore the importance of standardization and measurement equivalence in the context of automated assessments and reveal how cheating, response distortion, and retesting can impact an assessment's psychometrics. We also address how computer access and the "technology divide" can impact performance on the assessment.

Building High-Quality Assessments

While the specific format of technology-enabled assessments can vary widely, there is a core set of underlying measurement principles that should be applied universally, regardless of how the assessment tools are configured and administered. Without the foundation of solid psychometrics to drive these assessments, the advantages that technology brings will ring hollow and the organization may actually be worse off than if it hadn't implemented an online assessment in the first place.

Because the focus of this chapter and this book is on the assessment of talent in organizations, we will direct our discussion to those measurement criteria required to successfully assess and predict behavior in the workplace. While organizations may decide to buy or build their assessment programs, the measurement criteria described below apply to either decision.

The quality of any assessment can be evaluated by the extent to which it: (1) measures relevant criteria, (2) follows a clear set of assessment specifications, (3) provides a precise and consistent measure of the characteristics it is intended to measure, and (4) produces appropriate inferences (that is, prediction) of behavior and performance.

Measure Relevant Criteria

The first step in developing (or purchasing) a high-quality assessment tool is to clearly specify the constructs (knowledge, skill, ability, other personal characteristics; KSAOs) that need to be measured. This involves more than an informal review of job descriptions or anecdotal accounts of what it takes to be successful in a job. What is required, particularly when high-stakes testing (for example, selection) is involved, is a well-executed job analysis. Job analysis should serve as the foundation for any assessment program. Legal guidelines (Equal Employment Opportunity Commission [EEOC], 1978) and professional standards and principles (APA, 1999; *SIOP*, 2003) describe the importance of job analysis in the development of legally defensible, fair, and effective assessment programs.

There are a number of different approaches for conducting a job analysis that have evolved over the years and that are reflective of the dynamic nature of work and new organizational challenges. The choice of job analysis methods is driven by the purpose of the assessment (for example, training diagnostic versus hire/no hire decision), as well as practical and legal considerations, and there is no one preferred approach for all situations.

One way to determine how rigorous a job analysis should be to support a particular assessment application is to consider the level of risk involved should it be challenged. As the stakes increase, so does the level of rigor required in the job analysis. When assessment systems are challenged legally, the first area often investigated is the job analysis. At issue is how comprehensive and accurate the assessment criteria are to support the talent-related decisions. Unfortunately, many companies cannot produce solid job analysis data or documentation, and in many cases they must conduct “post hoc” analyses when faced with a challenge to their assessment program. It is always most efficient and cost-effective to conduct a robust job analysis as the first step in implementing any assessment program.

Develop Assessment Plan

Once the job analysis has been completed, the next step is to create an assessment plan that will clearly outline the attributes that need to be measured and identify the types of assessments appropriate for the targeted application. The assessment plan will establish the framework and specifications for determining: (1) the most appropriate item types and administrative format, (2) how to properly construct the assessments, and (3) how to ensure that the assessment results possess the required measurement properties.

As technology advances, the variation in testing formats becomes almost limitless. Emerging technologies that include interactive simulations, the use of avatars and virtual reality will all become readily available to creative test developers (Reynolds & Rupp, 2010). Computer adaptive testing (CAT), which is already well entrenched in larger testing programs, has made good use of both advancing technology and theory to

provide highly reliable and innovative measures with far fewer items administered than would be required with traditional assessments. It is therefore important to account for the implications of these ongoing developments at the assessment planning stage, since it will impact the number and the nature of the items required. For example, while CAT administers fewer items, it actually requires a much larger pool of items than traditional testing formats to ensure adequate calibration across a range of ability levels.

The assessment plan should also account for user demands, such as the need to limit the length and administration time while also “engaging” candidates in the experience. These sorts of requirements have to be balanced with measurement considerations, such as the need to ensure adequate construct coverage and reliable results.

Build Assessment Specifications

The most effective way to ensure that an assessment is constructed to meet user demands while also accurately measuring the targeted attributes, is to develop a comprehensive set of assessment specifications. These specifications serve as a blueprint for the test developers and should draw upon the job analysis to systematically identify the topic areas to be assessed by the test and determine the relative weight that should be afforded to various KSAO areas within the assessment battery or single test. The specifications should fully outline the content to be covered, the number of items to be included within each content area, the stimulus and response characteristics of the items (for example, stimuli presented as pictures with associated audio—responses presented in a forced-choice format) and the administrative format. Exhibit 2.1 shows an extract of how this component of the test specifications might be presented.

There are dozens of novel item types that have emerged over the past decade (Hambleton & Pitoniak, 2002; Zenisky & Sireci, 2002), and there is certainly no lack of creativity when it comes to leveraging technology to simulate tasks across a broad array of work environments. Theory about the targeted attribute should drive choices about the types of items that will best evoke examinees’ demonstration of that attribute. For example,

Exhibit 2.1. Extract from Practical Reasoning Test Specifications

Candidates will be presented with modules consisting of multiple pieces of information. This information will be presented in different formats (for example, tables, charts, graphs, text, etc.) and appear to come from different sources (for example, memos, newspapers, books, manuals, etc.). Candidates will be required to:

1. Answer specific questions about the details contained in the material;
2. Evaluate the consequences associated with the information presented;
3. Sift through the information to identify what is critical for taking action or making a decision;
4. Take action based on the information; and
5. Interpret and use the information to solve practical problems or situations.

Stimulus and Response Attributes

The information will focus on practical, business-related issues drawn from critical incidents provided by subject-matter experts. During the online, multimedia test, candidates take on the role of a first-line supervisor and are presented with a variety of "real-life," on-the-job situations. These situations take place in areas including an operations center, a plant control room, a work site, and a customer call center. Candidates must determine how they would respond to work situations based on information presented through live-action scenarios and interactive information resources. This test will include full-motion video wherein candidates are provided with access to an entire desktop as though they are sitting at their desks. Interruptions (for example, phone calls, voice messages) will be built into the process as they would be on the job.

Candidates will be required to comprehend, evaluate and apply the information to solve problems, make decisions, and/or take action. There will be a total of thirty items in this section. Each item will have four response alternatives. Each alternative will plausibly relate to the content of the item stem. The correct answer will be based on accurate interpretation of the materials presented. Distracter response alternatives for items will be based on inappropriate or inaccurate interpretation of the information.

Skills Assessed

Performance on this test will be determined by candidates' ability to:

1. Extract relevant information from tables, charts, graphs, and text to solve practical problems;
2. Access, evaluate, and utilize information contained in manuals or other reference materials to make decisions, answer questions, or provide input to others;
3. Synthesize information from various sources and communicate relevant information to others;
4. Understand and apply new information, procedures, or principles to perform the task at hand; and
5. Attend to and verify the accuracy and completeness of detailed information in documents or on the computer.

conceptualizing *emotional intelligence* as an ability would suggest using items that require the respondent to view photos or listen to recorded conversations and identify the emotions experienced by the actors. Those who conceptualize *emotional intelligence* as a personality trait, on the other hand, might use self-report or biodata items. The challenge for technology-enabled assessments will be to select from the wide range of potential stimulus and response options that are available to solicit a clear, job-relevant and efficient demonstration of the targeted attribute.

Given the tremendous array of options afforded by technology-enhanced platforms, it is generally useful to have a guiding framework in mind when building specifications for innovative item types. Parshall, Davey, and Pashley (2000) developed an item taxonomy that can be helpful when organizing assessment specifications. They arranged item types along five dimensions of innovation: item format, response action, media inclusion, level of interactivity, and scoring algorithm. *Item format* refers to the type of response that is evoked from the examinee. The two major types of item formats are selected response (for example, multiple-choice) and constructed response (for example, essay, video recording of answer). *Response action* refers to the mechanism used to provide responses (for example, laptop camera, keyboard, touch screens). *Media inclusion* refers to whether and how video and audio are incorporated into the assessment. *Level of interactivity* refers to the extent to which an item interacts with or adapts to examinee responses (for example, CAT versus traditional) and the final dimension, *scoring algorithm*, refers to how the examinee responses are translated into score results. Parshall et al.'s (2000) taxonomy covers the key issues that need to be considered when blueprinting item types and formulating an assessment plan.

It is also important when building assessment specifications to include the expected distribution of psychometric indices (for example, difficulty and discrimination levels) based upon the purpose of the test (for example, mastery versus selection). This is particularly important for CATs, where the accuracy of ability estimates depends on a wide range of item difficulties within the item pool. The assessment specifications should also take into account whether or not the assessments will be proctored. If the test user plans on UIT, a large pool of items will be required so that they can be replaced on a regular basis and also used to populate any planned verification tests (Dragow, Nye, & Tay, 2010; International Test Commission, 2006). Finally, details about how the items will be scored should be clearly designated.

Incorporate Face Validity. One of the real advantages of technology-enhanced assessments is their ability to simulate key aspects of the work performed. Face validity is an important characteristic that should, whenever possible, be built into the

assessment specifications. The reason that this is important is that, particularly in high-stakes testing, examinees who believe that they are being assessed on characteristics relevant to the purpose of the test are more likely to place credence on the measure and try their best (for example, blueprint reading items on a selection test for an architect job, customer service simulation items for a customer service job). Assessments that predict future job performance very well but don't look or feel like their intended purpose (interpretation of poetry passages for a technical job that requires reading comprehension) may give rise to a legal and/or labor relations challenge should the examinee perform below standard on the test. Including face validity is usually a fairly simple choice (one that we recommend to all test developers) and will be easier with technology-enhanced measures that can readily simulate realistic, work-related scenarios.

Conduct Editorial Review and Pretest the Items

Once the assessment items have been constructed, and before they are field tested, an editorial review should be conducted to ensure that the items are properly formulated (for example, item stems are phrased as complete sentences, distracters “look” and “sound” like the correct answer). In the event that assessments are translated and will be used in other countries and cultures, it will be necessary to not only conduct a review of how well the assessment has been translated (see section on Adaptation and Language Translation later in this chapter), it is also recommended that an editorial board be convened that represents each country where the assessment will be implemented. This board should be tasked with ensuring that the actual intention or meaning of each of the items carries forward to the target culture. This review should be complemented by a field test that will provide a second level of analysis as to the fidelity of the translation.

Once the editorial review has been completed, the newly developed assessments should be field tested (this is independent of and as a precursor to a validation study) to ensure that the instructions are clear, the items are working as intended (difficulty and discrimination), and that the measures are reliable. The assessment plan should include the methodology for

piloting these items in the actual setting for which they will ultimately be administered (for example, proctored small groups, un-proctored kiosks). A pilot is absolutely essential to evaluate the measurement properties on a representative tryout sample. This sample should include representation from groups protected by EEO laws (for example, gender, race) within the United States and multicultural/multilingual representation in the case of a global selection program. It will be particularly important that every examinee attempt every item so time limits are generous enough to minimize the number of “not-reached” items.

Exhibit 2.2 provides an overview of the sorts of analyses that, at minimum, should be conducted on the pilot sample.

Gathering Psychometric and Validation Evidence

Once the assessment has been properly constructed and field tested, it is necessary to establish the psychometric and validity evidence needed to make accurate behavioral inferences. The challenge, and the opportunity in the context of new assessment technologies, is to demonstrate that the measurement properties of novel item types and administrative formats justify their application. This section reviews classical and modern approaches for establishing reliability and provides recommendations for enhancing the precision of technology-driven measures. This section also discusses the strands of validity evidence that are needed to ensure adequate coverage of the targeted attributes and accurate prediction of job-related behaviors.

Establish Reliability

Traditional assessments create reliable scores by containing a large number of the best items (as shown by pilot testing) and then scoring each item independently. The challenge for simulations and other innovative item types is the need to yield as many independent *measurement opportunities* as possible. Reliability is typically lower for work samples and simulations since the time

Exhibit 2.2. Pilot Test Analyses

1. P-Values (Item Difficulty—Proportion who responded correctly to test items)
 - Ensure item p-values match overall goal of test (for example, mastery test would contain majority of items with relatively high p-values (.8); selection test would have average p-values in the .5 range)
 - Test items for which everyone responds correctly or incorrectly do not help us distinguish between test takers
 - Test items in which two or more response alternatives have high p-values may indicate the presence of more than one “correct answer”
 - Low p-value for the “correct” answer may indicate the correct answer is ambiguous or, in fact, incorrect
 - Test items that have low p-values for “correct” answers or that have two or more high p-values should be resubmitted to subject-matter experts for their review
2. Biserial or Point Biserial (Item Discrimination)
 - The Point Biserial (item-total score correlation) will be higher (closer to 1.00) when high-scoring examinees get the item right and low-scoring examinees get the item wrong
 - Ensure items possess good discriminating power (differentiate between high and low performers)
 - Positive, high-item total correlations are desirable
3. Distracter Analysis
 - Ensure test items possess only one correct answer
4. Review Overall Test Statistics
 - Mean. Describes overall difficulty (or easiness) of test
 - Standard Deviation. Describes distribution of test taker’s test scores
 - Reliability. Reliability should exceed .80
 - Standard Error of Measurement (SEM). SEM should be considered for use in “banding” around the cut-off to take into account test’s measurement error

requirements of these items tend to limit the number of questions that can be administered. If a CPA candidate is to fill out a tax form that has four sub-schedules, and if a mistake in any sub-schedule will produce the wrong answer on the tax form, then provision must be made for partial credit or else the entire tax form simulation will be essentially one “item” on a traditional assessment. Scoring algorithms must also allow for normal variations (for example, deductions can be summed in box 11 or itemized in boxes 11a through 11e). Because measurement opportunities often cannot be easily dropped (like independent multiple-choice items), psychometrically poor items are often kept but weighted zero in the scoring. While simulations and work samples tend to provide more measurement information than traditional multiple-choice tests, they offer less information per minute of testing time than multiple-choice items (Jodoin, 2003), and therefore their results will generally be less reliable for time-limited administrations.

The basic requirements of reliability transcend administrative format and apply to all forms of assessments for which the objective is to produce an accurate measure of the targeted attribute. The key question in the context of this chapter is whether and how much technological innovations and associated practices impact the scope and magnitude of measurement error. For example, one might argue that the increased administrative flexibility afforded by UIT would most certainly increase measurement error, but that might be offset by the precision of a set of items presented in an adaptive format. The critical issue here is determining the major sources of error, estimating their size, and, ideally, identifying strategies that can leverage the technology to improve reliability. As the stakes and consequences of assessment decisions increases, so does the importance of reliability.

There are two psychometric theories in use today that drive our assumptions and approaches for estimating reliability: random sampling theory and item response theory (Bejar, 1983). Random sampling theory—which continues to be popular and in wide use—includes both classical testing and generalizability theories. This theory defines measurement error as the extent to which an individual’s observed scores on an assessment randomly

deviate from his or her hypothetical true score. The objective here is to determine how well an observed score generalizes to the universe from which it is drawn and approximates the true score. The larger the measurement error, the less confidence we have in generalizing beyond the observed scores and specific test.

It should be noted that reliability and standard error of measurement (SEM) estimates that are calculated through these random sampling theory procedures only apply to the test scores and not to the assessment itself. That is, reliability is considered an attribute of the test data and not the assessment, so it is inappropriate to ever state that the assessment itself is reliable. In fact, the *APA Standards* (1999) state that when reliability is reported, it must be accompanied by a description of the methods used to calculate the coefficient, the nature of the sample used in the calculations, and the conditions under which the data were collected. All of these caveats are necessary due to the fact that the reliability estimates calculated through these procedures are sample dependent and, as a result, have a number of practical limitations when building or evaluating technology-enhanced assessments.

Use Item Response Theory (IRT) to Replace Single Index of Reliability

It is necessary in high-stakes testing to be able to determine how well a test discriminates along the ability continuum, particularly around the critical values used to set the cutoff scores (*APA Standards*, 1999). IRT allows us to calculate measurement error to this level of precision by replacing the concept of reliability with that of the *test information function*. The test information function tells us how precisely each ability level is being measured by the test. One of the challenges in using IRT that has prevented more widespread application of this theory, is the sample size requirement for calibrating item parameters. For example, for a sixty-item test, a sample size of one thousand is generally required for stable parameter estimates using the three-parameter model. This is generally not a problem for large testing programs but may be so for those applications that have only a few hundred cases. Fortunately, most technology-enhanced

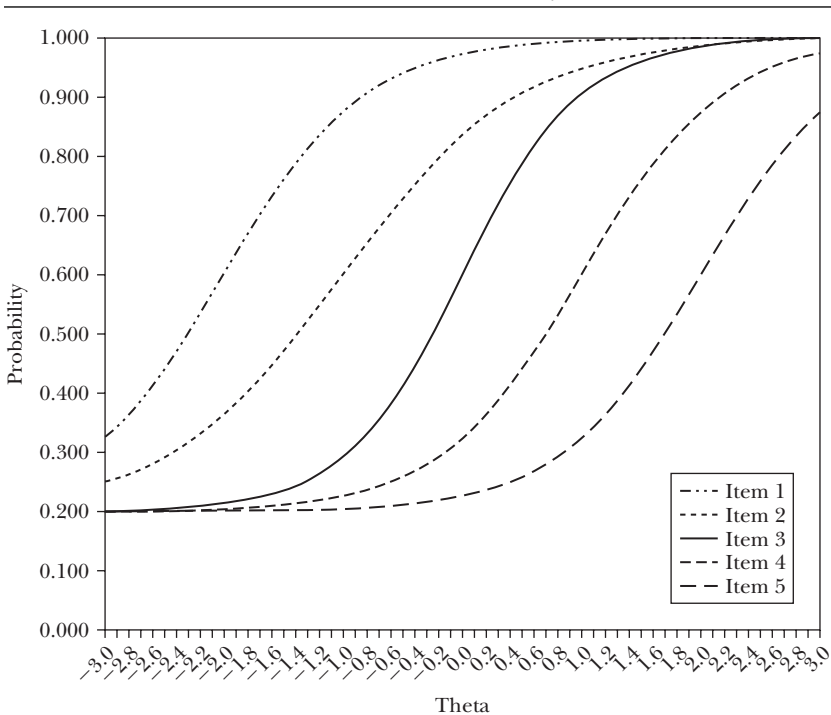
assessments are deployed because of high-volume hiring so the use of IRT becomes more feasible.

Under the IRT conceptualization, the relationship between ability (θ) and the probability of success on an item $P_i(\theta)$ can be expressed in the form of an item response function (IRF) as shown with five separate items in Figure 2.1. The probability of passing the item falls on the vertical axis, and the ability continuum (the “theta scale”) falls on the horizontal axis. As ability increases, so does the probability of passing the item.

This relationship between ability (θ) and the probability of success on an item $P_i(\theta)$ can also be expressed as:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[a_i(\theta - b_i)]} \quad (1)$$

Figure 2.1. Relationship Between Ability (θ) and the Probability of Success on an Item $P_i(\theta)$



The a parameter is the item discrimination index and represents the steepness of the IRF. The b parameter, which represents item difficulty, is defined as the point on an ability scale at which the probability of a correct response to an item is .5. The b parameter has the same metric (is on the same scale) as θ so the difficulty of an item can be directly compared to the ability of a test-taker. Item 1 in Figure 2.1 is the easiest and farthest to the left on the theta scale, while Item 5 is the most difficult and the farthest right. The c parameter indicates the probability that an examinee with very low ability will get the item correct and is often called the guessing parameter. It functions as the lower (or left-hand) asymptote of the IRF.

In terms of estimating an examinee's ability, θ , not all items are equally effective. IRT provides the *item information function*, $I_i(\theta)$, to show how effective an item is at measuring a given range of ability. The definition of item information is quite technical (the squared rate of change in the probability of a correct response divided by the variance of the item, as shown below). However, the use of information functions is quite simple.

$$I_i(\theta) = \frac{[P'_i]^2}{\sigma_i^2} \quad (2)$$

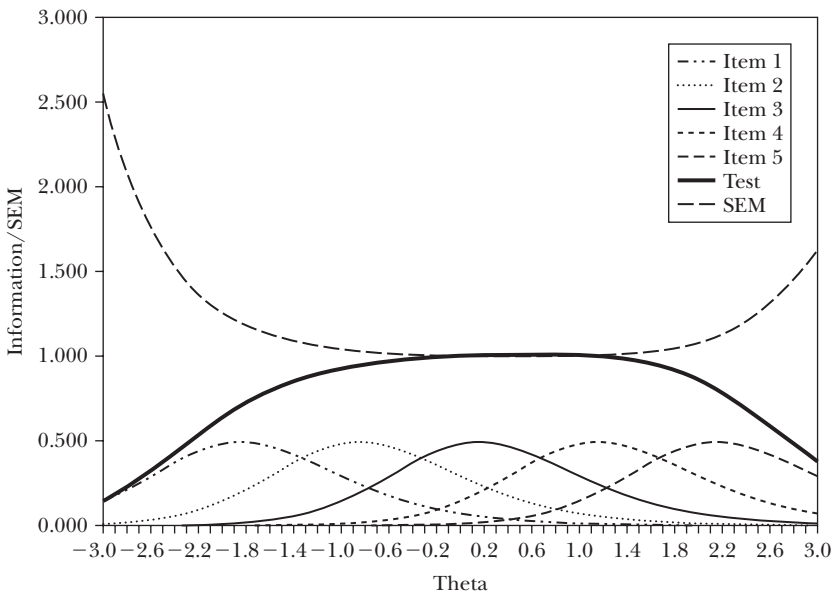
Figure 2.2 shows the item information functions for the items plotted in Figure 2.1.

Notice that where the IRFs rise steeply, information is high, while information is low where IRFs are flat—indicating that the item is ineffective at measuring examinees in that range of theta. Each item has a maximum degree of information and a range on the theta scale where it is effective. Item information functions sum to create the test information function, $I(\theta)$:

$$I(\theta) = \sum I_i(\theta) \quad (3)$$

Test developers should construct tests to have high information over the important ranges of the theta scale (or over the entire theta scale) by selecting those items yielding the most information. When IRT is used in this way, test length can be

Figure 2.2. Item Information Functions, Test Information Function, and SEM for a Five-Item Test



minimized without sacrificing measurement precision (especially using adaptive testing, described at the end of this section). In Figure 2.2, the bold line shows the test information (the sum of the individual item information functions). Although real tests would have more items, Figure 2.2 illustrates how tests can be constructed to have uniformly high information over the entire range of scores: The items must have a good spread of item difficulties and each item should have good item discrimination.

The degree of precision of the IRT test score, $\hat{\theta}$, can be calculated from the test information function. [Theta-hat, $\hat{\theta}$, is an estimate of the person parameter, θ , and is the IRT “score” for a test-taker.] The conditional standard error of measurement of $\hat{\theta}$ is the square root inverse of the test information function:

$$SE(\hat{\theta} | \theta) = \frac{1}{\sqrt{I(\theta)}} \quad (4)$$

Figure 2.2 also shows the relationship between the test information curve and standard error of measurement for a five-item test. The SEM is the “U-shaped” dashed line. The SEM curve is obviously a mirror image of the test information function. This means that imprecision/error of measurement is greater for scores at the edges of the score scale and is at a minimum across most of the score scale. Most tests have comparatively peaked test information functions because item difficulties tend to cluster around the center of the score scale. Good, general-purpose tests will look as close to Figure 2.2 as possible.

For pass/fail tests that have a known cut-score, the optimal assessment will have a test information function that peaks over the cut-score and may be quite low for other scores. This recognizes that on pass/fail assessments, only scores that determine whether a person passes or fails are important. On a driver’s licensing exam, for example, only the score that determines passing or failing is important to measure precisely; it is not helpful for that test to distinguish good from excellent (because both groups pass) or poor from very poor drivers (because both groups fail). Therefore, if we plan to use a single cutoff score in a selection context, a shorter test can be built by selecting only those items that are most informative at that specific ability level.

Computer adaptive testing combines advances in computer technology and IRT to create a very narrow, highly psychometric kind of artificial intelligence that can efficiently deduce the ability level of examinees from their responses with far fewer items than a traditional test. In fact, with a large pool of items calibrated using IRT, substantially shorter tests can produce more reliable scores. As testing is increasingly computerized and as item response theory becomes widely used, many assessment programs will encounter fewer barriers to its use and realize significant incremental benefits from adaptive testing.

It should be noted that many selection tests and non-cognitive assessments (for example, personality and attitude measures) have complex factorial structures and require multidimensional IRT models for item calibration. Multidimensional (MCAT; Segall, 1996) or bi-factor (BFCAT; Weiss & Gibbons, 2007) models provide a better basis for adaptively administering these assessments. Multidimensional models allow adaptive tests to leverage

the correlation among traits—when someone responds in an introverted manner, he or she is slightly more likely to be conscientious as well. For example, in one simulation study of the adaptive administration of the 16PF Questionnaire, a unidimensional CAT allowed a reduction of test length of about 25 percent with only slight loss of reliability (Mead, Segall, Williams, & Levine, 1999). However, test length on the MCAT could be reduced to about 50 percent with similar, small loss of reliability.

Bi-factor analysis is used for constructs with a main general factor and specific indicator factors (for example, general intelligence or personality measures like 16PF Extraversion, which is thought to be composed of Interpersonal Warmth, Liveliness, Social Boldness, Forthrightness, and Group Affiliation). In one application of BFCAT, Weiss and Gibbons (2007) examined a 615-item personality instrument that had an overall score and four content scores. On average, the BFCAT reduced test length by about 80 percent with slight loss of reliability.

Establish Validity

The fact that technology-enhanced assessments can be created to so closely simulate activities performed on the job sometimes raises questions by organizational stakeholders as to whether there is really a need to formally validate the tool. Since the assessment “obviously” measures elements of the job and validation studies can be a time-consuming and costly activity, what is the purpose of holding up implementation and delaying the dividends that the system could be paying? It is therefore not unusual to see validation placed low on the list of priorities by impatient stakeholders who may consider this activity more of a formality. However, despite the organizational pressures and advances in technology and measurement theory, the requirements for validation have not changed. The method may vary based upon the nature and purpose of the assessment (McPhail & Stelly, 2010), but validation is never optional.

According to the APA *Standards* (1999), “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests. Validity is, therefore, the most fundamental consideration in developing

and evaluating tests” (p. 9). In the talent selection context, the intended use of an assessment is to predict job performance. Therefore, we are interested in two facets of validity: (1) how well the assessment measures the criteria that underlie successful job performance and (2) how well the assessment actually predicts job performance. Guion (1998) refers to the first facet as *psychometric validity* (which subsumes content and construct validity) and to the second as *job relatedness* (that is, criterion-related validity).

Evaluate Psychometric Validity

In the case of simulations and work samples, the most frequently applied and generally most practical approach for gathering evidence of psychometric validity is through a content validity study. The objective here is to evaluate the extent to which the KSAOs measured by the assessment represent the targeted content domain, which is determined through the job analysis and fleshed out through the test specifications. A measure of a test’s content validity is generally not statistical (although expert ratings may be collected), but rather determined through agreement by subject-matter experts that the items used are representative of the domain from which they were sampled. The necessary ingredients for building evidence of content validity include a comprehensive job analysis, thorough test specifications, competent test construction and expert agreement that the test content is related to and representative of the content domain.

Gather Evidence Based on Internal Structure. For more traditional personality and multiple-choice cognitive ability tests, the interrelationships between items on the test—often assessed through factor analysis—can be an effective way to determine how well the structure of the assessment matches the intended framework. A review of the item statistics (item difficulty and discrimination) will also help determine whether the structure of the assessment supports the intended use. High item total correlations and internal consistency measures (for example, coefficient alpha) provide evidence that the test scores are systematically measuring some variable. If the content is based on a well-structured job analysis and the test has internal consistency, it is reasonable to

assume that the items are measuring the intended attribute without contamination (Guion, 1998).

An analysis of *differential item functioning* (DIF; Holland & Wainer, 1993) is another means for determining whether the items that comprise the assessment are operating as intended and support the assessment's internal structure. By reviewing DIF across different subgroups (for example, English- and Spanish-speaking shift supervisors for a multinational organization) with similar ability—or standing on an attribute—differentially functioning items can be identified for follow-up review and modification as necessary.

Evaluate Job Relatedness

The most direct way to evaluate how accurately an assessment can predict important job-related criteria is to conduct a criterion-related validation study. Evidence of job-relatedness is determined through a correlation between the assessment and the criteria of interest. Other forms of evidence can also be leveraged to support job relatedness under certain circumstances (see McPhail, 2007, for a detailed description of alternative validation strategies including transportability, validity generalization, and synthetic validity).

The choice of the performance criterion measures is of central importance in the validation study (*APA Standards*, 1999), and they must be held to the same psychometric validity standards used to evaluate the assessment measures (Guion, 1998). As is the case with assessment measures, the criterion measures must be based on a comprehensive job analysis and appropriately reflective of the multidimensional nature of job performance. Flaws in selection decisions can occur through too narrow a conception of the facets of job performance that contribute to success—and subsequently—missed opportunities to account for these facets by the assessment tools (Outtz, 2010).

Assessments should be validated under the actual conditions for which they will ultimately be administered. For example, if the intent is to administer the assessment in an un-proctored setting, the validation study should be set up to mirror these conditions. Likewise, if a verification test will be implemented to confirm the results on the un-proctored test, and it will be

administered under proctored conditions, the validation study of this test should occur in a proctored setting.

When a criterion-related validation study is properly conducted, the resulting evidence allows us to make informed decisions around how to maximize the prediction of performance, where to set passing scores to balance the goals of utility and fairness and how to implement a legally defensible selection program. The key criteria when evaluating criterion-related validity evidence are: (1) coverage of the important job performance criteria, (2) psychometric quality of test and criterion measures, and (3) relationship between predictor(s) and criteria (McPhail & Stelly, 2010).

Standardization and Equivalence

In a vault in the basement of the International Bureau of Weights and Measures on the outskirts of Paris, there sits a small cylinder of platinum and iridium—the *International Prototype Kilogram*, which has defined the meaning of “one kilogram” since it was manufactured in 1889. Copies of this standard exist in government bureaus around the world to enforce a standardization that allows a businesswoman in Beijing to know that the twenty kilograms of gold being offered by a dealer in London are equivalent to twenty kilos being offered in New York. Such standardization is essential for commerce and scientific progress in the physical sciences.

Standardization is also extremely important to psychological measures, where the construct being measured is unobservable and has no natural metric. If a personality test is being used to select workers, it is critical that it produce the same measurements on Monday and on Friday, this year and next, and when administered on computer or on paper. It is also critical, in many instances, that it produce interchangeable scores when administered to English-speaking Canadians and German-speaking Swiss and all the other languages used in locations where a multinational organization recruits professionals, managers, and salespeople. *Equivalence* is the degree to which standardization is maintained when an assessment is changed (for example, computerized). Thus, the topics of this section, standardization and

equivalence, are very important foundation topics for technological assessment.

Standardization Characteristics

High-quality assessments are standardized—they ask each respondent to react to the same set of carefully chosen questions or tasks under prescribed conditions designed to minimize irrelevant influences (for example, quiet rooms, adequate lighting, comfortable environment). Administration in a noisy, uncomfortable place might lower scores due to these distractions and *not* due to real differences in the knowledge or ability. Similarly, if a military assessment designed to measure performance under pressure (using loud recorded sounds, violent role players, etc.) were administered without such distractions, scores might well be significantly higher, *but not because the examinees were more tolerant of stress.*

Some researchers (for example, Weiss, 2007) have criticized web-based testing because standardization can be much harder, or impossible, with this media. However, some evidence (Buchanan, Johnson, & Goldberg, 2005; Stanton & Rogelberg, 2001) suggests that merely administering a test on a website does not preclude psychometric validity. If computerization makes reading the items harder or changes any other influential characteristic of the examination process, then the computerization itself may affect standardization. Because the characteristics that influence the examination process are not well understood, assessing equivalence is an important process.

Showing Equivalence

High-quality technology-enabled assessments are characterized by their equivalence across different conditions and groups. Measurement equivalence is related to standardization in that poor standardization, or violations of administration procedures, can produce non-equivalence. There are many other potential causes of non-equivalence. For example, research described below suggests that merely computerizing most kinds of assessments does *not* automatically cause non-equivalence. However, a bad interface design or very restrictive computer platform (for

example, a hand-held computer with a 4cm display or a tiny thumb keyboard) might introduce factors to the test that are irrelevant to the intended content. It would be inappropriate to compare people tested using paper and pencil to those tested with computerized tests that were not equivalent.

Relevant standards require test developers and users to show equivalence of paper and computerized forms of assessment. The APA *Standards* (1999) require that equivalence evidence be collected. The ITC *Computer-Based and Internet Delivered Testing Guidelines* (2005) are even more detailed, requiring that test developers show that computerized and paper forms have comparable reliabilities, correlate with each other at the level expected based on the reliability estimates, correlate comparably with other tests and external criteria, and produce comparable means and standard deviations or have been appropriately calibrated to render comparable scores (p. 11).

There are two main paradigms for researching equivalence: multiple groups and multiple measures and, as described above, there are three critical areas of equivalence. First, the computerized and paper forms should rank order test-takers similarly. This requirement ensures that computerized and paper forms have similar reliability and measure the same construct and can be shown statistically by correlating the scores of the computerized and paper forms of the test. Second, the mean and variance of test scores should be similar (either because of perfect raw-score equivalence or because form-specific norms are used). Finally, scores from computerized and paper forms of a test should have similar correlations with important external criteria, such as job performance.

If the mean or variability of scores on the computerized form are different from those of the paper form, then separate norms or *equating* (Kolen & Brennan, 2004) can be used to create interchangeable scores if the two forms have construct equivalence. Usually, the adjustments are fairly simple, such as adding or subtracting a few points.

Multiple-Groups Equivalence Designs

In the multiple-groups paradigm, one group takes one assessment (for example, computerized) and another group takes the

other assessment (for example, paper). If the groups are randomly assigned, then any important (“statistically significant”) difference in the mean scores for the two groups is taken as an indication of non-equivalence. The spread of scores for the two groups might also be compared to see whether one of the groups has a wider range of scores.

One serious problem arises if the groups are *not* randomly assigned to take one or the other form. If the groups are not randomly equivalent, then this design is seriously compromised because differences in test scores may well be due to group differences rather than with the form of the test taken. For example, if an attitude survey was administered on paper to day-shift employees and on computer to night-shift employees, what portion of the results are due to differences in the attitudes of day- and night-shift personnel? It is impossible to tell.

A technical problem with the multiple-groups equivalence designs is that hypothesis testing was designed to detect differences and is ill-suited to detecting equivalence (that is, standard hypothesis testing cannot be used to support the null hypothesis of no difference). Misusing hypothesis testing in this way has a number of unfortunate outcomes and should be avoided. While Rogers, Howard, and Vessey (1993) describe a framework for testing a hypothesis of equivalence, it would be best in these circumstances to discard hypothesis testing and rely on effect sizes (or equating).

A more fundamental problem with the multiple-groups approach is that we cannot correlate the scores on the two forms (for example, we do not have any information about whether people who scored well on the paper version also scored well on the computerized version). The correlation of scores on the two forms is *the* central issue in any equivalence research because it directly measures the degree to which the two forms of the assessment are reliably measuring the same thing. The multiple-groups paradigm is unable to address this question. Even worse, one could find that two quite different, and highly non-equivalent, assessments happen to have similar means and that two highly equivalent assessments happen to have different means. Thus, this design detects only one kind of non-equivalence and this kind of non-equivalence is the kind that is easily handled by separate norms or equating.

Thus, we are skeptical about equivalence research that uses a simple experimental approach and basic null hypothesis significance testing to compare paper and computerized groups. An alternative approach is to test for measurement equivalence (MEQ) using structural equation models (SEM; Meade, Michels, & Lautenschlager, 2007; Ployhart, Weekley, Holtz, & Kemp, 2002) or item response theory differential item functioning (IRT DIF; Raju, Laffitte, & Byrne, 2002). This approach can be used to test whether relationships between items and external criteria are the same across groups. Although the MEQ approach also cannot correlate scores across forms, it tests whether the items of the computerized and paper forms have identical psychometric properties (that is, the same difficulty and pattern of correlations with other items). Using this approach, it is assumed that if the paper and computerized forms are measuring different things, then the item psychometrics would not be exactly the same across the forms. The *Adaptation and Language Translation Issues for World-Wide Assessment* section below describes the SEM MEQ and IRT DIF approaches.

Multiple-Measurements Equivalence Designs

The alternative paradigm is the multiple-measure design, so-called because each volunteer is assessed with each of the forms (that is, measured two or more times). For example, all examinees might complete both the paper and computerized versions of a scale (a single group takes both forms of the assessment). Although this design has important methodological advantages (for example, allowing the researcher to correlate the scores on the two forms), there are unique problems that may arise through this design. The main issue is the influence of the repeated testing. It is best to administer parallel forms on different days, counterbalancing the order of administration.

What level of correlation shows equivalence? If the “true scores” of the test-takers are the same (to within a linear transformation) on the computerized and paper forms, then the observed correlation will be attenuated by the reliabilities of the forms. Equation 5 shows how estimated reliabilities can be used to estimate the true-score correlation of the computerized and paper forms (the so-called “disattenuated” correlation).

$$r(X, Y) = r(T_X, T_Y) \sqrt{r_{XX} r_{YY}} \quad (5)$$

In this equation, $r(X, Y)$ represents the “observed” correlation between the scores on the predictor (that is, test) X and the criterion, Y , $r(T_X, T_Y)$ represents the correlation between true scores and r_{XX} and r_{YY} are the reliabilities of X and Y . Because reliabilities are values less than one, the observed validity is always less than the true-score validity (the observed validity is said to be “attenuated by measurement error in X and Y ”).

If the estimated true-score correlation, $r(T_X, T_Y)$, is 1.0 then the construct being measured by the two forms is perfectly equivalent. [Note that even if the correlation is 1.0, the forms may have different means or variances and equating may be needed; however, if the equivalence correlation is low then no analysis can possibly produce equivalent forms.] Values below 1.0 indicate lower degrees of equivalence and, because they are correlations, are usually easily understood by psychologists and other test users. [Values above 1.0 should not occur; however, estimated correlations can exceed 1.0 due to sampling error. Byrne (1998) discusses this “boundary parameter” issue.] For example, values below .707 indicate that less than half of the variability in the scores on the paper and computerized forms are shared across the formats (see Mead & Drasgow, 1993).

Equivalence of Cognitive Assessments

Cognitive assessments have correct and incorrect answers and measure knowledge, skills, or abilities. In one of the earliest empirical comparisons of computerized and paper forms of an exam, military researchers (Sacher & Fletcher, 1978) administered vocabulary and logic tests to recruits in both computerized and paper formats. Their design allowed for the calculation of both the reliabilities of (both forms of) the test scores and the correlation of the scores across computerized and paper formats. The true-score correlation for 115 recruits was 0.95 and 0.87 for the vocabulary and logic tests, respectively. Although these researchers found other issues (for example, differences in response latency and answer changing), these correlations show

excellent comparability for the vocabulary test and good comparability for the logic test.

The logic test required recruits to answer six items per minute and so was considered fairly speeded, which likely impacted the comparability findings. Greaud and Green (1986) published an early and influential study of computerizing a speeded test, and they found poor equivalence. Thus, from the earliest research on this topic, speededness of the test emerged as a moderator of the comparability of computerized and paper forms.

The findings that speeded tests were less comparable should not be surprising because similar effects have been seen when seemingly small changes are made in the way that responses are recorded for speeded paper-and-pencil tests. For example, Boyle (1984) compared four groups who were all taking paper tests but using different kinds of optical marking answer sheets. He found that answer sheet formats requiring a single stroke were significantly different from a format that required a rather larger circle to be filled in—presumably a single stroke is a substantially different response than darkening a relatively large circle.

As more equivalence studies appeared in the literature, review articles also appeared to summarize the findings. In their influential narrative review of the literature, Mazzeo and Harvey (1988) suggested several possible moderators of equivalence, some of which have been subsequently discredited (for example, ability to change answers) and some which have been supported (for example, speededness). Bugbee (1996) provided another early narrative review that raised concerns about a lack of equivalence across media of administration in educational settings. However, a more recent narrative review by Paek (2005) concludes that K-12 students have access to computers in the classroom, frequently use computers for learning activities, and are comfortable with current technology. She concludes that the preponderance of the evidence supports equivalence except when long reading passages are present.

Mead and Drasgow (1993) published the first meta-analytic review and probably the most positive. For “timed power” forms (forms that were not highly speeded), they found a disattenuated correlation between paper and computerized formats of 0.97, which they interpreted as showing considerable support

for the equivalence of carefully developed computerized versions of cognitive ability tests that were not highly speeded (for example, the GRE).

When they examined highly speeded tests, they found a disattenuated correlation of only 0.72, which is a high correlation but clearly different from 1.0. (About half of the variance in the true scores of examinees were due to the computerization!) Also, equivalence of highly speeded tests was far more variable than for power tests. Mead and Drasgow interpreted this as support for speededness as a moderator of equivalence. Thus, it is particularly important to assess the equivalence of paper and computerized versions of highly speeded tests.

A few studies of computerization of speeded tests that have been published since the Mead and Drasgow (1993) meta-analysis have shown very good comparability between paper and computerized versions. Neuman and Baydoun (1998) showed essentially perfect true-score correlations for ten speeded clerical selection tests. Pomplun, Frey, and Becker (2002) studied computerized and paper versions of a speeded reading test and found a true-score correlation of 0.94. It is not clear whether computerized speeded tests are becoming more comparable to their paper counterparts (perhaps because of greater care taken by test developers, changes in the types of speeded tests studied, or because of advances in technology) or because of a file-drawer bias in published results, or some other reason. However, in a recent, carefully designed comparison of web- and paper-based speeded forms (Mead, 2010), we found a cross-mode true-score correlation of 0.80—close to the 0.72 value found by Mead and Drasgow (1993).

Equivalence of Non-Cognitive Assessments

Researchers have also examined the comparability of paper and computerized versions of non-cognitive predictors, such as attitudes, personality, measures of motivation, and automated interviews (for example, intake interviews). Early comparability concerns focused on changes in socially desirable responding, omitting items (Biskin, & Kolotkin, 1977), or anxiety (Canoune & Leyhe, 1985) caused by medium of administration. Of course,

almost everything about computers and our relationship to computerization is different today, as compared to the 1970s and 1980s when computers were comparatively primitive and uncommon.

In one large meta-analytic investigation of socially desirable responding on computerized, non-cognitive ability measures, Richman and her colleagues (Richman, Kiesler, Weisband, & Drasgow, 1999) examined differences between computerized and traditional formats across sixty-one studies and 693 means. They found an overall effect size of 0.02, which is very small, meaning that computerization matters very little.

Other recent, large-scale analyses have suggested fairly good comparability. Meade and his colleagues (Meade, Michels, & Lautenschlager, 2007) used a structural equations measurement equivalence framework to compare Occupational Personality Questionnaire (OPQ) personality scales in a large sample of undergraduates. Their results generally suggested that the scales of the OPQ functioned equivalently when administered on paper or on the Internet. Curiously, however, they found better equivalence when participants could choose the medium of administration than when they were assigned to a medium.

Mead and Blitz (2003) reported a meta-analysis of multiple-measures studies of comparability. They found 105 studies comparing paper- and computer-based versions of non-cognitive assessments, mainly attitude or personality scales. However, only *six* studies used the multiple-measures design. Across forty-one correlations from these studies, in a sample of $N = 760$, they found an overall true-score correlation of 1.02, which they interpreted as strong evidence for the comparability of non-cognitive abilities across administration modes.

Summary of Equivalence Issues and Recommendations

Research on assessments of both cognitive ability and non-cognitive constructs suggests that carefully developed computerized versions can measure the same construct as their paper counterparts—except for assessments with extensive reading or that are highly speeded, which may be noticeably less comparable. These results are good news because computerized tests that successfully measure the same construct should have the same criterion predictive

relationships shown for paper forms. However, a final question remains—are separate norms needed for the computerized and paper forms?

Mead and Drasgow (1993) meta-analyzed the standardized mean differences between paper and computerized forms of timed power tests. They found an overall mean of -0.03 , indicating that computerized power tests were very slightly more difficult than their paper counterparts. However, the estimated sampling error of these differences was 0.15 , indicating that one could easily sample a mean difference of 0.15 , 0.20 , or even 0.30 for a given assessment. Similar results were obtained by Richman and her colleagues (Richman, et al., 1999) in an analysis of the socially desirable responding on non-cognitive measures. Thus, we recommend that unless research has shown that a given computerization did not affect the norms of a paper form, the computerized form have its own norms.

Adaptation and Language Translation Issues for World-Wide Assessment

For large multinational employers, technology-enhanced assessment enables the global use of assessment systems on an unprecedented basis. Given the far-reaching talent consequences brought about by these technological advances, the need to properly adapt the assessment to the new context (for example, a new country, region, culture) cannot be overstated. Modifications to the assessment across contexts may range from minor issues (introducing UK English spelling and metric units) through the removal of cultural idioms to the translation of the assessment and instructions into a new language.

Some programs use translation/back-translation (TBT) to detect translation quality. Because TBT has not been systematically studied, its effect on translation quality is not empirically known, although its wide use (despite substantial cost) may suggest that TBT does have some value. However, there are several reasons to be skeptical that TBT is sufficient. First, the translators must be bilingual and thus likely to have had substantial experience with the other culture. Second, some terms may translate poorly, yet in a way that back-translates well. Finally, the content

of the instrument may interact with the culture of the respondents (Liu, Borg, & Spector, 2004; Ryan, Horvath, Ployhart, Schmitt, & Slade, 2000). A job satisfaction item that asks about one's boss might be perceived quite differently in high versus low power-distance cultures. Questions about co-workers in collectivist cultures may be affected by in- and out-group issues that matter little to individualistic respondents. So we strongly recommend that measures be pilot tested in the original and target culture(s) and measurement equivalence analyses be conducted to detect such issues.

Measurement Equivalence for Adaptations

There are two widely used frameworks for assessing the measurement equivalence of adapted tests, structural equations modeling (SEM) and IRT differential item functioning (DIF). Vandenberg and Lance (2000) provided an early review that clarifies how the SEM approach to measurement equivalence is far more nuanced (that is, complex) as compared to the IRT DIF approach, which focuses very closely on the equivalence of the item difficulty and psychometric quality across adapted instruments (for detailed comparisons, see Raju, Laffitte, & Byrne, 2002, and Stark, Chernyshenko, & Drasgow, 2006). The SEM approach is best when the response variables are fairly continuous (that is, item responses should be on 5- or 7-point Likert scales) and multivariate normal. Otherwise, an IRT DIF approach may be better. SEM may also be preferred because it can simultaneously assess equivalence across multiple groups (most IRT DIF approaches only analyze two groups, so they have to be used in pair-wise comparisons of multiple groups).

The SEM approach is quite flexible and is easily extended to assess the equivalence of item means (this model is sometimes called *mean and covariance structures*, or MACS; see Ployhart & Oswald, 2004). An interesting limitation of MACS analysis is that the item difficulties and group means cannot simultaneously be assessed. Analysis of changes in the item means requires that the analyst assume that the groups have equal means and analysis of differences in the group means requires that the analyst assume that the items have equal difficulty across groups. IRT DIF approaches have a clever solution—if *most* of the items

function equivalently, then IRT DIF approaches can separate the effects of a few items' difficulties changing from group differences. Empirical comparisons of the SEM/MACS and IRT DIF approaches suggest that they often reach similar conclusions when carefully similar analyses are conducted (see Stark, Chernyshenko & Drasgow, 2006) but there are also many instances of divergence due to different sensitivities of various IRT DIF statistics (see Raju, Laffitte, & Byrne, 2002).

The IRT DIF approach works well for categorical data, especially dichotomous responses from ability tests. IRT models are fit independently to each group and then a step called *iterative item linking* (Candell & Drasgow, 1988) is used to make the independent scalings comparable (incomparable scaling is very much like temperature measured in Celsius and Fahrenheit—the same construct but the temperatures cannot be compared until one converts to a common scale). Various IRT DIF methods can then be used to evaluate the comparable item scalings; see Raju and Ellis (2002) for a practical review of several IRT DIF approaches.

Cheating, Response Distortion, and Retesting

In this section, we focus on *the effect* of cheating, response distortion, and retesting on the psychometric properties of technology-enabled assessments. For a full discussion of cheating and response distortion, see Chapter 4 in this volume by Arthur and Glaze. *Cheating* on an assessment refers to any deliberate, malfeasant means of altering one's assessment score—that is, any attempt to obtain a higher score by improper, deceptive, or fraudulent means. *Response distortion* refers to cheating (most typically by inflating scores) on a self-report measure—for example, a person responding “Strongly Agree” on a personality survey item asking “I never miss deadlines” when, in fact, missing deadlines is a common occurrence for that person.

In theory, cheating and response distortion might completely destroy the validity of assessment scores. If all candidates completing an assessment obtained scores different from their natural score, the correlation of these scores with a criterion would probably be attenuated, perhaps to zero. Interestingly, because the reliability of assessment scores is affected by random error and because

cheating and response distortion might decrease random error, the reliability could actually seem to improve. However, this artifact simply shows that the validity of assessment scores is more important than the reliability of those scores.

Cheating and response distortion threaten validity in at least two ways. First, they may result in most people scoring about the same (technically, true-score variance is being diminished). When everyone scores very similarly, it is much harder to determine who are the best candidates—imagine a horse race where horses' noses are all within a few millimeters of each other; it would be difficult to determine the winner, even by photograph. We want assessments that allow individuals to express their natural differences in an area; cheating and response distortion act against this and diminish the value of assessment scores.

Also, if some people are cheating and others are not, the cheaters will tend to rise to the top of rankings of the candidates. Of particular concern are those low-ability candidates who obtain a high score by fraudulent means and rise dramatically to the top scoring band. If the top scoring candidates are selected, then they could disproportionately be cheaters (Zickar, Rosse, & Levin, 1996). If the assessment scores are (otherwise) valid, then that suggests that these cheaters will have poor outcomes (low tenure, poor performance, etc.). When these low potential performers are selected along with candidates who obtained legitimate high scores, and who are therefore high potential, the selected group's mean job performance will be lower and assessment scores will be less useful and have lower operational validity than it might otherwise have been.

So how bad is cheating and response distortion? Does it completely invalidate an assessment and automatically and invariably reduce the value of the assessment to zero? Not necessarily. In fact, validity coefficients are surprisingly robust to these issues (Ones & Viswesvaran, 1998; Rosse, Stecher, Miller, & Levin, 1998) because they take into account all of the scores from all of the candidates of an assessment. A small proportion of individuals obtaining higher assessment scores than they should (that is, resulting in lower job performance than their scores would otherwise indicate) does not necessarily produce lower validity coefficients, particularly when the overall impact of cheating

or response distortion is diluted over a large pool of candidates. Also, if all candidates distort their responses in the same way, the validity coefficient will be unchanged unless score changes cause “ceiling” or “floor” effects (where too many candidates get the best or the worst score), in which case the practical usefulness of the assessment might be severely compromised.

Clearly, applicant response distortion can seriously affect the norms, and practitioners should ensure that they use norms that were collected under conditions similar to the context in which the assessment is to be deployed. Applicant norms should always be preferred, especially for non-cognitive measures such as biodata and personality instruments, where response distortion is common under high stakes conditions.

The effect of cheating on the psychometric properties of assessments is difficult to quantify, as it is ultimately dependent upon the proportion of test-takers out of the total pool that actually cheated. This number in turn is dependent upon the level of exam security, the degree of organization among the cheaters, the difficulty of the exam, the controls put in place to minimize cheating, and the degree to which the outcome impacts candidates' lives. It is axiomatic that cheating reduces the usefulness of an assessment in predicting job performance. The consequences to an organization of even a single poor hire can be substantial when you consider the costs associated with training, low productivity, turnover, and the ultimate need to replace this individual. Multiply this by even a small number of low-ability cheaters hired into the organization and it becomes readily apparent that every effort should be taken to minimize opportunities to cheat on high-stakes exams. In particular, when high-stakes assessments are administered under unproctored conditions, it is highly recommended that the results be verified under secure, proctored conditions. The verification version of the assessment should be validated under proctored conditions.

Retesting

Retesting on the same form is widely assumed to be detrimental to exam security. For example, if an unqualified candidate knows that he or she will be retested with the same form, that

person may use the first assessment opportunity to memorize difficult items, which he or she will solve (or persuade friends to solve) at home and memorize so that he or she achieves an inappropriately high score. This becomes particularly problematic in unproctored testing situations where candidates may be allowed to take the test an unlimited number of times under assumed names before they actually submit their responses for scoring. Besides alternate forms, counter-measures include controlling exposure of items through item inventory control mechanisms and monitoring retest scores so that unusual score increases (for example, more than two standard errors of measurement) can be investigated (see Chapter 4).

In one provocative study, researchers arranged for actual returning candidates to randomly receive either the same or a different form of a radiography examination (Raymond, Neustel, & Anderson, 2008). They observed virtually the same score increase of about half a standard deviation ($d = 0.52$ for those who received the same form versus $d = 0.48$ for those who received an alternative form). They did notice a small difference for administration time; those who received the same form took slightly less time ($d = -0.02$) while those who received an alternate form took slightly longer ($d = 0.20$). The authors note that the examinees had no reason to expect to be retested with the same form—if same-form retesting were to become common, one could imagine greater exploitation by candidates. Also, the context of this study might be unique for two reasons: First, the exam was very easy to pass (slightly over 90 percent passed on the first try). And, second, the nature of most of the exam items involved scrutinizing medical images. Thus, candidates may not recognize items that they failed, and opportunities to memorize items and study them are fairly limited with this item type.

In a meta-analysis of many such studies of aptitude and achievement tests, Hausknecht and his colleagues (Hausknecht, Halpert, Di Paolo, Moriarty, & Gerrard, 2007) found that repeat examinees generally retested better, but those who retested using the same form improved twice as much as candidates completing an alternate form ($d = 0.45$ versus $d = 0.24$), although this difference decreased as the time between testing and retesting lengthened.

Very little is known about the effect of retesting on the validity of an assessment. Lievens, Buyse, and Sackett (2005) examined medical studies admissions tests and found that validity was higher for those retaking a knowledge test and passing on the second attempt than those passing on the first attempt. However, validity was lower for an intelligence test. The authors suggest that knowledge can be studied and so higher performance on the repeat administration of the knowledge test reflects greater learning, but higher scores on the intelligence test repeat administration simply reflects good luck, test-taking skills, and so forth that are unrelated to the criterion. More research, especially on the prediction of actual job behavior, is needed in this area. Candidates who retake an assessment can be expected to improve their scores. If the exam content can be “studied” then retesting may produce larger score changes and an identical form may inflate this effect.

Fairness of Technology-Enhanced Assessments

There is no doubt that web-based assessments expand the reach of organizations to access a larger, more diverse, candidate pool (Beaty, Grauer, & Davis, 2006). However, legitimate concerns have been raised that not everyone has the same access to, or comfort with, computer technology and this may impact assessment outcomes and introduce fairness concerns—along with measurement error (Tippins, Beaty, Drasgow, Gibson, Pearlman, Segall, & Shepherd, 2006). Since the application of technology-enhanced assessments will undoubtedly continue to expand at an exponential pace, it is prudent to examine the potential impact that this delivery option has on groups with limited access or familiarity with this sort of technology.

Internet Access and the Digital Divide

The “digital divide” is a term that is used to describe the gap between those individuals who have access to various telecommunications technologies and those who do not. The component of that technology that is most applicable to online assessments is high-speed Internet access or *broadband*. A recent

survey found that 63 percent of Americans have broadband at home and that broadband availability is available to more than 90 percent of households (Horrigan, 2009). This study also found that senior citizens and low-income Americans had the largest gains in broadband subscriptions between 2008 and 2009, while African Americans experienced a below-average broadband adoption growth rate in 2009, totaling 46 percent of current households.

Another study found that Internet usage rates (defined as occasional usage) vary by race: 71 percent of whites, 60 percent of blacks, and 56 percent of Hispanics. (Fox & Livingston, 2007). The rate for Spanish-dominant Hispanics drops to 32 percent. This study also examined the effects of education on Internet usage and found that Internet use is uniformly low for those who have not completed high school: whites (32 percent), Hispanics (31 percent), and African Americans (25 percent) and uniformly high (about 90 percent) for those who have completed college (Fox & Livingston, 2007). Internet usage rate also declines with age: 91 percent for 18 to 30 year olds, 90 percent for 31 to 42 year olds, 79 percent for 43 to 61 year olds, 56 percent for 62 to 71 year olds, and 29 percent for those 71 and older (Rainie, Estabrook, & Witt, 2007).

While it is obvious that Internet access and usage is becoming more widespread, it is still not universal and there do appear to be some race and age differences. This raises potential ethical and fairness concerns that the use of assessment technology could result in differential subgroup performance (Pearlman, 2009; Tippins et al., 2006). It is therefore incumbent on test users to review their situations and ensure that all candidates are treated fairly and afforded an equal opportunity to demonstrate their standing on the attributes being assessed. The demographic data on broadband access and use can be informative when analyzed against the targeted pool of applicants.

From the standpoint of the assessment, there are a number of ways to enhance familiarity with the technology and ultimately elicit the best possible performance from the candidate. For example, a tutorial can be incorporated that will show prospective candidates how to navigate through the screens and respond to the different types of test items. This tutorial should incorporate

sample items for each of the content areas being measured. An online narrator or “testing assistant” can also be programmed into the assessment to respond to common Q&As and even read the questions and responses if desired. In addition, help desk information should be made available to the candidates should they run into technical difficulties as they progress through the assessment.

For individuals who do not have access to, are uncomfortable with, or need special accommodations for online assessments, organizations should be prepared to offer supervised sessions whereby candidates can receive verbal instructions and assistance with technical issues. In the case of accommodation, it may be necessary in some circumstances to offer alternatives to the online assessment depending on the nature of the impairment (Tippins et al., 2006). Additionally, an equated paper-and-pencil version of the test could be made available if there is a large enough segment of the applicant pool that could benefit from this alternative.

While there is no doubt that, as access and familiarity with broadband increases over time, the impact of technology as a moderator of assessment performance will steadily dissipate. However, for the time being, organizations should analyze how the use of UIT or other technology-enhanced applications are impacting their candidate pools and target appropriate recruiting efforts to address any emerging gaps with the relevant labor market (Tippins et al., 2006).

Conclusion

The application of new technologies to the field of assessment has resulted in a tremendous amount of innovative practice and leading-edge research. Test users are able to leverage this assessment technology and supporting research to implement tools that are scalable on a global level and that can measure a broader array of attributes and behavior. Candidates can be assessed in multiple languages, in remote locations, for any level of job, and this can be accomplished with fewer items and greater precision than ever before. Technology has helped reignite the popularity

of assessment and, through its efficiencies and wide-scale application, can produce returns on investment that are hard for organizational leaders to ignore.

These advantages and large-scale applications have also led to unique challenges associated with established measurement practices. Some of these challenges, such as the use of UIT or the desire to shorten assessments, have resulted in innovative solutions such as CAT and the application of sophisticated theories such as IRT. However, there remain some basic measurement tenets that need to be applied universally, regardless of the content or technological medium within which the assessments are administered. In the rush to beat the competition in the war for talent, technology-enabled assessment systems may sometimes be “stood up” without the necessary attention to these measurement principles.

Whether the assessment system is purchased or developed from scratch, it still needs to be able to reliably measure the targeted attribute(s) and make accurate inferences about work behaviors. Any blueprint for building and implementing high-quality assessments must include at its core a thorough job analysis, detailed assessment plan, and well designed field research to establish the necessary psychometric and validation evidence. In addition, technology-enabled assessments must be implemented in a manner that ensures some level of standardization or the results may be suspect and of diminished value. Measurement error is tied to those irrelevant influences that interact with the test-taker, so every effort should be taken to establish and follow prescribed administration procedures.

Cheating, response distortion, retesting, and differential access to technology all impact the measurement accuracy of an automated assessment. It is therefore incumbent upon the test user to determine what level of impact each of these elements has in his or her own environment and to take appropriate action to address these sources of measurement error.

Good testing practice transcends the medium or application, and the best way to fully leverage emerging technology is to ensure such assessments have a solid measurement foundation.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Beatty, J. C., Grauer, E., & Davis, J. (2006). Unproctored internet testing: Important questions and empirical answers. Paper presented at the Practitioner forum conducted at the 21st Annual Meeting of the Society of Industrial and Organizational Psychology, Dallas, Texas.
- Bejar, I. I. (1983). Achievement testing: Recent advances. In J. L. Sullivan & R. G. Niemi (Eds.), *Quantitative applications in the social sciences*. (No. 07-036). Thousand Oaks, CA: Sage.
- Biskin, B. H., & Kolotkin, R. L. (1977). Effects of computerized administration on scores on the Minnesota Multiphasic Personality Inventory. *Applied Psychological Measurement*, *1*(4), 543-549.
- Boyle, S. (1984). The effect of variations in answer-sheet format on aptitude test performance. *Journal of Occupational Psychology*, *57*, 323-326.
- Breithaupt, K. J., Mills, C. N., & Melican, G. J. (2006). Facing the opportunities of the future. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances*, pp. 231-251.
- Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice*, *33*, 148-154.
- Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment*, *21*(2), 115-127.
- Bugbee, A. C., Jr. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, *28*(3), 282-99.
- Byrne, B. M. (1998). *Structural equations modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*(3), 253-260.
- Canoune, H. L., & Leyhe, E. W. (1985). Human versus computer interviewing. *Journal of Personality Assessment*, *49*, 103-106.
- Drasgow, F., Nye, C. D., & Tay, L. (2010). Indicators of quality assessment (pp. 22-59). In J. C. Scott & D. H. Reynolds (Eds.), *Handbook*

- of workplace assessment: Selecting and developing organizational talent.* San Francisco: Jossey Bass.
- Equal Employment Opportunity Commission, Civil Service Commission, United States Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, 43(166), 38290–38315.
- Fox, S., & Livingston, G. (2007). Latinos online: Hispanics with lower levels of education and English proficiency remain largely disconnected from the Internet [Electronic Version] <http://pewhispanic.org/files/reports/73.pdf>.
- Greud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement*, 10, 23–34.
- Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., & Pitoniak, M. J. (2002). Testing and measurement: Advances in item response theory and selected testing practices. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology* (3rd ed., Vol. 4, pp. 517–561). Hoboken, NJ: John Wiley & Sons.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., Moriarty, N. T., & Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. *Journal of Applied Psychology*.
- Holland, W. H., & Wainer, H. (1993). *Differential item functioning*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Horrigan, J. (2009). Home broadband adoption increases sharply in 2009 with big jumps among seniors, low-income households, and rural residents even though prices have risen since last year. [Electronic Version] www.pewinternet.org/Press-Releases/2009/Home-broadband-adoption-increases-sharply-in-2009.aspx.
- International Test Commission (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, 6(2), 143–171.
- Jodoin, M. G. (2003, Spring). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1–15.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer-Verlag.
- Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, 58, 981–1007.

- Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German job satisfaction survey used in a multinational organization: implications of Schwartz's culture model. *Journal of Applied Psychology, 89*(6), 1070–1082.
- Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (ETS RR-88–21). Princeton, NJ: Educational Testing Service.
- McPhail, S. M. (Ed.) (2007). *Alternative validation strategies: Developing new and leveraging existing validation evidence*. San Francisco: Jossey-Bass.
- McPhail, S. M., & Stelly, D. J. (2010). Validation strategies (pp. 671–710). In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Selecting and developing organizational talent*. San Francisco: Jossey Bass.
- Mead, A. D. (2010, April). Non-comparability of speeded computerized tests: Differential speededness? Paper presented at the Annual Meeting of the Society of Industrial and Organizational Psychology, Atlanta, Georgia.
- Mead, A. D., & Blitz, D. L. (April, 2003). Comparability of paper and computerized non-cognitive measures: A review and integration. Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, Orlando, Florida.
- Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458.
- Mead, A. D., Segall, D. O., Williams, B. A., & Levine, M. V. (1999, April). Multidimensional assessment for multidimensional minds: Leveraging the computer to assess personality comprehensively, accurately, and briefly. A paper presented at the twelfth annual conference for the Society for Industrial and Organizational Psychology, St. Louis, Missouri.
- Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*(2), 322–345.
- Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22*(1), 71–83.
- Ones, D. S., & Viswesvaran, C. (1998). The effects of socially desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245–269.

- Outtz, J. L. (2010). Addressing the flaws in our assessment decisions. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Selecting and developing organizational talent* (pp. 711–728). San Francisco: Jossey-Bass.
- Paek, P. (2005, August). *Recent trends in comparability studies*. PEM Research Report 05–05. Upper Saddle River, NJ: Pearson Educational Measurement.
- Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. van der Linder & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–148). Dordrecht: Kluwer.
- Pearlman, K. (2009). Unproctored internet testing: Practical, legal, and ethical concerns. *Industrial and Organizational Psychology*, 2(1), 14–19.
- Ployhart, R. P., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, 7, 27–65.
- Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. F. (2002, April). Web-based vs. paper and pencil testing: A comparison of factor structures across applicants and incumbents. Paper presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario.
- Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, 62, 337–354.
- Rainie, L., Estabrook, L., & Witt, E. (2007). Information searches that solve problems [Electronic Version] www.pewinternet.org/Reports/2007/Information-Searches-That-Solve-Problems.aspx.
- Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 156–188). San Francisco: Jossey-Bass.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), 517–529.
- Raymond, M. R., Neustel, S., & Anderson, D. (2008, March). The benefits of taking an identical version of a certification test on two occasions. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, New York.

- Reynolds, D. H., & Rupp, D. E. (2010). Advances in technology-facilitated assessment. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Selecting and developing organizational talent*. San Francisco: Jossey-Bass.
- Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*(5), 754–775.
- Rogers, J., Howard, K., & Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565.
- Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, *83*(4), 634–644.
- Ryan, A. M., Horvath, M., Ployhart, R. E., Schmitt, N., & Slade, L. A. (2000). Hypothesizing differential item functioning in global employee opinion surveys. *Personnel Psychology*, *53*(3), 531–562.
- Sacher, J., & Fletcher, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 403–419). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
- Stanton, J. M., & Rogelberg, S. G. (2001). Using internet/intranet web pages to collect organizational research data. *Organizational Research Methods*, *4*, 199–216.
- Stark, S., Chernyshenko, O. S. & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306.
- Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. J. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, *59*(1), 189–225.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

- Weiss, D. J. (2007). Adaptive—and electronic—testing: Past, present, and future. Invited address presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois.
- Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, *15*(4), 337–362.
- Zickar, M., Rosse, J., & Levin, R. (1996, April). Modeling the effects of faking on personality instruments. Paper presented at the 11th annual meeting of the Society for Industrial and Organizational Psychology, San Diego, California.

Chapter Three

IMPLEMENTING ASSESSMENT TECHNOLOGIES

Douglas H. Reynolds

If implemented effectively, technology-supported assessment systems can have substantial benefits: data about people can be collected efficiently, and better decisions can result from the insights generated from valid assessments. Technology can improve assessment in many ways, as the chapters in this book attest. Technology reinforces process consistency by guiding users through a workflow, thereby providing the benefits of standardization that are essential for good measurement and enhancing the procedural fairness that supports ethical and legal organizational decision making. These benefits do not accrue, however, if the implementation is poorly executed.

Mistakes are common when implementing new systems in organizations. For example, it is tempting to focus on the technological aspects of an implementation while underestimating the fact that people must change their behavior for the implementation to take hold. Managing the organizational change inherent with technological implementation is a likely prerequisite for success. The converse is also problematic; new technologies may be rolled out as a component of a well-planned change initiative, but if technical challenges are not anticipated and addressed, the initiative may fail as users become frustrated with poorly performing technologies.

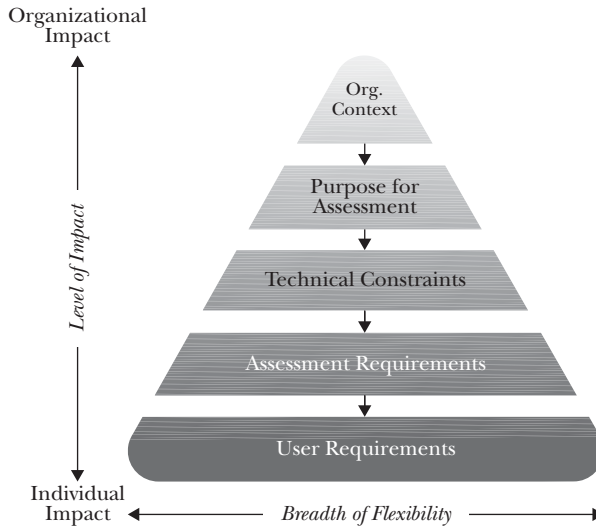
This chapter is structured on the premise that effective technology implementation involves the consideration and integration of multiple perspectives. Broadly considered, successful implementation of technology should recognize both the technological and the human issues associated with software deployment. Issues from each perspective should be examined and addressed to ensure that the technology is properly stitched into the fabric of the organization.

Strong implementations should begin with an analysis of the likely factors that will influence the deployment of software tools within the organization's context. The variables delineated in this chapter describe the major perspectives that should be examined as assessment technologies are implemented. These variables are not likely to apply in every situation, but it is important to consider each before deciding that its influence is minimal. Furthermore, these influences should not be considered in isolation to avoid overlooking the issues that arise when technology issues and human/organizational factors are interdependent.

Throughout the chapter it is assumed that the goal of effective software implementation should be the sustainability of the system within the organization, a criterion that has also been advocated by others (Kehoe, Dickter, Russell, & Sacco, 2005). The specific tools that are deployed will have their own intended benefits—an assessment that supports the hiring process should provide accurate insight into job candidates and lead to better selection decisions, assessments for development should spark targeted growth plans that have long-term impact. Regardless of the strength of these benefits, poorly implemented organizational programs will die on the vine before they show return. By playing close attention to the factors that drive sustainability, program implementers can put in place the foundational pre-requisites for the intended benefits of assessment software.

A Framework for Implementing Sustainable Technology

Before any attempt to develop or purchase assessment software is made, it is important to first understand the factors that will influence the effectiveness of the implementation. The framework shown in Figure 3.1 may be used as a starting point and roadmap toward understanding the context in which the assessment process

Figure 3.1. Implementation Framework

will operate. The figure also provides an organizing structure for the chapter, as each set of factors in the framework is described in more detail in the sections that follow.

This discussion assumes that an implementation team has accountability for the successful deployment of an assessment software system within an organization. In some cases, the implementation “team” may be a single person, typically for small programs of limited scope; however, the term “implementation team” is used throughout to signify the individual(s) who ensure that the software is deployed for the purposes intended by the organization.

For each level of the framework, implementation teams should consider three steps: first, identify the inputs that are necessary to understand the issues and requirements imposed by the variables at that level and who should best provide those inputs; second, define the options that are available for meeting the requirements and the implications of these options for other aspects of the implementation; and third, create outputs and documentation that describe decisions and recommendations associated with each level. Each of these steps is elaborated further as the levels are described.

The factors listed in Figure 3.1 are displayed to show the level of the major influences (for example, organizational to individual

impact) and the breadth of flexibility that is likely at each level. The length of the shapes shown in Figure 3.1 is intended to represent the typical level of flexibility that is likely for the conditions that exist in a large organization and for software that is readily available on the market.

Briefly stated, the critical variables that influence software implementation can be grouped into five categories, each representing a level in the framework. These categories include the organization's strategy within the larger environment (Organizational Context); existing or planned systems and processes for managing talent in the organization such as recruitment, selection, and development processes (Talent Systems and the Context for Assessment); the facilitating features and constraints imposed by the technology infrastructure, such as the bandwidth of the network and support for prerequisite software (Technical Facilitators and Constraints), the combined technical requirements suggested by the psychometric specifications and the software specifications for the system that will deploy the assessment (Assessment Requirements); finally, the user requirements and expectations for how the automated assessment process will operate (User Requirements). These variables will be examined in detail throughout the chapter.

Using the Framework

The framework can be used within any organizational environment by conducting an analysis of each category in the model. As shown in the figure, these variables may be grouped to reflect context, constraints, and requirements. The process for analyzing these variables within a given organization will be described as a Context, Constraints, and Requirements (CCR) analysis.

Just as it is necessary to conduct a job analysis to tie an assessment tool to job requirements and outcomes, it is important to analyze the critical aspects of the context, constraints, and requirements for assessment software to guide a successful implementation. The approach to this analysis can be simple (interviews with relevant stakeholders for each set of variables) or more complex (focus groups with users, feedback from reaction surveys). Used in this manner, the framework provides a template for defining the critical

issues, decision points, and design criteria that will be investigated as the implementation is planned. The purpose of the framework is to achieve four outcomes: (1) awareness building, (2) alignment and support, (3) planned flexibility, and (4) sustainability.

- *Awareness.* By working through the issues in the framework project implementers are able to better identify the multiple stakeholders to the envisioned system and their various perspectives and needs. Delineation of goals and constraints is another associated outcome.
- *Alignment and support.* By detailing facets of the strategic and operational environment, implementers can describe and reinforce essential alignments across organizational levels, units (departments, teams, etc.), and individuals. Strong alignment will help generate a cohesive and understandable vision of the system to help guide and motivate the various participants in the implementation process. Alignment is also critical when issues require support across organizational units and level to reach resolution.
- *Planned flexibility.* The framework is intended to reveal where various points of flexibility should be built into the system and where there are assumptions and limitations that constrain the system design. System flexibility also presents a paradox. As systems become more flexible and adaptable (for example, through software configurability), they become robust against changing needs and conditions, but they also become more complex and thus harder to implement. The art of effective implementation involves finding the right balance of constraints and flexibility points so the system is both manageable and adaptable.
- *Sustainability.* As noted, the ultimate outcome of effective implementation is sustainability of the software system over time and across fluctuations in business and organizational conditions. Assessment software that is widely adopted and used over time is a prerequisite to utility gains often cited for organizational assessments.

In sum, the effective implementation of software systems should be sensitive to the larger context within which they are deployed, respect the technical constraints operating in the organization, and

meet critical design specifications and user requirements. In most cases it is appropriate to investigate these conditions in detail before finalizing the psychometric and technical specifications for the new software system. The following sections review each level of the framework; key stakeholders, common issues, and typical options are noted for each.

Environmental and Organizational Context

An early step in the implementation of assessment software should involve the specification of the larger context and organizational rationale for the project. Organizations do not implement HR systems without good reasons to do so, and understanding these reasons and aligning tools and technologies to achieve desired outcomes is critical for effective technology implementation.

Understanding the Environment

The essential question at this level of the framework is how the technology implementation will support the strategy of the organization. Because strategy is typically developed in response to conditions in the larger environment, understanding these conditions is an important first step. Linkage to strategy has been recognized in other models that guide effective HR implementation (for example, Dorsey, 2002; Teachout & Hall, 2002); more elaborate models for ensuring the alignment of organizational strategy and technology strategy have evolved in the information technology literature (for example, Feurer, Chaharbaghi, Weber, & Wargin, 2000), and HR implementations benefit from these broader models also.

Understanding the linkage between environmental factors and organization strategy is critical for effective implementation planning. For example, organizations may need to respond to competitive pressures, new regulations, advancing technology, changes in resource and labor markets, and fluctuations in the economy. No organization is immune to its environment. Non-profit organizations may be affected by economic conditions just as for-profit businesses; similarly, educational institutions may be affected by demographic trends, funding levels, and competitive pressures from schools that recruit a similar student profile.

Healthy and adaptive organizations sense these environmental influences and set strategies to take advantage of the new opportunities provided by changing conditions. By documenting the connections between the environmental influences and the organizational strategies that justify investment in HR technology, implementation leaders are better able to align the efforts of the project team to the direction of the organization as a whole. This process also allows for the development of a strong business case for the effort, assuming the benefits of the technology are truly aligned with the organization's strategy. It also allows for critical questions to be asked if a misalignment with strategy is apparent. Projects that lack strategic alignment will often be subject to redefinition, derailment, or cancellation when the gap is exposed.

Although a conversation with the CEO might be an excellent method for understanding context and strategy, often other key stakeholders such as senior executives, HR leaders, line-of-business heads, and technology leaders can be just as valuable. Background research on organizational strategy can also be assembled using annual reports, strategic plans, and similar organizational communications. Based on this analysis, the implementation team should clearly state how the assessment will help facilitate organizational strategy. A rationale for the use of technology should also be evident from the work.

Organizational Characteristics

Various features of the organization can also play a strong role in determining the requirements for an assessment technology. Factors such as the size of the organization, its regional distribution, centralized or decentralized management structure, number of levels in the leadership hierarchy, and the industry in which it operates can all have an impact on technology requirements. For example, a globally distributed organization may have needs for a centralized software system that deploys assessments in multiple languages and tracks the results in a common database. A smaller organization with just a few locations within the same country may be able to address similar needs with stand-alone software at each location. Similarly, organizations with highly centralized management structures may be better able to design, implement,

and operate a global assessment process; decentralized organizations may have so many local requirements that separate systems for each region are more feasible and less expensive than attempting to stretch a one-size-fits-all solution to a broad range of needs.

Goals for the System

As noted by Fletcher (2005), the automation of HR functions has emphasized different benefits as the range of solutions has evolved. At a basic level, automation should drive efficient business processes. Beyond efficiency, automated systems should add insight into various aspects of the organization's operations and ultimately impact organizational strategy. At an early stage in the implementation process, the expectations and goals for the system should be declared. Tight alignment of these goals with the strategy and dominant characteristics of the organization should be assured to strengthen the rationale for the system.

An important output from this phase of the implementation is a vision statement for the automated system. The project vision may be a short document describing the strategic alignment with the broad goals of the organization, a brief description of system features, and the limitations of the system. A clear statement of what aspects of the system are to be considered in-scope and out-of-scope for the project should also be included. This list will help set boundaries for the specifications to be developed during later phases of implementation planning. Depending on the requirements of the organization, it may also be necessary to construct a financial business case to justify the expenditure for new technology. Business case analyses are common when justifying large-scale technology systems (for example, Miranda, 2002); these models may be adapted for use with the typically smaller-scale investments required to support assessment software.

Talent Systems and the Assessment Context

The next level in the implementation framework requires the implementation team to consider the direct purpose for which the assessments will be used. Technology-driven assessments have been deployed in a range of programs used to manage some aspect of an organization's talent pool. Assessments add value to

Example: Environment and Organizational Strategy Set the Context for Automated Assessment Systems

A large manufacturing organization was losing market share to competitors who were operating in lower-cost labor markets. Unwilling to close their domestic production facilities, the management team developed a strategy to lower production costs by using selective outsourcing and creating new efficiencies by implementing modern production technologies. If successful, the organization would be able to lower their costs while maintaining a competitive edge on quality and maintaining their loyal domestic market.

To drive this level of change, the organization realized it must develop and reinforce new leadership behaviors. Rapid innovation, driving efficiency, and managing constant change became top leadership objectives. The senior VP of human resources proposed a management assessment program for measuring key competencies related to the leadership objectives of innovation, efficiency, and change. An assessment platform was envisioned that would allow each of fifty-eight different locations to conduct an assessment of key competencies through a web-facilitated simulation. Resulting competency ratings would be held in a database that supports the generation of individual and group reports. Completion of leadership development events and multi-rater surveys would also be tracked in the system. The software was envisioned as a critical vehicle for communicating the new priorities to all managers, for tracking progress toward the development of new leadership skills, and for informing judgments about which areas of the organization held the strongest leaders and which needed additional support. The business rationale and the requirements for the system were described in a three-page document that served to gain the approval and funding for the effort. The document also provided a starting point for a request for proposals that was sent to potential vendors for the assessment technology.

organizational systems by providing accurate and objective information about people from which critical decisions may be based. Several summaries describe the variety of assessment technologies that may be deployed within the HR function (for example, Gueutal & Stone, 2005; Reynolds & Rupp, 2010; Rupp, Gibbons, & Snyder, 2008). The examples presented later in this book also provide ample evidence of the many ways that assessments are now delivered to users through technology.

The purposes for which assessments will be used have a strong impact on the nature of the assessments and on the technologies that support their deployment. Assessment purpose relates to the stakes involved with the process, as well as to the administration modes that are feasible, the roles of users, the expected outputs, and a variety of other features. Determining how the assessment will be administered is arguably the most important consideration at this phase of the implementation. Definitions of frequently used administration modes are shown in Exhibit 3.1.

The best choice of administration mode will be influenced by the purpose of the assessment and the stakes associated with the resulting decisions. An extensive discussion of the considerations involved with the mode of administration is provided by Tippins (2009) and a series of comments in the same volume. Some of the most common applications of assessment include applicant screening, job selection, and employee development programs. Implementation teams must determine the major technology implications for the purposes the assessments are likely to serve.

Assessments to Support High-Volume Screening

Many organizations have connected their job application process to the organization's website. These systems allow potential job applicants to review job information, complete screening assessments, submit a résumé, and/or complete a job application. When designed effectively, these tools allow for large numbers of applicants to be processed automatically so that the best applicants are identified quickly, flagged for recruiter review, and forwarded for additional steps in the selection process. Assessments used for this type of system should be amenable to exposure to high volumes of applicants so that the tool can be used as a first step in the application

Exhibit 3.1. Common Modes of Computer-Based Assessment Delivery

The International Test Commission's guidelines for computer-based tests (ITC, 2006) describe four categories of the most common modes of assessment administration. These categories vary based on the level of oversight and control asserted over the assessment process, including:

- *Open access.* The assessment can be accessed by the Internet from any location with no authentication of the user (that is, proof that the participant is who he or she claims to be) and no direct supervision of the assessment session.
- *Controlled delivery.* The assessment is made available only to known participants, but no direct authentication or supervision of the assessment session is involved.
- *Supervised delivery.* The identity of the assessment participant can be authenticated, and there is a degree of direct supervision over the administration.
- *Managed delivery.* The assessment session is highly controlled, often through the use of dedicated testing centers where there is oversight over authentication, access, security, the qualifications of the administrators, and the technical specifications of the computers used to deliver assessments.

process. Commonly, assessment systems of this type allow for the open access delivery mode.

Implementation teams should attend to the requirements imposed by assessment tools that are used for such a broad purpose. Screening assessments are often configured for each job for which they are used. For this reason, the technology system that supports these tools should have the flexibility to construct and deploy a range of question types and scoring rules. The assessment questions used for this purpose are often focused on basic job qualifications such as education levels, certifications, years of experience, and self-reported skills and knowledge. The scoring rules applied

to questions of this type might eliminate participants who don't meet pre-specified requirements, or they may simply give more credit to participants who possess a larger number of the qualities that are important for the work. Regardless of the specific scoring methodology, it is important that the specifications for the technology system include enough flexibility to accommodate the plan for how the assessment will be scored.

Assessments to Support Hiring Decisions

Contrasted with screening assessments, tests of knowledge, abilities, and traits are often used when making decisions among individuals who meet minimum qualifications to help determine who should receive job offers. These assessments tend to require more elaborate technology platforms to maintain appropriate testing conditions and security controls. Typically, these systems are administered under controlled, supervised, or managed administrative conditions, and the technology system must include supporting tools for the appropriate conditions.

For example, if the assessment is to be delivered under controlled conditions, an interface might be included for recruiters and HR staff to use for inviting participants into the assessment process. This feature might allow for generation of customized instructions that help introduce the participant to the requirements of the selection process. In contrast, if the assessment is to be supervised or managed (as is often the case with a high-stakes assessment), additional tools might be embedded that only allow the assessment to be delivered on pre-qualified computers within an established time window. Advanced systems may also include the capability to include a webcam for monitoring the participant or a biometric identification device, such as a fingerprint reader, to verify that the participant is not a surrogate. Implementation teams should therefore consider the security requirements associated with the assessments to be used in this manner.

Assessments to Support Employee Development

Leadership and employee development programs often include assessments to help participants understand their strengths and weaknesses and guide them toward development activities with the highest payoff. Technology-based assessments used for this

purpose can be administered under a range of conditions; however, controlled access is probably the norm.

Assessments used for development often impose different technology requirements than systems used for hiring because the range of assessment types is broader and the users are typically employees rather than applicants. For example, assessments used as part of a leadership development curriculum may include more in-depth measures, such as simulations, full assessment centers, and multi-rater surveys. Participants in development programs tend to be higher-level employees with expectations for the quality of the assessment experience, the level of insight provided by the assessment, and the usefulness of the outputs. These conditions place strong demands on the technology systems that support development programs (for example, broad bandwidth and special browser plug-ins may be required for the assessment to work correctly). The implementation team should identify these requirements early in the planning process so the technology systems can be appropriately configured to fit the intended purpose.

Plan for Additional Applications

As the purpose for the assessment is clarified, it is important for the implementation team to consider whether the usage may change over time. If, for instance, an assessment system designed to support a development program is later used to inform promotion decisions, a new set of technology requirements may emerge. The system may need to deploy a parallel form of the assessment so that participants don't see the same assessment twice, access to outputs and reports may need to be restricted to hiring managers only, and scores used for development would need to be tracked separately from those used for promotion. Use of a technology platform that does not allow for these flexibilities may lead program administrators to work around a system that is poorly suited for its desired purposes. Work-around processes can damage a program when, to continue the example, participants learn that an assessment they took for purely developmental reasons is now being used as an input to their promotions. Mid-stream changes in the purpose for an assessment process can affect the credibility of the larger talent program; ensuring the

technology platform includes the necessary flexibilities to handle the likely range of purposes will help to avoid this problem.

The outputs to be created by the implementation team for this level of the framework can include a specification of the talent processes that will use the technology-based assessment. The specification should include a description of how the assessment will add value to the process, the types of assessments that could be included in the process, and the technical requirements associated with the appropriate mode of assessment administration.

Technical Facilitators and Constraints

In the next level of the framework, the implementation team must investigate the available technical environment and consider how its characteristics will affect the deployment of the new program. Beyond the software itself, there are a variety of factors that influence how an automated assessment system will operate. For example, the physical network infrastructure and configuration, processes for maintaining data integrity and security, the availability of support, and the requirement to integrate with other software-based systems are important to consider as the software is implemented. Each of these facets of the technical environment should be adequately investigated before preparing an implementation plan. Depending on their characteristics and their interaction with the software requirements, these facets may operate as facilitators or constraints for the assessment process. A few of the most prominent technical variables are briefly described in the following sections.

Infrastructure Characteristics

An important source of influence on the implementation of assessment software stems from the technical infrastructure upon which the software must operate. Incompatibility between the software and these aspects of the system will reduce the chances that the software will meet its goals. By analogy, consider the software system to be akin to the cars on a railway train and the hardware and network infrastructure are the rails upon which the cars will ride. It is important to note that even though infrastructure characteristics are considered to be constraints, it is also possible to

change these features. However, just as it is easier to change railway cars than it is to change the direction of the track, in most cases it is considerably less difficult to adjust the software requirements to fit the technology infrastructure than vice versa.

Examples of infrastructure considerations include the connection bandwidth required by the assessment, the configuration of the network firewall, and the capabilities of the computers to be used for the assessment.

Bandwidth: Do the Internet connections along the route to the assessment participant support the demands of the assessment content? For example, if an assessment contains video clips, does the assessment location have broadband access? How many computers will be sharing the same incoming connection? If assessments are to be delivered on a single computer on a scheduled basis, then the connection speed may not be an issue; however, if several computers are used at the same time, it is possible that the speed will be dramatically slower as the participants view large content files at the same time. Conversely, has the content and deployment system been designed to be compatible with the likely connection speed of the end-user? Systems used for large-scale recruitment are often designed for applicants with dial-up connections to ensure the process is inclusive.

Firewalls: Firewalls are designed to monitor the flow of information moving across a network and to exclude information that does not meet pre-defined rules or criteria. These rules could involve the exclusion of content with various keywords, network traffic that attempts to use certain communication ports, or information that has been sent from an unapproved source. As assessment systems are implemented, network engineers who control access to the local network should be consulted regarding the prevailing firewall restrictions. Firewalls are an important consideration any time an assessment is to be deployed internal to an organization, such as when assessments are administered for development or promotion to current employees.

Local access computers. The computer to be used by the assessment participant must also be configured to handle the requirements of the assessment software. Occasionally the assessment system may require a minimum processor speed or peripheral hardware such as speakers and a microphone. However, more typically it is the

software on the local computer that presents compatibility concerns. Checking the compatibility of the Internet browser and plug-in extensions, such as Flash or Silverlight, with the requirements of the assessment software can save frustration later in the deployment. It is not uncommon for organizations to delay upgrades of browser software to ensure compatibility with existing systems, so assessment software that requires the latest browser versions may not operate properly on the intended computers.

Software Deployment Model

Another important facet of the technical environment relates to how the software will be deployed. There is an interaction between the technical features of the network and the deployment model for the assessment software. Software might be deployed as a stand-alone program, installed locally on the computers that run it. Although this deployment model is not as common as it used to be, it may still be appropriate for simple assessments that require little maintenance, are rarely upgraded, and do not require scores to be integrated with broader systems. Questionnaires used for self-insight or tests of basic skills that are used as a screening tool for small applicant pools might be deployed in this manner.

More commonly, assessment software is deployed on centralized servers and accessed by the users over the Internet. This configuration can be set up within an organization and made available only to users within the same network (so-called “behind-the-firewall” installations) or on servers operated by a third party and accessed by users over the Internet. Third-party hosting may be accomplished through a variety of models such as systems that are dedicated to a particular client organization or software-as-a-service models whereby many clients share the same hosted software system.

An important distinction between these deployment models is how they are maintained and supported over the lifetime of the deployment. Stand-alone software is the most difficult to maintain, because the software must be upgraded on each computer on which it resides. Hosted software is easier to maintain; however, if the software has been specially built or configured for a single client organization, maintenance costs tend to be higher. When software-as-a-service models are used, the support and maintenance

tend to be less burdensome because many of the software provider's clients reside on the same system. A change made centrally to upgrade the software is thus applied to all clients at once. Implementation teams must consider the viable deployment models before deciding between software systems, because software providers do not typically support all deployment models. The choice of a deployment model is thus linked to the choice of whether to build or buy software—a topic to be explored later in the chapter.

Technical Support

Assessment software can require various levels of technical support. For nearly all systems, some level of support should be available for the technicians who install and operate the system. Who can clients turn to if they discover undocumented incompatibilities with existing software on the assessment computers? What is the process for handling technical errors experienced by users? Typically, software providers have support staff available to handle issues of this nature. A common configuration is to employ generalists who are trained to handle recurring problems at the initial contact point and a well-defined escalation rule for issues that are to be handled by the engineers who wrote the software.

End-user support is also important for many systems. For large-volume assessment tools, end-user support is often provided by on-site administrators when the assessment is deployed in a supervised or managed (proctored) environment. User support for open access and controlled delivery systems (unproctored) must be handled differently because users are remote and the volume of users can be huge. These systems should be designed to be intuitive and simple to use, the software should be robust to common variations in user computer configurations, and on-screen help should be available that guides users through common issues. When user help is required, the client organization often prepares its recruiters or HR staff to handle simple problems and the software provider maintains a support center to handle problems that are escalated from the client's staff.

Special conditions exist when a system is used internationally. In these circumstances it may be necessary to make arrangements to have support staff available who are able to address

issues in the language of the user. This may necessitate the use of bilingual support personnel who are able to address the user in his or her native language and also communicate effectively with the engineers in situations in which issues must be escalated for technical evaluation. The desired support model should be included in the implementation plan. Typically, the specific support services to be offered by the software provider will be described in a service-level agreement (SLA) attached to the contract with the provider.

Adjacent Technology Systems

The requirements of other software systems may also impose constraints on the technical aspects of the assessment system. Many HR systems are moving toward larger constellations of software tools that have been configured to manage organizational talent across multiple phases of the employee lifecycle. For example, recruiting and selection tools may be connected directly to the organization's human resources information system (HRIS) so that a record may be automatically created for employees after they are hired. This system may also be connected to a process for setting performance targets and appraising performance at defined intervals. If the assessment system to be deployed must integrate with other software tools, the requirements for this integration should be investigated as the software implementation is planned.

Several models for integration are possible, each posing different technical requirements. A simple method of integration involves the exchange of information between software systems. For example, a system that collects online job applications may trigger an assessment for the most qualified candidates by transferring a candidate identification number to the assessment system. This information allows the assessment to recognize the candidate as an approved user when he or she logs in. Upon completion of the assessment, results may be transferred to an applicant tracking system where the next step in the process may be initiated. Each integration step in this example may be accomplished merely by transferring information in a predefined format, between the components of the process.

More elaborate integrations may involve the exchange of instructions between systems. Consider in the prior example, if the

assessment were to vary based on answers to an earlier job application, such as the possession of a college degree or a technical certification. These responses might qualify the applicant to take a higher level of assessment, and in turn be considered for higher-level jobs. In this instance, the assessment system alters the way it operates (by providing an alternative assessment) based on instructions from the job application system. System interoperability can become complex, potentially driving a range of alterations in how the software appears to the user. In many cases these alterations are invisible to the user, allowing several subsystems to appear as one.

When implementing multiple software systems that must work together, it is tempting to assume that the integration will be straightforward; however, this assumption should be avoided in most cases. Although software systems often adhere to standard integration approaches (for example, they may use HR-XML, a standard that provides a common language for database fields and software functions), many do not. Detailing the integration requirements should be a step in the investigation of technical constraints and facilitators. A more detailed treatment of the options and methods involved in integration for assessment systems can be found in Reynolds and Weiner (2009).

Technical Requirements for the Assessment

Ideally, the software requirements imposed by the assessment itself should be considered after the organizational context and technical constraints are known. Once the context and purpose for the assessment are clear, specifications for the assessment may be gathered by the implementation team. The options are wide-ranging, depending on the type of assessment to be deployed. Multiple-choice assessments used for screening of job applicants for high-volume jobs may have little in common with online assessment center delivery systems. Each type of assessment may in turn have many different configuration options. It is important that all requirements specific to the tool be gathered so that the technology system can be chosen or constructed to support them. Table 3.1 shows common options for the delivery of online assessments.

Many software functions are available to support assessment deployment; the list shown in Table 3.1 is intended to provide a

Table 3.1. Common Features of Online Assessments

<i>Feature Set</i>	<i>Purpose</i>	<i>Common options</i>
<i>Content Display</i>	Presents assessment stimuli to participants and collects responses	Multiple-choice test questions; graphical response methods (for example, drag-and-drop controls); free-text response.
<i>Timing</i>	Restricts the time that assessment content is available to participants to pre-defined limits	Ability to time the delivery of individual questions, subtests, and entire tests; ability to display or hide the timer to participants; ability to record the participant's completion time.
<i>Content Rotation</i>	Allows for the presentation of alternate content across participants	Support for the deployment of parallel forms of the assessment, randomization of the delivery of test forms, presentation of experimental, un-scored, assessment content.
<i>Advanced Test Construction Support</i>	Deploys randomized, adaptive, and other forms of rule-based question delivery	Randomization of question presentation within subsection, support for adaptive delivery of questions (for tools developed from an item response theory model), support for branched presentation (where questions are delivered in a logical order based on prior responses).
<i>Branding and Appearance</i>	Provides flexibility to conform the screen to organizational requirements	Ability to change screen colors, adjust layout features such as the location of title bar and menus, and to add logos.
<i>Practice Questions and Online Help</i>	Introduces the participant to the assessment platform and provides assistance when issues are encountered	Support for short videos or onscreen animations that instruct the participants on the use of the assessment platform; practice questions with feedback; online help text for common questions; online chat with system administrators; phone support for unusual issues.

(Continued)

Table 3.1. Common Features of Online Assessments (Continued)

<i>Feature Set</i>	<i>Purpose</i>	<i>Common options</i>
<i>Security</i>	Ensures that participants do not inappropriately gain access to assessment content	Inclusion of role-based access restrictions to administrative functions; ability to set up specific appointment times for participant access to the assessment; proctor functions such as assessment reset and cancellation; access and monitoring tools such as fingerprint readers and webcams.
<i>Disability Accommodation</i>	Provides adjustments to allow for the accommodation of disabled participants	Online options for requesting accommodations; administrator control over the assessment timer to allow for appropriate time extensions; compatibility with screen-reader programs.
<i>Data Recording and Reporting</i>	Captures relevant participant responses and generates score reports	Support for recording and storing participant responses as well as other assessment-relevant behaviors (response latency, incomplete/unviewed questions, etc.); generation of scale scores, test scores, and decision criteria; production of end-user (for example, participant, recruiter, hiring manager) reports and aggregate summaries of participant performance.
<i>Data Archiving, Extraction, and Deletion</i>	Allows for database maintenance and data export for other purposes	Tools for setting archiving rules, processes for reactivating archived records; access points for the generation of data extracts for research purposes (for example, export of participant responses to each question as well as computed scores); flexibility to delete participant records when deletion criteria are met.

sampling of common options. Specific requirements should be determined through consultation with the experts who designed the assessments. Future research and development will certainly add more features to this list, and new features should be researched to ensure that they do not unduly influence the behavior to be assessed before they are implemented for operational use.

The output from this level of the implementation framework should include a set of specifications for the assessment. These specifications should consider the manner by which the assessment will be deployed alongside of the requirements for the specific assessment tools. Implementation teams often develop lists of must-have and nice-to-have features for the assessment delivery system as an outcome of this level of implementation planning.

User Requirements

Planning for the needs of various user groups is the final category of requirements to be determined as the software is designed or configured. This level of the framework requires the implementation team to define user roles, understand the requirements associated with each role, and ensure that the necessary features for each role are included in the software.

For many assessment programs used in organizations, a range of roles can be considered. Well-designed software will support the common roles associated with the operation of an assessment program as well as allow for the construction of unique user roles. This flexibility is often essential because new roles will emerge as a tool is used. Ideally, configuration options are available within the software that allow for new roles to be constructed. Roles are created by providing access to specific functions within the software through a user-rights management system. By assigning the ability to either view or edit a particular page, the system administrator can support new roles as the need arises.

For example, in an assessment system that is used for recruiting across multiple regions, a new region may adopt the use of the assessments, thereby necessitating the creation of a new administrator role. This role is created by assigning user rights to the specific assessment tools to be used in the region. Further, suppose this new region is very small, and the local hiring manager also

serves as the test administrator; user rights are then configured such that the functions associated with test administration (candidate invitations, test reset and retake controls, etc.) and the functions associated with the needs of the hiring manager (candidate test reports, interview guides, etc.) are combined into just one role for this region.

Ascertaining user requirements must be done carefully. Techniques for requirements gathering can include interviews and focus groups with user groups (for example, recruiters, test administrators, hiring managers, and candidates). As user groups are interviewed, it is best to provide the subjects with at least a general understanding of the contexts, constraints, and other requirements for the system. If this is not done, users may develop unrealistic expectations for the system, and these unmet expectations may lead to active resistance as the system is implemented.

The range of user roles and functions to be deployed in assessment software is closely associated with the purpose of the assessment. Table 3.2 shows common purposes, the associated roles, and required functionality for each role. Once the roles for the system are understood, the implementation team should summarize the most essential roles that are envisioned for the system and the configuration flexibilities required to support new roles that may emerge as the system operates.

Executing the Implementation Plan

Once the implementation team has completed an investigation of the context, constraints, and requirements (that is, the CCR analysis) for the assessment software, a plan should be finalized that contains the critical results from the analysis. This planning document should contain the results from the investigation of each level of the framework. The plan will include a description of the vision for the system, how it fits within the talent management process, any significant technical constraints, and the requirements imposed by the desired assessment and the various users who will have a role in its operation. The documentation of these results need not be elaborate. Often a simple, clear statement of the major findings at each level of the framework is sufficient. In fact, elaborate specifications can often become a barrier to rapid development. New models

Table 3.2. Example User Roles and Associated Technology Requirements, Shown by Assessment Purpose

<i>Purpose for Assessment: High-volume online job application and screening</i>	
<i>Sample Roles</i>	<i>Associated Technology Requirements</i>
<i>Administrator</i>	User-configured questions and scoring Tools to post job openings and recruiting requisitions Tools to create new system roles
<i>Candidate</i>	Access to assessments after submission of an application that meets pre-defined criteria Features that allow candidates to update their profiles after initial submission Access to message board where recruiters and hiring managers may leave confidential messages
<i>Recruiter</i>	Tools for inviting participants to the assessment Dashboards where candidate progress on the assessments can be tracked Access to score reports and aggregate summaries that indicate the effectiveness of various recruitment sources
<i>Purpose for Assessment: Internet-based testing to support hiring decisions</i>	
<i>Sample Roles</i>	<i>Associated Technology Requirements</i>
<i>Proctor</i>	Features for restricting access to the tests to only pre-registered candidates Tools to restart a test if a problem occurs Ability to adjust test delivery conditions (time limits, font sizes, etc.) to accommodate a test-taker disability
<i>Hiring Manager</i>	Access to candidate test results Access to interview support tools Access to a questionnaire about recruit quality and new hire performance

(Continued)

Table 3.2. Example User Roles and Associated Technology Requirements, Shown by Assessment Purpose (*Continued*)

<i>Purpose for Assessment: Internet-based testing to support hiring decisions</i>	
<i>Sample Roles</i>	<i>Associated Technology Requirements</i>
<i>Test Researcher</i>	<p>Ability to export question responses for psychometric analysis</p> <p>Tools to change scoring rules for future participants</p>
<i>Purpose for Assessment: Assessments for leadership development</i>	
<i>Sample Roles</i>	<i>Associated Technology Requirements</i>
<i>Participant</i>	<p>Access to multiple assessments within the program such as multi-rater surveys, simulations, and tests</p> <p>Access to templates for creating an individual development plan (IDP) and associated development resources, such as web-based training courses</p>
<i>Mentor/Coach</i>	<p>Ability to provide secure access to the IDP to a designated coach or mentor</p> <p>Access to participant development plans, when permission has been granted</p> <p>Ability to enter comments and suggestions into the IDP</p> <p>Ability to recommend participants for high-potential pools or other succession management programs</p>
<i>HR Program Manager</i>	<p>Administration tools for tracking the progress of a cohort through the development program</p> <p>Ability to link the assessment program to a learning management system to allow participants to register for courses associated with development needs</p> <p>Ability to assign coaches/mentors to program participants</p>

for writing software such as agile development (Martin, 2002) may only require a high-level vision and specification to guide the initial development effort.

Once the plan is complete, the implementation team will focus on execution of the plan. Two important challenges for the team will be deciding how the software will be acquired and managing the implementation itself.

Acquire the System: Build, Buy, or Rent

Ideally, the factors in the implementation framework have been analyzed before a specific software system is purchased or developed. From the consideration of the environmental and organizational influences that drive the need for automation to the specific requirements of each type of system user, these factors all play a role in the decision regarding how to acquire the assessment software. If the software is to be built specifically for the organization, the results from the CCR analysis can be used as a set of functional specifications for the developers to use as they design the software. If the software is to be acquired from a third party, the CCR results may be easily turned into a request for proposals and a set of criteria that may be considered as alternative systems are evaluated. These options for acquiring the software have different risks, benefits, and costs, so the decision regarding how to proceed should be carefully deliberated.

Building Custom Software

The benefits of this approach are clear: those who have the resources to build and deploy their own software are able to construct features that fit their needs with precision. They are not beholden to another organization for maintenance and support, they can ensure compatibility with other organizational systems and processes, and they retain ownership rights over the end product so ongoing fees such as licensing and hosting costs can be kept to a minimum. The downsides of custom software can be severe, however.

Often custom software is not optimal until several version releases are complete. The first iteration that is deployed is frequently met with a variety of user requests for system enhancements;

hard-to-identify bugs also emerge as the number of users expands. Need for additional versions can be problematic when the engineers who created the initial version have moved on to other projects. Designers of custom software should plan for several versions, because experience with the system will reveal necessary enhancements required for the tool to sustain in the organization. Long-term maintenance can also be an issue as the surrounding technical environment changes (for example, as new browsers are released); the software must be kept up-to-date or eventually it becomes obsolete. If maintenance releases are infrequent, the organization runs the risk that the original programmers have left and the necessary skills to update the system will be unavailable. Under these conditions, many organizations choose to deploy software that is built and maintained by third-party providers.

Acquiring Third-Party Software

Alternatively, organizations can often acquire software from vendors who will configure their product for specific needs. In these cases it is wise to stay within the capabilities of the vendor's standard product, because customization of vendor software leads to the same disadvantages as those described for custom-built software. Well-designed software should allow client organizations to configure their specific application within common boundaries of practice. If the vendor has a good understanding of the requirements imposed by assessments and of the common configuration needs across organizations, the software should allow enough flexibility to meet fundamental needs. Unless a particular assessment need is unique, it is often best to find ways to meet the requirements within the configuration options available in the vendor's standard platform. On rare occasions, ownership rights to a copy of the software may be provided to the organization; more typically the software is run in an environment hosted by the vendor, with the client organization "renting" the software by paying periodic fees for licensing, hosting, and maintenance. This approach has the advantage of centralizing the responsibility for maintaining and upgrading the software within the vendor organization, where there is a presumed financial motive to keep the platform competitive.

Managing a Successful Implementation

When software implementations fail, it is often for reasons that bear little direct relation to the technology itself. Rather, poor implementation is likely to result from a failure to understand the human side of the organizational change required to sustain a software-driven business process. Organizational psychologists are well suited for understanding the prerequisites for effective change, and extensive literature exists on the implementation of organizational change and innovation in organizations (cf. Hedge & Pulakos, 2002).

Managing organizational change in the context of a software implementation requires attention to several factors. Building ownership, accountability, and alignment of roles, as well as effective communication across constituents, is essential for effective implementation. Measurement of progress along the way is also important for ensuring success.

Stakeholder involvement is paramount: as noted throughout this discussion, obtaining input from those involved in the myriad aspects of a software deployment is essential for success. Selling the project at the top and getting the support needed to finance and prioritize the work will require input from senior management. Organizational leaders can help identify the linkages between environmental trends, business strategy and direction, and the specific drivers behind the use of assessment and the need for automation (Shupe & Behling, 2006; Weill, Subramani, & Broadbent, 2002). They also serve as visible role models when the time comes to vocally support the changes required for the new program to succeed. An example of stakeholder support in one organization that adopted an assessment process is shown in the case study below.

Beyond the senior stakeholders, each step in the implementation framework suggests important coordination points between the implementation team and other organizational resources required to support the process. Information technology and support staff, recruiters, HR administrators, and hiring managers are just a few of the functions in an organization that may take on new responsibilities as the assessment system is implemented and operated. For each group, a clear specification of their role, defined accountabilities, and adequate training will be required to ensure the assessment process operates effectively. Research has shown that training alone does not ensure successful implementation; users must also

be aware of how they will be supported in their roles as the software is deployed (Marler, Liang, & Dulebohn, 2006).

Measurement of progress is also essential to a successful implementation. By defining interim milestones and both proximal and distal outcomes, the implementation team can chart their progress and make necessary adjustments along the way.

Implementation Case Study: Assessment for M&A Integration

When two large banks merged, a team was formed to manage the integration of the two workforces. The integration team settled on the deployment of an assessment to be administered to all current employees in both companies whose role could be affected by the merger. The assessment was designed to provide a common and objective measure that allowed information about people to be incorporated into the decision-making process. Sensitivity to the new assessment was acute because of the high stakes associated with creating a combined organization.

To ensure the proper message was sent to all employees, the most senior leaders in each function completed the assessment process before asking their subordinates to participate. This experience allowed senior management to assist with the formal communication about the steps in the process, how the information would be used, and the limitations and restrictions associated with the resulting data. Informal communication from these senior stakeholders was also a benefit, because they could share their personal experience with skeptical employees. By coordinating the communication from senior leaders with the rollout of the assessment program, the implementation team was able to reinforce accountability across the organization. When pockets of resistance were identified, senior managers could be enlisted to help support the program by clarifying the organizational commitment to using the process in a consistent manner. Through this process, the implementation team was able to deploy assessments to support the decisions required to combine the two workforces.

Proximal measures (measures of variables that are likely to be impacted soon after the implementation of technology) could include reactions from users in each of the roles required by the system. For example, applicant reactions to assessment and selection processes are often evaluated to gauge perceptions of fairness, appropriateness, ease of use, and perceived validity. The study of the factors that influence applicant reactions has been a growing literature in recent years (see Chapter 6 in this volume, for a current overview). When assessments are deployed using technology, a wider range of questions should be asked of other user roles also (for example, system administrators, proctors, recruiters, hiring managers). Other common proximal measures of system effectiveness are shown in the left side of Table 3.3. Proximal measures are usually easier to collect than more distal measures, and they are critical for a meaningful evaluation of the assessment process because they should reflect the influence path to the distal variables that show organizational impact.

Table 3.3. Common Criteria Used to Evaluate the Effectiveness of Automated Assessment Processes in Organizations

<i>Proximal Measures</i>	<i>Distal Measures</i>
Completion rates for the assessment	Performance/productivity of employees after six months to one year on the job
Volume of technical support calls	Turnover and tenure rates among employees hired using the assessment process
Time to complete the assessment	Promotion rates, time-in-grade, and similar measures of growth within the organization
Number of interviews conducted by hiring managers	Engagement and organizational commitment levels of new hires
Time to hire new employees	Unit-level performance for groups using the assessment, compared to those who do not
New hire satisfaction with the hiring process	
Performance in training	
Manager satisfaction with their new hires	
Time-to-productivity for new employees	

Distal measures often provide evidence that will help justify the value of the implementation as well as to guide future adjustments (see the right side of Table 3.3 for examples). Although distal measures have the disadvantage of being affected by many variables beyond the assessment itself (for example, economic and environmental factors that influence the organization's performance), they tend to be important variables to stakeholders. Distal variables are also likely to be associated with a monetary benefit, and many implementation teams attempt to communicate their findings in terms of dollars. However, this approach can have serious drawbacks because utility models often result in large numbers that beg for skepticism. An alternative approach is to focus on the benefits associated with the implementation, where the influence path from the use of the assessment to a distal measure is most clear. Stakeholders can then easily convert the results to financial benefits themselves, thereby enhancing their buy-in to the results.

Measures of implementation progress should demonstrate the intended benefits of the assessment, such as process efficiency, improved insight when making decisions about people, and strategic impact on the organization. The effectiveness of the implementation process should also be judged based on the use of the program. Adoption can be tracked by evaluating such factors as the number of departments or regions that use the software, the number of users in each role, and the number of exceptions to the use of the software (a negative indicator).

Concluding Thoughts

Technology-based assessments can have a strong payoff in organizations if they are implemented in a manner that allows them to sustain. By approaching the software deployment as an organizational change, and by using a structured framework to understand the environment surrounding and supporting the software, organizational psychologists are well-suited to guide effective implementations. The framework presented in this chapter, along with an analysis of the context, constraints, and requirements for assessment software, is intended to provide a starting point for practitioners who are involved with these implementations.

Implementing assessment systems that are to be a regular component of the process for managing talent in an organization can require substantial effort. Looking beyond the first-release jitters and bugs that are common in software deployments, assessment programs that do not adapt as the organization changes won't survive to generate the benefits they were intended to provide. To remain sustainable, the processes involved with implementation should continue on a regular basis. Technology changes rapidly, and the environmental conditions that provide the justification for the use of assessments can change even faster. To maintain a software-based assessment over time, it is necessary to reevaluate the context, constraints, and requirements for the software and make necessary adjustments. Calibration to the larger environment allows the benefits of assessment to accrue as long as the program is sustained.

References

- Dorsey, D. W. (2002). Information technology. In J. Hedge & E. Pulakos (Eds.), *Implementing organizational interventions: Steps, processes, and best practices*. San Francisco: Jossey-Bass.
- Feurer, R., Chaharbaghi, K., Weber, M., & Wargin, J. (2000). Aligning strategies, processes, and IT: A case study. *Information Systems Management, 17*, 23–34.
- Fletcher, P. A. K. (2005). From personnel administration to business-driven human capital management: The transformation of the role of HR in the digital age. In H. Gueutal & D. Stone (Eds.), *The brave new world of eHR: Human resources in the digital age*. San Francisco: Jossey-Bass.
- Gueutal, H. G., & Stone, D. L. (Eds.). (2005). *The brave new world of eHR: Human resources in the digital age*. San Francisco: Jossey-Bass.
- Hedge, J. W., & Pulakos, E. D. (Eds.). (2002). *Implementing organizational interventions: Steps, processes, and best practices*. San Francisco: Jossey-Bass.
- International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*, 143–172.
- Kehoe, J. F., Dickter, D. N., Russell, D. P., & Sacco, J. M. (2005). e-Selection. In H. Gueutal & D. Stone (Eds.), *The brave new world of eHR: Human resources in the digital age*. San Francisco: Jossey-Bass.
- Marler, J. H., Liang, X., & Dulebohn, J. H. (2006). Training and effective employee information technology use. *Journal of Management, 32*, 721–743.

- Martin, R. C. (2002). *Agile software development: Principles, patterns, and practices*. Upper Saddle River, NJ: Prentice Hall.
- Miranda, R. (2002). Needs assessment and business case analysis for technology investment decisions. *Government Finance Review*, 18, 12–16.
- Reynolds, D. H., & Rupp, D. E. (2010). Advances in technology-facilitated assessment. In J. Scott & D. Reynolds (Eds.), *The handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent*. San Francisco: Jossey-Bass.
- Reynolds, D. H., & Weiner, J. A. (2009). Online recruiting and selection: Innovations in talent acquisition. Malden, MA: Wiley-Blackwell.
- Rupp, D. E., Gibbons, A. G., & Snyder, L. A. (2008). The role of technology in enabling third-generation training and development. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 495–499.
- Shupe, C., & Behling, R. (2006). Developing and implementing a strategy for technology deployment. *The Information Management Journal*, 40, 52–57.
- Teachout, M. S., & Hall, C. R. (2002). Implementing training: Some practical guidelines. In J. Hedge & E. Pulakos (Eds.), *Implementing organizational interventions: Steps, processes, and best practices*. San Francisco: Jossey-Bass.
- Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial-Organizational Psychology: Perspectives on Science and Practice*, 2, 2–10.
- Weill, P., Subramani, M., & Broadbent, M. (2002). Building IT infrastructure for strategic agility. *Sloan Management Review*, 44, 57–65.

Chapter Four

CHEATING AND RESPONSE DISTORTION ON REMOTELY DELIVERED ASSESSMENTS

Winfred Arthur, Jr.,* and Ryan M. Glaze

In this chapter we provide the reader with a succinct review of the state of the literature on cheating and response distortion on remotely delivered ability and knowledge, and nonability and noncognitive assessments, respectively. This review and associated discussion is organized around five questions: (1) What is cheating and response distortion? (2) How pervasive and extensive are they? (3) How are they detected and how effective are said detection techniques? (4) What does one *do* with either the information, cheaters and dissimulators, or both, once they have been detected? and (5) How does one deter these behaviors (cheating and response distortion) and how effective are

*Correspondence concerning this chapter should be addressed to Winfred Arthur, Jr., Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843-4235. Email should be sent to w-arthur@neo.tamu.edu.

said deterrence techniques and methods? (It is important to note that we do not discuss the effect of cheating and response distortion on the psychometric properties of remotely delivered assessments; for a review of these issues, the interested reader is referred to Chapter 2 by Scott and Mead on the Foundations for Measurement in this volume.) We conclude the chapter with a few summary statements and a brief discussion of directions for future research. One such important summary statement is that the proctored versus unproctored delivery of assessments has more profound effects and implications for ability and knowledge tests than it does for noncognitive and nonability measures primarily because proctoring is not a means for controlling for response distortion on the latter type of assessments. Another equally important summary statement is that the amount of research on unproctored Internet-based ability and knowledge tests is very limited, and given the concerns about cheating on unproctored Internet-based ability and knowledge tests, it is definitely not commensurate with the use of these types of assessments in practice.

For the purposes of this chapter, we use the terms “remotely delivered assessments” and “unproctored Internet-based tests” synonymously and interchangeably. Within this nomenclature, collectively, remotely delivered assessments are typically characterized by the absence or a reduced level of proctoring—that is, the presence of a human proctor to ensure that test-takers follow and abide by (ethical) testing rules and standards. This characteristic justifiably raises misgivings about the accuracy and validity of test scores obtained from this mode of testing because in the absence of a proctor or proctoring, there are concerns that test-takers, left to their own devices, are likely to engage in illicit activities to present themselves in as favorable a light as possible—especially in the context of high-stakes testing. Thus, the opportunity for malfeasant behaviors is a major psychologically-based source of construct irrelevant variance in that these behaviors (may) produce test scores that do not accurately reflect the test-taker’s standing on the constructs of interest.

We define “malfeasant behavior” as deliberately falsifying or misrepresenting one’s responses on a test, assessment tool,

or device in an attempt to distort one's true standing on the constructs of interest. Malfeasant behaviors may take one of two forms—cheating or response distortion. *Cheating* is associated with ability and knowledge tests and entails the use of illicit aids to obtain and produce the keyed or correct answers to test items. Cizek (1999) defines cheating as “an attempt, by deceptive or fraudulent means, to represent oneself as possessing knowledge [or ability]. In testing specifically, cheating is violating the rules” (p. 3). *Response distortion*, on the other hand, is associated with noncognitive measures and refers to deliberately falsifying one's responses to self-report items, taking the form of faking, impression management, and other forms of non-truthful responding with the goal of presenting oneself in as (socially) favorable a light as possible (Paulhus, 2002). Consequently, in the present chapter, we use the term “cheating” to refer to malfeasant behaviors on ability and knowledge tests and “response distortion” to refer to malfeasant behaviors on noncognitive and non-ability measures.

Cognitive Ability and Knowledge Tests

Ability and knowledge tests and measures are characterized, and thus are distinguishable from noncognitive self-report measures, by the fact that they have demonstrable incorrect and correct or best answers. To this end, in situations in which test security is a concern, proctoring is the primary means by which cheating is curtailed. Cheating on knowledge and ability tests may take the form of employing illicit aids such as cheat sheets, calculators and dictionaries, surrogate test-takers (for example, a smart friend), or preknowledge of test items. Cheating is commonly inferred by either direct observation or statistically. In the context of unproctored Internet-based tests, the latter approach is more germane and typically calls for a repeated administration of the test—the first under unproctored conditions and the second under proctored conditions. Differences in performance on the two administrations are then used to make inferences about the absence or presence of cheating. Needless to say, said repeated administrations may not always be feasible or practical.

Amount of Cheating

Although the number of published studies on the prevalence of cheating in employment testing and settings is very limited, cheating in educational settings is very well documented and seems to be quite widespread as well. For example, reviews by Cizek (1999) and Whitley (1998) summarized research indicating that 50 percent or more of all college students reported cheating on an exam at least once during their college education. Indeed Chapters 2 and 3 of Cizek's book present a trove of information on the frequency and perceptions of cheating and common methods of cheating; and Chapter 4 discusses cheating in postgraduate and professional contexts. Although the focus of most of the writing and data has been on cheating in academic contexts (primary, secondary, higher, and postgraduate and professional education), suffice it to say that, even with the limited data that are available it is not unreasonable to assume that if unchecked the proclivity to cheat on employment tests and exams is probably just as high as that which has been observed in educational settings.

For instance, it has been reported that 45 percent of U.S. job applicants falsify their work histories (Automatic Data Processing, Inc., 2008). Consistent with this, both Hense, Golden, and Burnett (2009) and Beaty, Fallon, Shepherd, and Barrett (2002) reported higher scores on the unproctored versions of their tests. The respective standardized mean differences for these two studies were 0.32 and 0.52. Furthermore, using a 3 standard error of measurement (SEM) operationalization of cheating, Beaty, Fallon, Shepherd, and Barrett identified 9 percent of the top seventy-five scorers as having cheated on the unproctored administration. Likewise, using a design in which applicants completed an unproctored Internet-based cognitive ability test first as job applicants and then as research participants (average retest interval of 429 days), Arthur, Glaze, Villado, and Taylor's (2010) 1 SEM operationalization of cheating identified 7.77 percent of their sample as having cheated with these individuals being distributed across the entire range of test scores. It is also worth noting that from one perspective, their 7.77 percent

may represent an underestimate of the prevalence of cheating since the cognitive ability test was designed to be very speeded to prevent cheating. In addition, since the (research-based) retest was on a volunteer basis and consequently, via self-selection, it may have had a larger proportion of individuals who did not cheat on the first, high-stakes assessment. Finally, in a finding in education that has resonance for unproctored Internet-based testing, Whitley's (1998) review found that an honor code coupled with subsequent unproctored exams resulted in higher levels of self-reported cheating on the unproctored exams.

To summarize, ability and knowledge tests are susceptible to cheating for several reasons. First, in organizational, educational, and other high-stakes settings (for example, professional certification or licensure), test-takers' scores play an important role in whether they will be hired, admitted, or otherwise obtain or achieve their desired outcome. Second, the transparency and valence of ability and knowledge test items are clear—that is, these test items have demonstrable incorrect and correct or best answers. So, unlike personality measures, the desired response is a matter of fact. Consequently, although additional research is needed to further document its prevalence and extent, it is reasonable to conclude that cheating can and does occur in applicant and other high-stakes organizational testing and that it is likely to be exacerbated with the use of unproctored remotely delivered assessments since the absence of proctors creates a permissive environment for cheating. We obviously do not claim or intend to imply that proctored testing is immune from cheating—to the contrary, data such as that reviewed and summarized by Cizek (1999) and Whitley (1998) would suggest otherwise. Since proctoring is the primary means by which rule compliance is enforced, it is not unreasonable to posit that in the absence of proctors those who are motivated to do so may use surrogates and advisors and illicit reference materials, calculators, and dictionaries in an effort to increase their ability or knowledge test scores. The resultant expectation is that the presence of cheating will result in elevated test scores.

Differences in Cheating in Proctored and Unproctored Internet-Based Settings

Two noteworthy issues in the discussion of differences in cheating in proctored and unproctored Internet-based settings are the methods and frequency of cheating. Methods of cheating that would appear to be common to *both* proctored and unproctored Internet-based tests include cheating by taking, giving, or receiving information from others, and cheating through the use of forbidden materials and information. Cheating via the use of surrogate test-takers would also seem to be common to both although one could reasonably speculate that it is easier to accomplish this with unproctored Internet-based tests than proctored tests. Finally, Cizek (1999) describes a category of cheating methods that involve “taking unfair advantage of the person(s) giving the test or the circumstances of the testing process. For example, students can take advantage of vague, ambiguous, or uncontrolled test administration protocols, their instructor’s willingness to help . . .” (p. 48). As described, this category of cheating methods would seem to be particularly more germane to proctored than unproctored settings.

A methodological aspect of cheating that characterizes only unproctored Internet-based testing pertains to attacks on the technology used to support Internet-based testing. Thus, pirate and hacker attempts to access test content, scoring keys, and test score data and, subsequently, making them available to test-takers (Burke, 2009), is a threat that is less germane in proctored non-Internet-based settings. Another aspect of unproctored Internet-based testing that may differentiate it from proctored testing is the international scope of the former, which derives primarily from its often touted advantage of being a “test anywhere-test anytime” method. With the increasing emphasis on corporate globalization, and the concomitant globalization of organizations’ human resource management systems and practices, the use of Internet-based assessments is consonant with this business model. However, an unintended but not surprising result of this is that the security threats and concerns that confront a specified Internet-based assessment tool are consequently also international in scope. So, for instance, as part of an effort to maintain

the integrity of their Internet tests, SHL conducts web patrols to detect test security breaches and test fraud; and reflective of the international scope of security threats, Burke (2009) reports that of the eighteen (out of thirty) websites that SHL detected and classified as high-risk in an eighteen-month period, four were sites that were operating in England and fourteen were in China.

Concerning frequency and prevalence, because proctored settings, by their very nature, use human observation and supervision to curtail, prevent, and detect these cheating methods, one would expect the prevalence of cheating—at least overt cheating—to be lower in proctored settings. Thus, in proctored settings, one would expect cheating to be associated with the extent to which test-takers can or believe that they can outwit human proctors, who because they are neither perfect or infallible, are unable to prevent or detect all cheating attempts. Thus, said cheating attempts would also be expected to be more covert and clandestine. In contrast, one would expect cheating in unproctored settings to be more open and overt and less clandestine. Therefore, the logical inference is that cheating in unproctored settings may be more prevalent and frequent, which is consonant with Whitley's (1998) finding that students were more likely to cheat when they thought there was relatively little risk of being caught and when they anticipated large rewards for success on the test.

In summary, proctored and unproctored settings share some common cheating threats. Similarly, using several telling examples, Drasgow, Nye, Guo, and Tay (2009) make a compelling case that in reference to cheating, proctored assessments cannot necessarily be considered the "gold standard". Nevertheless, because the use of human observers and proctors is the primary means by which several threats and specified cheating methods are prevented and detected, one would expect the use of cheating methods in unproctored settings to be more overt and less clandestine, and concomitantly, one would also expect the frequency and prevalence of cheating in unproctored settings to also be higher. There are also aspects of cheating, such as attacks on the test hosting server or site and the global and international nature of cheating-related issues and concerns, that are more applicable to unproctored Internet-based testing. However, the critical question is that, although cheating does occur, is there

any clear, overwhelming *empirical evidence* that conclusively demonstrates that the amount of cheating is more than trivial? And furthermore, is it qualitatively more than that which occurs with proctored tests? As Drasgow, Nye, Guo, and Tay (2009) emphasize, “Cheating on proctored tests is not difficult, and there is evidence to suggest that it may be pervasive in some situations” (p. 48). Based on our review of the literature, at the present time it seems we simply assume—and maybe justifiably so—that cheating on unproctored tests is more prevalent. However, we must acknowledge that it is just that, a reasonable supposition, and that there is very limited empirical research that has investigated and conclusively and overwhelmingly demonstrated this.

Detection

Although there is agreement about the need to be concerned about cheating on unproctored Internet-based tests, given its elusive nature, it is difficult to directly measure this behavior. Thus, in both research and applied settings, rarely do we actually see or observe applicants or test-takers cheating, nor is there typically direct evidence that they did. As a result, the techniques used to detect cheating are not directly behavioral, but instead, cheating is determined and operationized in terms of inferences that are made from test scores. In the absence of directly observing test-takers cheating, techniques for detecting cheating in unproctored Internet-based testing take the form of (1) statistical detection, (2) score comparison and verification testing, and (3) technological detection.

Statistical Detection

Statistical detection methods assess (1) the similarity of (pairs of) scores or response patterns (of errors and correct responses) or (2) deviations from some known or expected distribution of scores which may be probabilistic or actual (retest) scores. Because they originate from educational testing, statistical detection methods focus primarily on answer copying as a particular source of cheating. In this regard, Cizek (1999; see also Hanson, Harris, & Brennan, 1987; Saupe, 1960) notes the distinction between *chance* and *empirical* statistical detection methods;

adding that, over time, detection methods have moved away from empirical to chance methods. Chance methods “compare an observed pattern of responses by a pair of examinees (one or both of whom are suspected to be cheating) to a known distribution, such as the binomial or standard normal distribution. Empirical methods [on the other hand] compare the probability of an observed pattern of responses by a pair of examinees to a distribution of values derived from other independent pairs of students who took the same test. Distributions of empirical indices for suspected copiers are compared to distributions of statistics obtained under conditions where cheating could not have occurred” (Cizek, 1999, p. 137). Regardless of the specific approach used, individuals with a large index (for example, B , K , g_2) are subsequently flagged as having cheated. The interested reader is referred to Cizek (1999) for a more detailed presentation of these statistical detection methods that apply and are used almost exclusively for detecting whether a test-taker copied from another. In summary, statistical detection methods are mostly used in educational testing settings and may have limited applicability to unproctored Internet-based testing as used in organizational and employment testing contexts—with the military being an exception to this. A final reason for its limited applicability in (civilian) organizational settings is that statistical detection is a post hoc approach that requires relatively large sample sizes that are not present in most organizational settings.

Score Comparison and Verification Testing

In unproctored Internet-based testing of the sort that is commonly used in organizational settings, copying the answers of an adjacent test-taker is less of a concern; the focus is more on rule violations pertaining to the use of illicit aids such as calculators and dictionaries, preknowledge of test items, and the use of surrogate test-takers. So, in an effort to utilize unproctored Internet-based testing *and* maintain test utility and validity, the use of proctored retesting to verify and confirm unproctored test scores has been advocated (International Test Commission [ITC], 2005; Tippins, Beaty, Drasgow, Wade, Pearlman, Segall, & Shepherd, 2006). Proctored retesting can take the form of a full length retest (repeating the original test or an alternate form of

equal length) or an abridged retest which uses a shorter form of the original test. Regardless of the specific approach used, score differences between the two administrations—that is, when the proctored score is lower than the unproctored score—are then used to make inferences about cheating.

Psychometric theory and research indicates that retesting is generally associated with *increases* in test scores. For instance, Hausknecht, Halpert, Di Paolo, and Moriarty-Gerrard's (2007) meta-analytic results indicate that test-takers increase their retest scores both with coaching ($d = 0.64$), and without coaching ($d = 0.21$). Consequently, within some standard error of measurement, to the extent that a test-taker's proctored *retest score is lower* than his or her first unproctored score, the unproctored score is considered to be suspect because, psychometrically, it is expected to be the same or higher, not lower, than the first test score. In summary, the primary issues in verification testing are (1) score equivalence and (2) the techniques and thresholds that one uses to determine that the unproctored and proctored scores are different and, thus, warrant the suspicion or conclusion of cheating.

A number of empirical studies have investigated these very issues. Hense, Golden, and Burnett (2009) report the results of a between-subjects design in which applicants who took a video-based job simulation when it was first rolled out as a proctored test obtained lower scores than those who took it when it was later reintroduced as an unproctored Internet-based test ($d = 0.32$). As the authors acknowledge, although the higher scores on the unproctored version are consistent with a cheating hypothesis, in the absence of random assignment, alternative explanations such as different testing conditions, locations, and test-takers for the two versions cannot be ruled out. In addition, group mean differences do not identify specific individuals who may have cheated; they merely indicate that, at the group level, unproctored test scores are higher than proctored test scores.

In a study that is less susceptible to the preceding methodological threats, Beaty, Fallon, Shepherd, and Barrett (2002) report the results of a multi-stage assessment process in which applicants who were not screened out on the basis of minimal requirements ineligibility took a speeded (twelve minutes for

fifty-four items) unproctored Internet-based cognitive ability test. The top seventy-five scorers from the unproctored administration were then invited to complete a parallel form of the test under proctored conditions. The average retest interval was four weeks. The results of this within-subjects design were consistent with the cheating hypothesis—the unproctored scores were higher than the proctored scores ($d = 0.51$). In addition, using a 3 SEM operationalization, six individuals (9 percent) had proctored scores that were 3 SEMs lower than their unproctored scores and so their unproctored scores were flagged as suspect.

Arthur, Glaze, Villado, and Taylor (2009, 2010) also used a within-subjects design. However, they compared two unproctored conditions such that on the first administration, test-takers completed the unproctored Internet-based speeded cognitive ability test (twenty minutes for 120 items) as job applicants (high-stakes) and on the second administration volunteered to retake the same unproctored test for research purposes (low-stakes). The average retest interval was 429 days. However, unlike Hense, Golden, and Burnett (2009) and Beaty, Fallon, Shepherd, and Barrett (2002), Arthur, Glaze, Villado, and Taylor's findings were more consonant with a psychometric than a cheating explanation since the retest scores (although low-stakes) were *higher* than the initial high-stakes scores ($d = 0.36$). In addition, using a 1 SEM operationalization, only 7.77 percent of the sample (twenty-three out of 296) had Time 1 scores that were 1 SEM lower than their Time 2 scores. Arthur, Glaze, Villado, and Taylor (2010) attributed this low incidence of cheating to the very speeded nature of the test, which was intentionally designed as such to increase the temporal costs associated with and, hence, deter cheating. However, the authors also acknowledge that, although speededness may deter some forms of cheating, it does not curtail the use of surrogate test-takers. Finally, since the (research-based) retest was on a volunteer basis and, thus, via self-selection, it may have had a larger proportion of individuals who did not cheat on the first high-stakes assessment.

In a two-step testing process in which candidates first completed an unproctored Internet-based speeded perceptual speed and accuracy test and then retested on a parallel form under

proctored conditions on average a month later (both high-stakes), Nye, Do, Drasgow, and Fine (2008) obtained results that were quite similar to Arthur, Glaze, Villado, and Taylor (2010). Specifically, the proctored retest scores were higher than the unproctored initial test scores ($d = 0.22$). In addition, using a regression-based approach, Nye, Do, Drasgow, and Fine identified only four individuals (out of 856) who showed differences larger than the 1.96 cutoff.

The research reviewed above used different approaches—namely SEM- and regression-based—to infer whether differences between unproctored and proctored test scores were indicative of cheating or not. Along these lines, Guo and Drasgow (2009) present a Z-test and a likelihood ratio test as alternative means of comparing the consistency of performance across unproctored and proctored testing conditions and subsequently, identifying the dishonest test-taker. Based on the results of a simulation study, Guo and Drasgow conclude that (1) both test statistics have a high power to detect cheating at low Type I error rates and (2) compared to the likelihood ratio test, the Z-test is more efficient and effective.

The use of score comparisons and verification testing raises a number of additional noteworthy issues. First, proctored retesting obviously adds additional steps to the recruitment and testing process and thus, to some extent diminishes the efficiencies and cost effectiveness advantages often touted in support of unproctored Internet-based tests (that is, “test anywhere-test anytime”). Second, as Pearlman (2009) notes, retesting paradoxically constitutes “a tacit admission by the organization that UIT results cannot be relied upon, which may have negative indirect consequences” (p. 15).

Third, as previously alluded to, retesting raises interesting issues about determining whether the unproctored and proctored test scores are from the same person. From an applied perspective, one could argue that the importance of this depends on which test score is intended to be the score of record. Thus, the criticality of determining whether the two scores are from the same person or not is largely a function of whether the unproctored score is the score of record, such that it is less of an issue and concern if the proctored score is the score of record.

Fourth, related to the preceding is the question of *which* of the two test scores should be used as the score of record from a decision-making perspective—the first (unproctored) or the second (proctored) score, or some combination of the two? Although this would appear to be a critical issue with important psychometric and applied implications (for instance, Arthur, Glaze, Villado, & Taylor [2010], Beaty, Fallon, Shepherd, & Barrett [2002], and Nye, Do, Drasgow, Fine [2008] reported retest correlations of .78, .41, and .63, respectively, between the initial and retest test scores of their tests), we were unable to locate any studies that addressed or investigated the issue of which score to use as the operational score. Nevertheless, Burke (2009) notes that SHL uses the first score as the score of record and treats the verification test score (“short tests administered in a proctored setting” [p. 36]) in a manner that is analogous to a fake-good check on a personality measure. However, to the extent that the second (proctored) administration uses either the same or an equivalent test, a reasonably good case could be made for using *that* as the score of record and the first as just an initial screen—but in the absence of any empirical research, it is difficult to decide which approach is better and subsequently, to make any firm recommendations.

A fifth issue is that of measurement equivalence. For instance, Lievens, Reeve, and Heggstad (2007) reported that proctored retesting is associated with changes in the factor structure of ability tests such that test scores based on a retest of general mental ability are less saturated with general mental ability and, subsequently, less predictive of grade point average than initial test scores. These results suggest that the use of proctored retesting may threaten the construct-related and criterion-related validity of test scores gathered using this approach. Consequently, this is an issue that should be considered when deciding whether to use the unproctored (first administration) or the proctored (second administration) scores as the operational assessment scores.

In summary, the empirical research reviewed above indicates that verification testing and score comparisons are viable approaches for detecting score differences that may subsequently be interpreted as cheating. In operational, as opposed to research terms, this will entail retesting that is also high-stakes using either

a repeated administration of the exact same test, a parallel form (same length but equivalent form), or an abridged form. Another critical issue pertains to how one operationalizes “cheating”, that is, the threshold or cut-point at which one considers a score difference to be suspect. The research reviewed above used both regression-based and SEM-based approaches, and in the latter, a range of SEM bandwidths. In addition to these two approaches, Guo and Drasgow (2009), present a Z-test and likelihood ratio test to detect cheating. Finally, retesting raises issues pertaining to measurement equivalence and, similarly, which of the two sets of scores to use as the operational test score.

Technological Detection

A range of technological innovations are geared at several facets of cheating such as verifying the identity of test-takers and also monitoring their behavior during the testing process. So, for instance, biometric identification systems such as fingerprint, iris, and retina scans may be used to verify the identity of the test-taker. In addition, webcams may be used to monitor test-takers during the test administration. However, from one perspective, although assessments administered under these conditions may be remotely delivered, they are strictly speaking not unproctored since they are being proctored—albeit electronically and technologically instead of via the direct presence of human proctors. Hence, these are really new ways of monitoring test-takers that do not require the physical presence of a proctor in the same room as the test-taker. Nevertheless, they do require the same level of human vigilance and as a result, are susceptible to the same threats as direct human proctoring because someone has to monitor the webcam monitors and the data and information that is generated by the other technologies.

Other layers of technological security for unproctored Internet-based tests include authentication via keystroke analytics, electronic monitoring and control such as real-time data forensics, browser lockdown, keystroke monitoring, and operating system and Internet access controls (Foster, 2009). For example, keystroke analytics is a biometric authentication system that is based on the premise that people have unique typing patterns. Specifically, the system requires test-takers to enter a password

repeatedly prior to the start of the test to establish a keystroke pattern. The system then periodically requests the same password to be retyped, and the typing pattern is compared to the established keystroke pattern. If the established keystroke pattern cannot be reproduced, the inference is that another individual is now taking the test. Real-time data forensics is a technique of flagging test-takers who display suspect item response or response latency patterns. For example, if a test-taker correctly answers several difficult questions then incorrectly answers several easy questions, the program flags the test for future review. Keystroke monitoring is a technology that allows the system to record test-takers' attempts to use illicit keys (for example, alt-tab, print screen) and alert a remote proctor or suspend or stop the test. Other security options include software programs that limit test-taker's access to available information (browser lockdown, operating system and Internet access control). Web patrols may also be used to detect Internet sites that compromise test materials (Burke, 2009).

In summary, although technological detection techniques hold some promise, they have some disadvantages, including cost, invasiveness, and applicant reactions to the testing process. And some even require some level of human proctoring, albeit technologically mediated. In addition, some of these methodologies are currently state-of-the-art and may consequently not be a viable option (from a cost and expertise perspective) in most small-scale testing programs. Hence, it is not too surprising that they are currently not as widely used and as commonplace as, for example, verification retesting. Nevertheless, it is important to note that technology changes rapidly and it is likely that new, and perhaps more available and cheaper, means of detecting (and preventing) cheating will be developed. Finally, it should also be noted that there is a dearth of research that empirically investigates the effectiveness of these techniques.

Once Detected, What Does One Do with and About Cheats and Perpetrators?

The endeavor of trying to detect cheating begs the question, "What does one do with and about cheats and perpetrators once they have been detected?" This ostensibly simple question raises

several interrelated issues. The first pertains to “What constitutes positive evidence of cheating?” Clearly, simple score differences, even after taking into account some specified amount of measurement error (which encapsulates sources of error variance due to factors such as illness and practice), is unlikely to meet the evidentiary threshold; a condition that is equally applicable to most of the technological detection methods as well. Indeed, Cizek (1999) discuss several instances when academic administrators have been very hesitant to press academic dishonesty charges when human proctors have accused students of cheating on exams on the basis of their having *observed* them doing so—it would seem that even “charges” based on only observation may not be sufficient. Ironically, even in large-scale academic testing such as those undertaken by ACT and ETS, statistical evidence is brought to bear only when some other trigger (for example, observation) provides a strong reason for flagging cases for subsequent statistical analysis (Cizek, 1999). Consonant with this, in describing the ITC’s (2005) position on this issue, Bartram (2009) notes that “[a]s in a proctored environment, there needs to be positive evidence of cheating rather than just circumstantial evidence” (p. 13).

A second related issue is that of false positives or identifications. Within this context, it is important to emphasize that none of the prevailing detection methods actually *detect* cheating. They only provide evidence or data that permit inferences about the presence of cheating. Thus, the conclusion that cheating has occurred on the basis of these methods is almost always probabilistic and requires an inference. Given these limitations, confronting suspected cheaters and perpetrators, although necessary in academic testing, is neither required nor necessary in employment testing. However, an employment-related decision has to be made on the basis of the specified test scores. In verification testing, this decision is probably moot if the proctored retest is used as the operational score or record. On the other hand, it is a much bigger concern if the unproctored initial test score is the operational score.

So, if we do not have to confront them, can we fail them for having cheated? Whereas the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National

Council on Measurement in Education [NCME], 1999) does not currently speak specifically to unproctored Internet-based testing, it has several standards that serve as professional guidelines concerning “test irregularities” and how to deal with them. Standards 8.11, 8.12, and 8.13 appear to be particularly germane. For instance, Standard 18.11 states that:

“In educational testing programs and in licensing and certification applications, when it is deemed necessary to cancel or withhold a test-taker’s score because of possible testing irregularities, including suspected misconduct, the type of evidence and procedures to be used to investigate the irregularity should be explained to all test-takers whose scores are directly affected by the decision. Test-takers should be given a timely opportunity to provide evidence that the score should not be canceled or withheld. Evidence considered in deciding upon the final action should be made available to the test-taker on request.” (p. 89)

In contrast to the current *Standards*, the *ITC Guidelines* (ITC, 2005) speaks *specifically* to unproctored Internet-based tests and the pertinent guidelines that address the issue of cheating are Guidelines 43, 44, and 45. For instance, Guideline 45.3 explicitly stipulates the need for verification testing noting that:

“For moderate and high stakes assessment (for example, job recruitment and selection), where individuals are permitted to take a test in controlled mode (i.e., at their convenience in non-secure locations), those obtaining qualifying scores should be required to take a supervised test to confirm their scores. Procedures should be used to check whether the test-taker’s original responses are consistent with the responses from the confirmation test. Test-takers should be informed in advance of these procedures and asked to confirm that they will complete the tests according to instructions given (for example, not seek assistance, not collude with others, etc). This agreement may be represented in the form of an explicit honesty policy which the test-taker is required to accept.” (pp. 20–21)

So, like the *Standards*, the *ITC Guidelines* also stipulates that test-takers should be informed in advance of the retesting procedures.

In conclusion, in our opinion, it would seem that the most efficacious and straightforward way to deal with cheats and perpetrators is to have a verification retest and use this score as the score of record. Test-takers should also be informed in advance of this along with the fact that the unproctored test is just an initial screen. This approach preempts the need to confront test-takers about having cheated.

Deterrence

Given the practical, ethical, and legal challenges associated with not only the detection of cheating, but also what to do with and about applicants suspected of cheating, the proverb, “an ounce of prevention is better than a pound of cure” is particularly germane here. Approaches to deter cheating can be conceptualized as having one of two foci—(1) to discourage cheating attempts and (2) to make it more difficult to engage in cheating. Approaches that fall into the first category include monitoring and the saliency of other detection techniques (including web patrols), and warnings and threats. The second category consists primarily of test design characteristics and features such as the use of multiple test forms, computerized adaptive tests, and speeded tests.

Monitoring

Because we define an unproctored Internet-based test or remotely delivered assessment as one that does not entail the direct monitoring or supervision of a human proctor in the same physical setting or location as the test-taker, our focus here is on electronic monitoring primarily in the form of webcams, keystroke analytics, and keystroke monitoring. Foster (2009) describes new technology-enhanced ways of monitoring remotely delivered assessments that do not require the physical presence of a proctor in the same room with the test-taker. But by virtue of needing a human to monitor the data obtained from these systems (for example, monitoring webcam monitors or other data), these technology-based monitoring systems are ultimately susceptible to the human-related concerns and limitations that characterize traditional proctored tests.

It is also worth noting that, although they may not be designed or intended to specifically do so, it would seem that the saliency of detection techniques such as biometric identification, web patrols, and webcams, coupled with test-takers' beliefs about their effectiveness, are also likely to serve a deterrent function as well. That being said, it should be noted that there is a paucity of empirical research investigating the effectiveness or efficacy of these techniques in deterring or discouraging cheating.

Warnings and Threats

Like the saliency of detection techniques, warnings and threats may also serve to discourage and deter cheating. Conceptually, warnings and threats may take the form of informing test-takers about the presence and effectiveness of detection techniques and the consequences of detection. Thus, it is not unreasonable to posit that informing job applicants that they will be retested under proctored conditions to verify their test scores may serve a deterrent function. Consonant with this, as previously noted, the *ITC Guidelines* (ITC, 2005) recommends informing test-takers in advance of the expectations and consequences of detected cheating. However, we again note that there is a paucity of empirical research investigating the effectiveness or efficacy of this technique in the context of unproctored Internet-based tests.

Test Design Characteristics and Features

One source of cheating arises from a breach of test security resulting in the preknowledge of test items. Thus, an unproctored Internet-based testing program that continuously and frequently administers the same test form (especially if its length is relatively short) is likely to be most susceptible to this cheating threat because of the high item overlap rate (the number of overlapping items encountered by test-takers divided by the length of the test; Chang & Zhang, 2002; Drasgow, Nye, Guo, & Tay, 2009). Strategies to make it more difficult for cheating to occur from this source include limiting the number of administrations, using multiple test forms, and using computerized adaptive tests (Foster, 2009). Thus, Burke (2009) describes "a randomized testing model through which equivalent but different tests are constructed from item response theory calibrated item pools"

(p. 36). So with this and other similar approaches, such as those in which items are selected and presented based on estimates of the test-taker's ability, the likelihood that multiple test-takers will see a large number of common items is reduced (Dragow, Nye, Guo, & Tay, 2009).

However, these strategies are not without their costs. For instance, limiting the number of administrations is at odds with the continuous testing advantage (that is, "test anywhere-test anytime") that is frequently cited as a primary attraction of unproctored Internet-based tests. In addition, the use of multiple forms and computerized adaptive testing require large item pools. So, in spite of their efficacy, these strategies may not be practical in most small-scale testing programs.

Another approach to making cheating more difficult and, thus, deter it is to use speeded tests, which by virtue of their time constraints, have the potential to curtail the expected malfeasant behaviors. Of course, this approach is predicated on the assumption that a speeded administration is consonant with the job-relatedness of the test. Another caution against their use—in the United States at least—would be in situations in which accommodations for additional time associated with the Americans with Disabilities Act must be made. However, within these boundary conditions, assuming there is no preknowledge of test content, possible modes of cheating (for example, using additional aids or helpers) are time dependent. That is, if individuals do not have preknowledge of the test content and are not using surrogate test-takers, then the time constraints under speeded conditions should make it more difficult to engage in and subsequently deter test-takers from engaging in many of the noted malfeasant behaviors.

Consistent with this reasoning, Arthur, Glaze, Villado, and Taylor's (2010) results indicated that for a speeded unproctored Internet-based cognitive ability test (twenty minutes for 120 items), only 7.77 percent of the sample (twenty-three out of 296) had Time 1 (high-stakes) scores that were 1 SEM lower than their Time 2 (low-stakes) scores. In addition, consistent with a psychometric, instead of a cheating explanation, the retest (low-stakes) scores were *higher* than the initial high-stakes scores ($d = 0.36$; average retest interval = 429 days). In a two-step testing process in

which they used an unproctored Internet-based speeded perceptual speed and accuracy test, Nye, Do, Drasgow, and Fine (2008) obtained results that were quite similar to Arthur, Glaze, Villado, and Taylor's results. Specifically, the proctored retest scores were higher than the unproctored first test scores ($d = 0.22$; average retest interval = 1 month). In addition, on the basis of a regression-based approach, they identified only four individuals (out of 856) whose test score differences warranted some concern. However, it is worth noting that, whereas speededness may deter some forms of cheating, it does not curtail the use of surrogate test-takers or preknowledge of test items.

Finally, Foster (2009) describes an alternative test item format, the Foster Item, in which response options are presented serially instead of simultaneously and presentation ceases once an item has been answered either correctly or incorrectly. Although Foster does not describe it in these terms, it would seem that this item format shares several characteristics in common with the constructed-response format (for example, see Arthur, Edwards, & Barrett, 2002; and Edwards & Arthur, 2007). Concerning its effectiveness, Foster notes that "early research indicates that the Foster Item performs as well or better than its traditional multiple-choice counterpart with significant security advantages" (Foster, 2009, p. 32).

In summary, there are several approaches and techniques to deterring cheating. These techniques focus primarily on either discouraging cheating or making it more difficult to do so. However, as is characteristic of cheating on unproctored Internet-based tests in general, there is very limited empirical research that has investigated the efficacy and effectiveness of these approaches. So, in the absence of extensive empirical support, we only have their conceptual merit to go on—and said conceptual merit appears to be reasonably sound.

Noncognitive Tests and Measures

We use the terms "noncognitive" and "nonability" to broadly refer to the class of tests and measures for which there are ostensibly no true correct or incorrect answers to the items on the measure. These measures also typically entail self-reports.

Thus, although we may frequently make specific references to personality measures (because the extant literature has most frequently and extensively focused on this class of noncognitive measures), our discussion of specified issues is equally applicable to other self-report measures, such as measures of integrity, interests, attitudes, and other noncognitive constructs (Alliger & Dwight, 2000; Grubb & McDaniel, 2007; McFarland & Ryan, 2000). Furthermore, consonant with the construct/method distinction (Arthur & Villado, 2008), some testing *methods* have received some research attention in terms of dissimulation. These include resumes, job application blanks (for example, Wood, Schmidke, & Decker, 2007), employment interviews (for example, Delery & Kacmar, 1998; Ellis, West, Ryan, & DeShon, 2002; Levashina & Campion, 2006), biographical measures (for example, Schmitt & Kuncze, 2002), and assessment centers (McFarland, Yun, Harold, Viera, & Moore, 2005).

So, in the absence of true correct or incorrect answers coupled with the inability to verify the accuracy of test-takers' responses, noncognitive measures are recognized as being susceptible to test-takers' dissimulation and response distortion which may take the form of self-deception or impression management efforts (Edens & Arthur, 2000). Consequently, whereas cheating is the malfeasant behavior of interest with ability and knowledge tests, dissimulation or response distortion—in the form of social desirability responding—is the primary malfeasant behavior of interest with noncognitive and nonability measures.

Paulhus (1986, 2002) highlights the distinction between self-deception and impression management as facets of social desirability responding. Social desirability responding is the tendency to over-report socially desirable personal characteristics and to under-report socially undesirable characteristics. It entails the tendency to choose specified responses even if they do not represent one's true tendency or opinion. As a facet or dimension of social desirability responding, self-deception occurs when an individual unconsciously views himself or herself in an inaccurately favorable light; this typically entails a lack of self-awareness. In contrast, impression management or deliberate response distortion refers to a situation in which an individual consciously presents himself or herself falsely to create a favorable impression. Our focus here

is on intentional response distortion (that is, impression management), as opposed to self-deception.

A variety of terms and labels are used to describe response distortion in the extant literature. Some of these include social desirability, faking, dissimulation, impression management, lying, honesty, frankness, claiming unlikely virtues, denying common faults and unpopular attitudes, exaggerating personal strengths, response fabrication, good impression, and self-enhancement. Although there may be subtle distinctions among some of these descriptive labels, for the purposes of this chapter, we use the term “response distortion” to collectively refer to them and subsequently define it as a conscious and deliberate attempt on the part of test-takers to manipulate their responses in order to create an overly positive impression that deviates from their true standing on the trait or characteristic of interest (Ellingson, Sackett, & Connelly, 2007; McFarland & Ryan, 2000; Zickar & Robie, 1999). Response distortion is commonly conceptualized as systematic error variance (Arthur, Woehr, & Graziano, 2001; Tett, Anderson, Ho, Yang, Huang, & Hanvongse, 2006; cf. Uziel, 2010). Thus, job applicants are assumed to distort their responses because it assists them in attaining valued outcomes such as jobs and promotions. Hence, response distortion is posited to be determined by one’s motivation, and ability, along with specified situational factors such high- versus low-stakes testing (McFarland & Ryan, 2000, 2006; Snell, Sydell, & Lueke, 1999; Tett, Anderson, Ho, Yang, Huang, & Hanvongse, 2006).

Amount of Response Distortion

Paralleling and analogous to efforts taken to maintain test security and prevent cheating and other sorts of malfeasant behaviors in ability testing contexts, the prevailing view in both the academic and applied literatures is that applicants do distort their responses and answers on noncognitive and nonability self-report measures, although the focus has particularly been on personality measures (Anderson, Warner, & Spector, 1984; Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Cellar, Miller, Doverspike, & Klawnsky, 1996; Ellingson, Sackett, & Connelly, 2007; Hough, Eaton, Dunnette, Kamp, & McCloy,

1990; Levin & Zickar, 2002; Schmit & Ryan, 1993). So, for example, Anderson, Warner, and Spector (1984) reported that 45 percent of applicants faked on their experience with bogus task statements. Likewise, large mean differences between applicants' and incumbents' scores, typically more than a standard deviation in size, have been reported (Bott, O'Connell, & Doverspike, 2007; Rosse, Stecher, Miller, & Levin, 1998). And as Fluckinger, McDaniel, and Whetzel (2008) conclude in their review of faking in personnel selection, "noncognitive tests can be faked, and they are faked in high-stakes settings" (p. 105).

So, given that applicants distort their responses (Donovan, Dwight, & Hurtz, 2003), process-oriented models of response distortion posit that the prevalence of this behavior is related to a number of factors such as knowledge of the construct being measured (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; McFarland & Ryan, 2000), where response distortion is likely to be higher when test-takers "know" which constructs are job-relevant along with the valence of the items. This factor is also related to the transparency (direct/indirect) of the items and hence, the measure (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; McFarland & Ryan, 2000). An additional factor is the value of the outcomes associated with testing. From the test-taker's perspective, this generally takes the form of whether testing is considered to be high- or low-stakes—that is, the extent to which the test-taker desires the job or other outcome that is to be awarded on the basis of the test scores.

In summary, although there may be continued debate about the extent, magnitude, and effect of response distortion (for example, see Dilchert, Ones, Viswesvaran, & Dellar, 2006; Edens & Arthur, 2000; Ellingson, Sackett, & Connelly, 2007; Hogan, Barrett, & Hogan, 2007; Hough & Oswald, 2008; Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmitt, 2007a, 2007b; Ones, Dilchert, Viswesvaran, & Judge, 2007; Smith & Ellingson, 2002; Tett & Christiansen, 2007), one of the most consistent findings in the extant faking literature is that higher means are almost always obtained for fake-good experimental designs and applicant samples compared to honest experimental conditions and incumbent samples (Fluckinger, McDaniel, & Whetzel, 2008), with the resultant conceptual

expectation that malfeasant responding results in elevated scores on desirable characteristics and depressed scores on negative ones with said scores being inaccurate representations of the test-taker's standing on the noncognitive constructs of interest. In addition, some empirical evidence indicates that test-takers who distort their responses are evenly distributed across the score range, with a slight trend of more malfeasant responders in the upper quartiles (Arthur, Glaze, Villado, & Taylor, 2010).

Differences in Response Distortion in Proctored and Unproctored Settings

The critical question is whether the unproctored remote delivery of noncognitive measures results in greater levels of response distortion compared to proctored testing. As previously noted, response distortion is posited to be determined by one's motivation and ability, along with specified situational factors. Hence, proctored and unproctored settings could be conceptualized as differing on the extent to which they make it easy or difficult to fake. So, in an unproctored environment, access to illicit aides may create a relatively more permissive environment compared to proctored testing as test-takers may collaborate with other individuals (for example, surrogate test-takers or advisors) in an effort to inflate their test scores. However, it is unlikely that test-takers will engage in these behaviors if they are confident in their ability to elevate their test scores using their own personal schema of a desirable personality profile.

In response to concerns about response distortion, several techniques have been proposed for preventing or minimizing malfeasant responding on personality and other noncognitive measures. These include the use of forced-choice responses, empirical keying, warnings of verification, and response elaboration (for example, see Hough [1998] for a review). However, glaringly absent from this list of techniques is the use of test proctors because the presence or absence of test proctors has little or no bearing on controlling for response distortion. So, given the preponderance of research that indicates test-takers can effectively distort their responses under proctored conditions (Viswesvaran & Ones, 1999), there is little reason or impetus

for them to behave any differently under unproctored conditions (for example, use surrogate test-takers). Thus, the magnitude and extent of response distortion should be similar for both proctored and unproctored Internet-based noncognitive measures (Arthur, Glaze, Villado, & Taylor, 2010; Bartram & Brown, 2004; Gupta, 2007; Kaminski & Hemingway, 2009; Templer & Lange, 2008). In summary, in contrast to ability and knowledge tests, concerns regarding response distortion on noncognitive and nonability tests (for example, personality) are the same in both proctored and unproctored testing conditions.

Detection

Strategies or attempts to deal with response distortion generally take one of two forms—(1) detection and (2) deterrence. Detection techniques may take the form of score comparison and verification testing, the use of lie scales, inconsistency responding, response latencies, and statistical detection and control. Deterrence strategies include the use of forced-choice response formats, empirical keying, warnings, verifications, and threats, elaboration, and profile matching and the use of nonlinear models. These strategies and their efficacy and effectiveness are briefly reviewed below. It should be noted that the techniques and associated literature are based primarily on proctored tests. But as previously noted, because one would not expect differences in response distortion as a function of proctoring or lack thereof on noncognitive measures, these techniques (for example, use of lie scales) are equally applicable to and germane for unproctored testing as well. In addition, because of the technology via which they are delivered, there are some detection and deterrence techniques that are available for unproctored tests (for example, response latencies, and interactive prompts or cautions) that are unavailable for typical paper-and-pencil tests in proctored settings.

Score Comparison and Verification Testing

As with ability and knowledge tests and measures, one could conceivably use verification retesting and subsequent score comparisons as a means to detect response distortion on noncognitive

measures. However, since proctoring is not a means by which one controls for response distortion, the expectation is that the levels of response distortion under unproctored and proctored conditions would be similar. To this end, although we were unable to locate any studies that undertook a within-study comparison of proctored and unproctored noncognitive measures, between-study comparisons of studies that have focused on either proctored or unproctored noncognitive measures provide some preliminary evidence. Thus, the results of both of Arthur, Glaze, Villado, and Taylor's (2010) studies supported the supposition that unproctored noncognitive measures display levels of response distortion that are similar to those reported for proctored measures in the extant literature (for example, Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). So, on the basis of Arthur, Glaze, Villado, and Taylor's (2009; 2010) data, one can reasonably conclude that unproctored personality measures display mean score shifts between high- and low-stakes testing conditions that are similar to those reported for proctored measures. In addition, the magnitude of the high- versus low-stakes score elevation and the percentage of individuals identified as having elevated high-stakes scores are also similar to those reported for proctored tests (Griffith, Chmielowski, & Yoshita, 2007).

In summary, because proctoring is *not* a means by which response distortion is controlled, verification retesting and subsequent score comparisons have different efficacy implications for cognitive versus noncognitive measures. Specifically, the effectiveness and meaningfulness of this detection technique with noncognitive measures is of very limited and questionable value and hence of limited utility as well. Consequently, it is not surprising that verification retesting and score comparison is not used or extensively discussed as a detection technique for noncognitive tests and measures.

Lie Scales

Because response distortion is theorized to be a consciously motivated behavior driven by a complex array of individual and situational factors, the propensity and tendency to engage in response distortion has been conceptualized as a discernable individual difference. Specifically, attitudes toward faking and

subjective norms predict intentions to fake which in turn, predicts faking behavior (McFarland & Ryan, 2006). Consonant with this, socially desirable responding has been demonstrated to differentiate individuals on their tendency to systematically describe themselves favorably (Paulhus, 2002; cf. Uziel, 2010) as well as their ability to fake (McFarland, & Ryan, 2000).

Concomitant with its conceptualization as an individual difference variable, the extant research also indicates that socially desirable responding can be reliably and validly measured and several approaches have been taken to accomplish this. The first entails a reliance on direct evidence of the tendency to engage in response distortion. An approach to obtaining this direct evidence is by means of a measure external to the focal personality measure—what we broadly refer to as “lie scales”. Examples of lie scales are the Unlikely Virtues Scale (Hough, 1998), and the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960).

A second approach is to embed the lie scale in the focal measure. With this approach, lie scale items are interlaced into the focal measure. Examples of personality measures that use this approach include the California Psychological Inventory (Gough & Bradley, 1996), the Hogan Personality Inventory (Hogan & Hogan, 1992), the Occupational Personality Questionnaire (SHL Group, 2000), the Personality Research Form (Jackson, 1999), and the Inwald Personality Inventory (Inwald, 1992; The interested reader is referred to Uziel [2010] for an informative article on the rethinking of social desirability scales and what they measure.)

Regardless of which approach is used, researchers and practitioners examine test-takers' scores on the lie scales to determine whether the measure was answered honestly (Kuncel & Borneman, 2007). If a predetermined scale score is exceeded, then it is inferred that the test-taker may not have responded truthfully to the focal measure (for example, see Table 1 of Goffin and Christiansen [2003] for recommendations regarding what actions to take for high levels of social desirability responding as specified for several major personality inventories). Concerning the efficacy and effectiveness of lie scales in detecting response distortion, the research evidence indicates that inter-correlations among personality dimensions and lie scales increase under instructions to fake (for example, Michaelis &

Eysenck, 1971). Likewise, Stanush (1996) demonstrated that when test-takers are instructed to fake, there is a stronger relationship between lie scales and other personality inventory scales ($r = .34$), compared to test-takers instructed to answer truthfully ($r = .09$). Furthermore, instructions to fake resulted in elevated scores for both lies scales ($d = 0.82$) and scores on the Five Factor Model (FFM) personality factors ($d = 0.29, 0.53, 0.29, 0.36,$ and 0.28 for agreeableness, conscientiousness, extraversion, emotional stability, and openness, respectively). The larger effects for elevated scores obtained for conscientiousness and emotional stability are consonant with more recent research as well (for example, Birkeland, Manson, Kisamore, Brannick, & Smith, 2006).

In summary, lie (social desirability) scales effectively detect intentional distortion (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). However, most, if not all, of this research has been conducted under proctored settings. Nevertheless, for reasons noted earlier, these relationships are expected to be directly applicable to unproctored settings as well; a conclusion that is consonant with data reported by Arthur, Glaze, Villado, and Taylor (2010).

Inconsistency Responding

An alternative approach to obtaining direct evidence of response distortion is to examine response patterns to indirectly assess response distortion. Indirect approaches have historically focused on detecting response styles, but may also be used to capture the presence of response sets. Response styles are tendencies to respond to items in specific ways regardless of their content (for example, acquiescence [agreeing with every statement]). In contrast, response sets are tendencies to respond to test content with a particular goal in mind (for example, answering in a manner that is socially desirable). Response styles are biases that are consistent across time and test content, whereas response sets are specific to a given situation. Controlling for response styles is usually done by including matched pairs of items with one item of a pair reversed. Inconsistent responses across item pairs is considered to be indicative of a response style (for example, agreeing to all items) or a response set (for example, agreeing to all items in the hopes that said responses are the most desirable). As an example, the Guilford-Zimmerman

Temperament Survey (Guilford, Zimmerman, & Guilford, 1976) uses this approach to assess the presence of response styles and sets.

In summary, inconsistency responding could be viewed as a variation of the lie scale approach to detect response distortion—albeit a more indirect one. Consequently, there is reason to expect that its efficacy and effectiveness at detecting response distortion would be similar to that observed for lie scales.

Response Latencies

It has frequently been shown that participants take longer to respond in fake good conditions compared to those responding under honest conditions (for example, Dwight & Alliger, 1997; Holden, 1995; Holden, Kroner, Fekken, & Popham, 1992; Leonard, 1996; Robie, Brown, & Beatty, 2007; Vasilopoulos, Cucina, & McElreath, 2005; cf. Robie, Curtin, Foster, Phillips, Zbylut, & Tetrick, 2000; Vasilopoulos, Reilly, & Leaman, 2000). The conceptual basis for this effect is that it takes longer to construct a distorted response than to answer honestly. And because unproctored Internet-based tests are by definition computer-administered, albeit via the Internet, assuming the test platform technology lends itself to it, then another possible means of detecting response distortion is to use response latencies. So, although the precision of the measurement of response times over the Internet would be a particularly important technological concern, response latencies as an indicator of response distortion would appear to be an avenue worthy of future research consideration.

Statistical Detection and Control, and What Does One Do with Fakers?

Regardless of the specific response distortion detection technique that is used, at some point a cutoff score has to be established above which test-takers are presumed to be intentionally distorting their responses and below which they are not. This is particularly true for the use of lie scales. Consequently, determining a cutoff score above which applicants are considered to be distorting their responses is an important step in detecting response distortion. Standardized tests and measures will have the specified cutoff scores reported in their test manuals. But regardless

of whether one is using a standardized or an in-house test or measure, the issue of misclassifications—either in the form of false positives or false negatives—is critical. Hence, when deploying a cutoff score, it is important to consider the proportion of each type of misclassification that the specific score will produce. The mean plus three standard deviations has been used as a cutoff score (for example, Hough, 1998), although other values including the mean plus one or two standard deviations, and the mean plus 1 standard error of the measurement have been used as well (for example, Arthur, Glaze, Villado, & Taylor, 2010; Griffith, Chmielowski, & Yoshita, 2007; Hogan, Barrett, & Hogan, 2007).

Once detected, the question becomes what does one do with those who intentionally distort their responses? Two strategies may be used to reduce the effect of response distortion once it has been detected (Hough, 1998). The first entails removing applicants who are identified as distorting their responses from the applicant pool. Hough (1998) examined the effect of removing applicants who scored more than three standard deviations above the mean on an unlikely virtues scale and concluded that such a procedure reduced the effect of response distortion without affecting criterion-related validities.

The second strategy for addressing response distortion is to statistically adjust or correct the focal construct scores of either all applicants or only those identified as having distorted their responses based on their response distortion (for example, lie scale) score. The function of statistical control techniques is to statistically remove the irrelevant systematic error introduced by intentional response distortion from the focal construct scores. That is, statistical control techniques attempt to correct test scores so that they are not influenced by response distortion. To be effective, statistical control techniques require that specified relationships be present. Specifically, there must be (1) a relationship between the response distortion scale scores and the focal construct (for example, personality) scale scores; (2) a relationship between the focal construct scale scores and the job performance scale scores; and (3) an absence of a relationship between the response distortion scale scores and job performance. In the absence of these specific relationships, statistical control will be unable to accurately adjust test scores.

Although statistical control techniques have been examined using applicant samples (Hough, 1998; Ones, Viswesvaran, & Reiss, 1996), it is unclear whether these techniques have ever been employed in *operational* contexts. Nevertheless, as previously noted, studies that have examined statistical control techniques have concluded that they reduce the effects of response distortion without affecting criterion-related validities (Hough, 1998; Ones, Viswesvaran, & Reiss, 1996). In spite of these favorable results, statistically adjusting test scores based on lie scale scores as a means of controlling for response distortion remains controversial. Often, the necessary assumptions regarding the interrelationships between lie scale scores, personality scale scores, and job performance are tenuous at best. Consequently, support for the efficacy and effectiveness of this approach has generally not been very favorable and is mixed (Christiansen, Goffin, Johnson, & Rothstein, 1994; Ellingson, Smith, & Sackett, 2001; Goffin & Christiansen, 2003; Hough, 1998; Ones, Viswesvaran, & Reiss, 1996; Schmitt & Oswald, 2006).

A second concern with statistical control or score adjustments is that true variance may be removed from personality scale scores if response distortion is a true individual personality difference (Ones, Viswesvaran, & Reiss, 1996). A final concern questions the ability of statistical control techniques to correct responses that are intentionally distorted. As Cronbach (1990) wrote, "Once test users take a wrong course, there is no going back to the choice point" (p. 521). Phrased another way, if the test-taker tells lies to the tester, there is no way to convert the lies to truth.

In summary, based on the extant literature, the statistical control of response distortion does not appear to be strongly recommended and should probably be avoided. A number of researchers have concluded that correcting applicant test scores for socially desirable responding is neither a particularly effective nor viable approach to dealing with response distortion (see Hough & Furnham [2003] and Smith & Robie [2004] for summaries of this work). For instance, research such as Ellingson, Sackett, and Hough (1999), has shown that the detection and correction approach for handling response distortion does not work (that is, "corrected" scores are not accurate reflections of honest scores). More recently, Lönnqvist, Paunonen, Tuulio-Henriksson,

Lönnqvist, and Verkasalo's (2007) results indicate that (using uncorrected scores) even rank-order stability is maintained for applicants who are initially tested as applicants, and then tested three years later as incumbents (cf. Arthur, Glaze, Villado, & Taylor, 2010).

Deterrence

As with ability and knowledge tests, the methods and techniques that have been developed and used to deter response distortion can take several forms. For instance, features such as forced-choice responding (all response choices are designed to be similar in terms of social desirability) and empirical keying to create subtle indicators of the focal construct can be built into the test during its design and construction to minimize and deter response distortion (Hough, 1998). The use of firm test instructions and warnings have also been investigated as a means of deterring and minimizing response distortion (McFarland & Ryan, 2000; Vasilopoulos, Cucina, & McElreath, 2005). These strategies, which vary in their degree of effectiveness at deterring response distortion, are briefly reviewed below.

Forced-Choice Response Formats

Forced-choice response formats entail the use of equally desirable response options or items and force the test-taker to choose between them. In spite of their intended goal of reducing response distortion, the research evidence suggests that forced-choice strategies are not immune to intentional distortion (Hough, 1998). Waters (1965), in a review of forced-choice inventories, concluded that individuals are able to distort their responses. And so although it is posited that forced-choice scales may reduce response distortion (Hirsh & Peterson, 2008; Jackson, Wroblewski, & Ashton, 2000; Snell, Sydell, & Lueke, 1999), they do not eliminate response distortion (Christiansen, Burns, & Montgomery, 2005; Converse, Oswald, Imus, Hedricks, Roy, & Butera, 2008; Heggstad, Morrison, Reeve, & McCloy, 2006). Furthermore, ipsative scales make comparisons between applicants difficult (Hough, 1998; Heggstad, Morrison, Reeve, & McCloy, 2006; McCloy, Heggstad, & Reeve, 2005).

Empirical Keying

Empirical keying, which is commonly associated with biodata inventories, is a scoring procedure that focuses on the prediction of an external criterion using keying procedures at either the level of the item or item-option. In using empirical keys, the valence of specified items and item-options is less transparent and permits the design of a more subtle assessment of the non-cognitive constructs of interest. Consequently, subtle scales are posited to be less susceptible to response distortion than obvious, more transparent scales. However, although they may minimize it, empirical keying does not eliminate response distortion since research suggests that empirically keyed measures are still susceptible to response distortion (Hough, 1998; Kluger, Reilly, & Russell, 1991; Viswesvaran & Ones, 1999).

Warnings, Verification, and Threats

Warnings, verification, and threats take the form of informing test-takers that their answers will or can be verified (McFarland, 2003; McFarland & Ryan, 2000; Pace & Borman, 2006; Vasilopoulos, Cucina, & McElreath, 2005). Assessments of the effectiveness of this deterrence strategy have yielded mixed results (Converse, Oswald, Imus, Hedricks, Roy, & Butera, 2008; Dwight & Donovan, 2003; Hough, 1998) because of the possible unintended consequences resulting from their use (Robson, Jones, & Abraham, 2008; Vasilopoulos, Cucina, & McElreath, 2005). For instance, Vasilopoulos, Cucina, and McElreath found that a warning strategy had the unwanted consequence of increasing the complexity of the personality measure. They found that a personality measure preceded by a warning was correlated with a measure of cognitive ability and concluded that the warnings of verification made responding so complex that the personality measure was measuring cognitive ability to a limited degree.

Interactive Prompts or Cautions

Because of the technology with which they are administered, the inconsistency responding approach previously described can be extended into a deterrence strategy. Specifically, pairs of items for which one would expect test-takers to respond consistently can be designed into the test and whenever a test-taker responds

to the second item of a pair inconsistently, the system could be programmed to give some sort of interactive prompt to encourage or caution the test-taker to pay attention to the item content and respond accordingly. Although this approach is novel, it can to some extent be considered to be a variation of the warnings approach discussed above. It is also worth noting that distorting consistently to *both* elements of an item pair makes it difficult, if not impossible, to detect distortion using this approach.

Elaboration

Elaboration is a strategy whereby test-takers are asked to elaborate on their responses. For example, Schmitt and Kuncze (2002) asked applicants to provide an elaboration only if they had endorsed specific responses (sometimes, often, or very frequently) to a question (such as, How often have you rearranged files [business, computer, personal] to make them more efficient in the past year?). The efficacy of this strategy is based on the premise that by asking test-takers to elaborate on and provide detailed follow-up responses to their answers, they are less likely to distort their responses because they would also have to concoct an elaboration as well. Although elaboration strategies appear to be somewhat effective, these strategies may also introduce unwanted consequences. For example, Schmitt, Oswald, Kim, Gillespie, Ramsay, and Yoo (2003) found that requesting respondents to elaborate only if they endorsed specific responses resulted in lower overall test scores. Thus, it appears that elaboration may discourage all applicants from endorsing responses that require additional work. However, Schmitt, Oswald, Kim, Gillespie, Ramsay, and Yoo report that the correlation between test scores and performance were equivalent across elaboration and no elaboration conditions.

Profile Matching and Nonlinear Models

Profile matching entails the matching or fitting of a pattern of applicant scores across multiple dimensions or constructs to some ideal or standard profile. Thus, the use of profile matching or profile similarity indices (which are used extensively with measures such as the Guilford-Zimmerman Temperament Survey [GZTS; Guilford, Guilford, & Zimmerman, 1978]), are

an attempt to compare two sets of multiple personality dimensions (for example, profiles) representing, for example, an applicant and an “ideal” employee, via a single score or index that provides information on the degree of congruence, similarity, or match between the two profiles. Profile similarity indices used in congruence research can be classified into one of two categories—those representing the correlation between the two profiles and those based on the sum of differences between profile elements (personality variables or dimensions; Edwards, 1993). Edwards (1993) presents a detailed description and review of specific indices of these two types of profile similarity indices along with a discussion of methodological problems associated with their use in congruence research, including discarding information regarding the absolute level of the profiles, along with the direction of their difference, and with correlations, the magnitude of the difference as well. He also notes that profile similarity indices mask which elements are responsible for the differences between the profiles. Given these methodological problems, Edwards recommends polynomial regression procedures and shows how they may be used to avoid the problems with profile similarity indices while capturing the underlying relationships profile similarity indices are intended to represent. (The reader is referred to Edwards for a more in-depth, detailed coverage of these issues. Also see Kristof [1996] for additional discussion of these issues and some limitations associated with polynomial regression analysis.)

Inherent in the use of profile matching is an implicit, if not explicit, recognition of the nonlinearity of the specified relationships. The issue of response distortion in the extant literature and its associated discussion therein is predicated on the assumption that the test scores are being used in a linear fashion such that higher scores on the focal constructs are generally deemed to be linearly better (cf. Arthur, Glaze, Villado, & Taylor, 2010). However, one can envisage several conceptually and theoretically sound scenarios whereby the relationship between personality variables and job performance is better conceptualized as being nonlinear. For instance, one can envisage a situation in which moderate levels of agreeableness may be related to effectiveness in customer relations, with low and high levels of agreeableness, on

the other hand, being somewhat counter-productive (Graziano & Eisenberg, 1997; Graziano, Jensen-Campbell, & Hair, 1996). Likewise, it *is* possible to be too conscientious to perform certain roles effectively as is reflected in the obsessive-compulsive label. Murphy (1996) comments on this possibility when he notes that an individual who is high on conscientiousness “might be so conventional and rule-bound that he or she cannot function in anything but the most bureaucratic setting” (p. 22).

Accordingly, the relationships among various personality constructs and job performance may be better conceptualized, under some circumstances, as being nonlinear. Thus, assuming there is empirical support for such an approach (see Day & Silverman, 1989; Robie & Ryan, 1998; Robins, 1995; Scarborough, 1996; Sinclair & Lyne, 1997 as examples of studies that have explored nonlinearity in the relationships between personality variables and job performance), profile matching, coupled with its underlying use of nonlinear models, may mitigate concerns about response distortion, specifically, the uniform elevation of scores. So, to the extent that the specific profile or score configuration is unknown to test-takers, coupled with the fact that the ideal profile or score configuration is usually organizationally job-specific, one would expect profile matching approaches to be less susceptible to the ubiquitous effects of response distortion, especially that which takes the form of across the board elevation of scores, particularly on dimensions that are fairly transparent. In summary, although we were unable to locate any empirical research that investigated the efficacy and effectiveness of the use of nonlinear models as a deterrent technique, response distortion may be less of an issue or concern with this approach because dissimulation to obtain high construct-level scores does not necessarily imply or result in the successful faking of the optimal profile. So, because scores are not being used in a linear fashion, response distortion may be less of a concern.

Conclusion

It is widely accepted that the remote delivery of assessments in the context of personnel selection and related employment decision-making is increasingly becoming a common practice

(Tippins, 2009). However, associated with this practice are continuing concerns about the veracity of the assessment or test scores that are obtained from this mode of assessment (for example, see focal article and commentaries in *Industrial and Organizational Psychology: Perspectives on Science and Practice* [2009]). Hence the primary objective of this chapter was to provide the reader with a succinct review of the state of the literature on cheating and response distortion on remotely delivered ability and knowledge and nonability and noncognitive assessments, respectively. In this regard, we distinguished between two forms of malfeasant behaviors—cheating and response distortion. We defined cheating as being associated with ability and knowledge tests and entailing the use of illicit aids to obtain and produce the keyed or correct answers to tests. On the other hand, we defined response distortion as being associated with noncognitive measures and referring to deliberate falsification of one's responses to self-report items with the goal of presenting oneself in as favorable a light as possible.

Within the context of the preceding framework, we sought to answer a number of questions, the answers to which can be summarized as follows. First, within the context of ability and knowledge tests, some subset of test-takers will engage in cheating if the testing environment is permissive to such attempts. Although proctored testing cannot be considered the “gold standard” in curtailing cheating, unproctored tests would seem to be more permissive to overt and less clandestine cheating attempts compared to proctored testing. Test-takers may gain illegal access to test content (via hacking or pirating behaviors), collude (for example, use a smart friend as a surrogate), or access illicit information. So, although there is limited empirical evidence that speaks to the levels of cheating on unproctored Internet-based ability and knowledge tests, it is not unreasonable to posit that cheating will occur if left unchecked.

Proctoring is the primary means by which cheating attempts on ability and knowledge tests are detected. Proctoring serves multiple purposes which consist of (1) verifying test-takers' identities, (2) detecting the use of illicit materials (for example, crib sheets), and (3) detecting test-takers sharing information during the test administration. Thus, in the absence of a test proctor, these

sources of cheating must be curtailed using alternative methods and technologies. As a summary statement, the efficacy and effectiveness of these alternative methods and technologies are not well understood. For example, the use of webcams as a means of electronic monitoring may be intuitively appealing, but in the absence of empirical data, their efficacy remains unknown. Furthermore, it is unclear how feasible the use of some of these technologies are for unproctored testing (for example, iris and retina scans). That is, although it may be reasonable to expect test-takers to have access to webcams, it may be unreasonable to expect them to have access to biometric identification equipment. In the absence of bona fide detection techniques, the use of proctored retesting is the only viable method of detecting cheating. However, proctored retesting reduces the cost-effectiveness and other advantages of unproctored testing (test anywhere at any time). Furthermore, there is currently no consensus on the extent to which unproctored and proctored test scores must diverge to raise concerns and warrant corrective action.

Regardless of the specific methods or technologies used to detect it, the issue of taking corrective action when cheating is detected is a complex one. For instance, one must first determine what constitutes evidence of cheating. The academic and educational testing literature reflects concerns about accusing or confronting test-takers about cheating (Cizek, 1999). Furthermore, professional and ethical guidelines require test-takers to be informed of testing irregularities, and the implications of test-takers reactions to being accused of cheating has not been investigated. On a related note, there is a dearth of information regarding the legal implications of canceling or correcting applicants' test scores because of suspected cheating.

Given the complexity and difficulty of detecting cheating and compiling convincing evidence that cheating has occurred, attempting to deter cheating may be more efficacious. Approaches to deterrence can take one of two forms: (1) increasing the saliency of detection methods and (2) using test design characteristics that make it more difficult to cheat. Increasing the saliency of detection techniques can take the form of overt monitoring and warnings and threats. Although the efficacy of these methods is unknown, it is reasonable to posit that warnings

(especially of proctored retesting) may curtail cheating attempts. Regarding test design characteristics, the use of speeded tests (Arthur, Glaze, Villado, & Taylor, 2010), measuring constructs that are difficult to cheat on (Nye, Do, Drasgow, & Fine, 2008), limiting item exposure (Drasgow, Nye, Guo, & Tay, 2009), and guarding against preknowledge of test content would seem to be effective strategies. However, these methods engender some disadvantages. Using speeded tests is limited to situations in which a speeded test is consonant with the job-relatedness of the test, and only a limited number of constructs (for example, perceptual speed) may be particularly resistant to some forms of cheating. In addition, limiting item exposure and preknowledge of test content may be administratively difficult.

So, as a summary statement, it is reasonable to assume that cheating will be quite widespread on unproctored ability and knowledge tests if left unchecked. Methods for curtailing cheating include deterring and detecting cheating behaviors. However, most of the methods and techniques for curtailing cheating behavior have limited or no empirical support. Furthermore, there are theoretical, practical, ethical, and legal issues that are yet unresolved regarding deterring and detecting cheating, the evidence required to take corrective action, and the proper corrective action to take.

For noncognitive and nonability tests and measures, response distortion on unproctored Internet-based tests seems to be no more pervasive or extensive than that for proctored testing since proctoring is not a method for curtailing response distortion. Concerning methods for detecting and deterring response distortion, by virtue of the technology by which they are administered, there are methods that are unique to unproctored Internet-based tests, and those that are common or applicable to both proctored and unproctored Internet-based tests. The latter include the use of forced-choice response formats, empirical keying, warnings, verifications, and threats, elaborations, lie scales, and indicators of inconsistency responding. Response distortion detection and deterrence approaches that are unique to unproctored Internet-based tests include the use of response latencies and interactive prompts or cautions. Compared to the literature

on cheating on unproctored ability and knowledge tests, there is relatively more empirical literature investigating the effectiveness of these methods, but most, if not all, of it has been conducted in the context of proctored tests. However, as previously noted, there is no conceptual reason to suspect that these findings would not generalize to unproctored settings. Finally, as with cheating on ability and knowledge tests, there are unresolved issues regarding deterring and detecting response distortion on noncognitive measures, the evidence required to take corrective action, and the proper corrective action to take.

In addition to the preceding summary statements, there are a number of noteworthy observations. First, it is important to note that the proctored versus unproctored delivery of assessments has more profound effects on and implications for ability and knowledge tests than it does for noncognitive and nonability measures primarily because proctoring is *not* a means for controlling for response distortion on the latter type of assessments. As a result, for noncognitive tests and measures, response distortion issues are common across *both* unproctored and proctored settings (Arthur, Glaze, Villado, & Taylor, 2010; Bartram & Brown, 2004; Gupta, 2007; Kaminski & Hemingway, 2009; Templer & Lange, 2008). However, because of its technological mode of administration, there are some issues, such as the use of response latencies for detection and interactive prompts as a deterrent, that may be more commonly associated with the remote delivery of noncognitive and nonability measures than they are with ability and knowledge tests.

Second, although we reviewed and discussed them in a singular fashion, for most, if not all, the detection and deterrence techniques for both ability and noncognitive measures, multiple methods *can* be used in conjunction. That is, the use of these methods and techniques is not mutually exclusive. So, for instance, for noncognitive measures, response latencies may be used in addition to lie scales, forced-choice response formats, and empirical keying. Consequently, research investigating the efficacy and effectiveness of various combinations of techniques would be informative. However, even in the absence of said research, it does not seem unreasonable to posit that the conjunct use of

various techniques could be more effective than the use of any one method by itself.

Third, the amount of research in this domain is very limited and is definitely not commensurate with the use of these types of assessments in practice. As a result of this, we were unable to directly comment on or evaluate the effectiveness and efficacy of several of the detection and deterrence methods such as the use of verification retesting (as a deterrent) and webcams (as a detection method). Future research should investigate, for example, the extent to which verification retesting deters cheating on initial testing and applicant reactions to this method. As previously mentioned, the use of verification retesting implicitly undermines the perceived veracity of initial unproctored test scores. Furthermore, in the absence of data to the contrary, it is not unreasonable to posit that requiring test-takers to own or obtain webcams may reduce the number of unproctored test-takers. This problem would be exacerbated if ownership of webcams varied as a function of status on a protected class variable.

Fourth, the distributional placement (position in the score distribution) of malfeasant responders (cheaters and those who distort their responses) is a critical issue. Arthur, Glaze, Villado, and Taylor (2010) argue that the impact of malfeasant responding on employment decisions is partially a function of distributional placement. That is, if the preponderance of malfeasant responders is in the low end of the distribution, malfeasant responding may have minimal effects. However, if malfeasant responders are evenly distributed across the score range, or reside in the upper end of the distribution, then cheating and response distortion become relatively more critical issues. Similarly, Arthur, Glaze, Villado, and Taylor provide data that suggests that those suspected of cheating on a cognitive ability test were evenly distributed across the score range, with a slight trend of having more cheaters in the upper end of the distribution. In contrast, Impara, Kingsbury, Maynes, and Fitzgerald (2005) found that cheating occurs across the score range with the exception of those at the high end of the distribution. In the context of nonability tests, Arthur, Glaze, Villado, and Taylor found that response distortion occurred across the score range, with a slight trend towards having more response distortion

in the upper end of the score distribution. Griffith, Chmielowski, and Yoshita (2007) examined the proportion of applicants who were suspected of distorting their responses that would be hired under varying selection ratios. Their results indicated that with a 50 percent selection ratio, 31 percent of the test-takers in the hiring range would *not* have been hired using their honest scores; this number increased to 66 percent with a 10 percent selection ratio. Similarly, inflating self-report SAT scores appears to be negligible for test-takers in the upper and midrange of SAT scores, however those in the lower end of the distribution tended to inflate their SAT scores (Newman & Lyon, 2009).

Fifth, there are a number of less commonly used approaches for detecting response distortion that are nevertheless, worth noting. For example, Zickar and Drasgow (1996) present an item-response theory-based theta-shift model for detecting response distortion. Specifically, they proposed that test-takers who distort their responses respond to a subset of items honestly but respond to some items in an inaccurately favorable manner. Thus, test-takers whose response patterns reflect a theta-shift are suspected of response distortion. As a summary statement, the theta-shift model resulted in a lower number of false positives compared to a social desirability scale. Zickar and Robie (1999) provided convergent evidence for the theta-shift model by examining response distortion at both the item- and scale-level of analysis. Their findings suggest that for test-takers with the same level of the underlying construct, those who distorted their responses were more likely to endorse a more extreme positive response. Both of these studies demonstrate the viability of item-response theory-based methods for detecting response distortion.

In conclusion, it is widely recognized that merit-based public-sector selection and promotion testing, especially municipal safety forces (police and fire), is extremely litigious. Thus, one could say that unproctored Internet-based testing has really “arrived” and overcome its security threats and concerns when it is widely used and accepted in this type of testing environment and setting. This would currently appear to be the case for bio-data and training and experience measures, but less so, if not virtually nonexistent, for ability and knowledge-based assessments.

References

- Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement, 60*, 59–72.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Anderson, C. D., Warner, J. L., & Spector, C. E. (1984). Inflation bias in self-assessment examination: Implications for valid employee selection. *Journal of Applied Psychology, 69*, 574–580.
- Arthur, W., Jr., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed-response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology, 55*, 985–1008.
- Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2009). Unproctored internet-based tests of cognitive ability and personality: Magnitude of cheating and response distortion. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 39–45.
- Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*, 1–16.
- Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection. *Journal of Applied Psychology, 93*, 435–442.
- Arthur, W., Jr., Woehr, D. J., & Graziano, W. G. (2001). Personality testing in employment settings: Problems and issues in the application of typical selection practices. *Personnel Review, 30*, 657–676.
- Automatic Data Processing Inc. (2008). *2008 ADP screening index*. Retrieved October 4, 2009, from www.adp.com/media/press-releases/2008-news-releases/adp-annual-pre-employment-screening-index.aspx.
- Bartram, D. (2009). The International Test Commission guidelines on computer-based and internet-delivered testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 11–13.
- Bartram, D., & Brown, A. (2004). Online testing: Mode of administration and the stability of OPQ 32i scores. *International Journal of Selection and Assessment, 12*, 278–284.

- Beatty, J. C., Jr., Fallon, J. D., Shepherd, W. J., & Barrett, C. (2002, April). Proctored versus unproctored web-based administration of a cognitive ability test. Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.
- Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317–335.
- Bott, J., O'Connell, M., & Doverspike, D. (2007). Practical limitations in making decisions regarding the distribution of applicant personality test scores based on incumbent data. *Journal of Business and Psychology, 22*, 123–134.
- Burke, E. (2009). Preserving the integrity of online testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 35–38.
- Cellar, D. F., Miller, M. L., Doverspike, D., & Klawnsky, J. D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five factor model. *Journal of Applied Psychology, 81*, 694–704.
- Chang, H-H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387–398.
- Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267–307.
- Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47*, 847–860.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*, 156–169.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
- Crowne, D. P., & Marlowe, D. A. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349–354.
- Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology, 42*, 25–36.

- Delery, J. E., & Kacmar, K. M. (1998). The influence of applicant and interviewer characteristics on the use of impression management. *Journal of Applied Social Psychology, 28*, 1649–1669.
- Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: Born to deceive, yet capable of providing valid self-assessments. *Psychology Science, 48*, 209–225.
- Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*, 81–106.
- Drasgow, F., Nye, C. D., Guo, J., & Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 46–48.
- Dwight, S. A., & Alliger, G. M. (1997). Using response latencies to identify overt integrity test dissimulation. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, St Louis, Missouri.
- Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1–23.
- Edens, P. S., & Arthur, W., Jr. (2000). A meta-analysis investigating the susceptibility of self-report inventories to distortion. Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, Louisiana.
- Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology, 46*, 641–665.
- Edwards, B. D., & Arthur, W., Jr. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92*, 794–801.
- Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92*, 386–395.
- Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 215–224.
- Ellingson, J. E., Smith, B. D., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86*, 122–133.

- Ellis, A. P. J., West, B. J., Ryan, A. M., & DeShon, R. P. (2002). The use of impression management tactics in structured interviews: A function of question type? *Journal of Applied Psychology, 87*, 1200–1208.
- Fluckinger, C. D., McDaniel, M. A., & Whetzel, D. L. (2008). Review of faking in personnel selection. In M. Mandal (Ed.), *In search of the right personnel* (pp. 91–109). New Delhi: Macmillan.
- Foster, D. (2009). Secure, online, high-stakes testing: Science fiction or business reality? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 31–34.
- Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and initial survey of researchers. *International Journal of Selection and Assessment, 11*, 340–344.
- Gough, H. G., & Bradley, P. (1996). *CPI manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
- Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 795–824). San Diego, CA: Academic Press.
- Graziano, W. G., Jensen-Campbell, L. A., & Hair, E. C. (1996). Perceiving interpersonal conflict and reacting to it: The case for agreeableness. *Journal of Personality and Social Psychology, 70*, 820–835.
- Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*, 341–355.
- Grubb, W. L., & McDaniel, M. A. (2007). The fakability of Bar-On's Emotional Quotient Inventory Short Form: Catch me if you can. *Human Performance, 20*, 43–59.
- Guilford, J. P., Guilford, J. S., & Zimmerman, W. S. (1978). *Manual for the Guilford-Zimmerman Temperament Survey*. Hanover, PA: Sheridan Press.
- Guilford, J. S., Zimmerman, W. S., & Guilford, J. P. (1976). *The Guilford-Zimmerman Temperament Survey handbook: Twenty-five years of research and application*. San Diego, CA: EdITS.
- Guo, J., & Drasgow, F. (2009). Identifying dishonest job applicants in unproctored internet testing: The Z-test and the likelihood ratio test. Manuscript submitted for publication.
- Gupta, D. (2007). Proctored versus unproctored online testing using a personality measure: Are there any differences. Unpublished doctoral dissertation. Denton, TX: University of North Texas.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying* (ACT

- Research Report Series No. 87-15). Iowa City, IA: American College Testing.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty-Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.
- Heggstad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24.
- Hense, R., Golden, J. H., & Burnett, J. (2009). Making the case for unproctored internet testing: Do the rewards outweigh the risks? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 20-23.
- Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. *Journal of Research in Personality, 42*, 1323-1333.
- Hogan, R. T., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.
- Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270-1285.
- Holden, R. R. (1995). Response latency detection of fakers on personnel tests. *Canadian Journal of Behavioral Science, 27*, 343-355.
- Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63*, 272-279.
- Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*, 209-244.
- Hough, L. M., Eaton, N. L., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581-595.
- Hough, L. M., & Furnham, A. (2003). Use of personality variables in work settings. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 131-169). Hoboken, NJ: John Wiley & Sons.
- Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 271-290.

- Industrial and Organizational Psychology: Perspectives on Science and Practice*. (2009), 2(1).
- Impara, J. C., Kingsbury, G., Maynes, D., & Fitzgerald, C. (2005). Detecting cheating in computer adaptive tests using data forensics. Paper presented at the 2005 annual meeting of the National Council on Measurement in Education and the National Association of Test Directors, Montreal, Canada.
- Inwald, R. E. (1992). *Inwald Personality Inventory technical manual* (rev. version 2A). New York: Hilson Research.
- International Test Commission. (2005). *International guidelines on computer-based and internet delivered testing: Version 2005*. Downloaded electronically October 12, 2009, from [http://www.intestcom.org/Downloads/ITC percent20Guidelines percent20on percent20Computer percent20- percent20version percent202005 percent20approved.pdf](http://www.intestcom.org/Downloads/ITC%20Guidelines%20on%20Computer%20-%20version%202005%20approved.pdf).
- Jackson, D. N. (1999). *Personality Research Form manual* (3rd ed.). Port Huron, MI: Sigma Assessment Systems.
- Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance*, 13, 371–388.
- Kaminski, K. A., & Hemingway, M. A. (2009). To proctor or not to proctor? Balancing business needs with validity in online assessment. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2, 24–26.
- Kluger, A. N., Reilly, R. R., & Russell, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology*, 76, 889–896.
- Kristof, A. L. (1996). Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology*, 49, 1–49.
- Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, 15, 220–231.
- Leonard, J. A. (1996). Response latency as an alternative measure of performance on honesty tests. Unpublished doctoral dissertation. Tampa, FL: University of South Florida.
- Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment*, 14, 299–316.
- Levin, R. A., & Zickar, M. J. (2002). Investigating self-presentation, lies, and bullshit: Understanding faking and its effects on selection

- decisions using theory, field research, and simulation. In J. M. Brett & F. Drasgow (Eds.), *Psychology of work* (pp. 253–276). Mahwah, NJ: Lawrence Erlbaum Associates.
- Lievens, F., Reeve, C. L., & Heggstad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*, 1672–1682.
- Lönnqvist, J.-E., Paunonen, S., Tuulio-Henriksson, A., Lönnqvist, J., & Verkasalo, M. (2007). Substance and style in socially desirable responding. *Journal of Personality, 75*, 291–322.
- McCloy, R. A., Heggstad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multi-dimensional forced-choice items. *Organizational Research Methods, 8*, 222–248.
- McFarland, L. A. (2003). Warning against faking on a personality test: Effects on applicant reactions and personality tests scores. *International Journal of Selection and Assessment, 11*, 265–276.
- McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across non-cognitive measures. *Journal of Applied Psychology, 85*, 812–821.
- McFarland, L. A., & Ryan, A. M. (2006). Towards an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36*, 979–1016.
- McFarland, L. A., Yun, G., Harold, C. M., Viera, R., Jr., & Moore, L. (2005). An examination of impression management use and effectiveness across assessment center exercises: The role of competency demands. *Personnel Psychology, 58*, 949–980.
- Michaelis, W., & Eysenck, H. K. (1971). The determination of personality inventory factor patterns and inter-correlations by changes in real-life motivation. *Journal of Genetic Psychology, 118*, 223–234.
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*, 1029–1049.
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729.
- Murphy, K. R. (1996). Individual differences and behavior in organizations: Much more than *g*. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 3–30). San Francisco: Jossey-Bass.

- Newman, D. A., & Lyon, J. S. (2009). Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability, and diversity. *Journal of Applied Psychology, 94*, 298–317.
- Nye, C. D., Do, B., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment, 16*, 112–120.
- Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995–1027.
- One, 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.
- Pace, V. L., & Borman, W. C. (2006). The use of warnings to discourage faking on noncognitive inventories. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 283–304). Greenwich, CT: Information Age Publishing.
- Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire* (pp. 143–165). New York: Springer-Verlag.
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of the construct. In H. Braun, D. Jackson, & D. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Lawrence Erlbaum Associates.
- Pearlman, K. (2009). Unproctored internet testing: Practical, legal, and ethical concerns. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 14–19.
- Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21*, 489–510.
- Robie, C., Curtin, P. J., Foster, C. T., Philips, H. L., Zbylut, M., & Tetrick, L. E. (2000). *Canadian Journal of Behavioural Science, 32*, 226–233.
- Robie, C., & Ryan, A. M. (1998). Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance. In D. Ones (Chair), *Multiple predictors, situational influences, and incremental validity*. Symposium presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, Texas.
- Robins, K. W. (1995). Effects of personality and situational judgment on job performance. *Dissertation Abstracts International, 55(9-B)*: 4155.

- Robson, S. M., Jones, A., & Abraham, J. (2008). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance, 21*, 89–106.
- Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Human Performance, 83*, 634–644.
- Saupe, J. L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement, 20*, 475–490.
- Scarborough, D. J. (1996). An evaluation of back propagation neural network modeling as an alternative methodology for criterion validation of employee selection testing. *Dissertation Abstracts International, 56*(8-B), 4624.
- Schmit, M. J., & Ryan, M. A. (1993). The big five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*, 966–974.
- Schmitt, N., & Kuncze, C. (2002). The effects of required elaboration of answers to biodata questions. *Personnel Psychology, 55*, 569–587.
- Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology, 91*, 613–621.
- Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., & Yoo, T. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. *Journal of Applied Psychology, 88*, 979–988.
- SHL Group. (2000). *Expert report: OPQ32 Version n*. Boulder, CO: Author.
- Sinclair, R. R., & Lyne, R. (1997). Non-linearity in personality-performance relations: Models, methods, and initial evidence. In J. Hogan (Chair), *Non-cognitive measures of job performance*. Symposium presented at the annual conference of the Southwestern Psychological Association, Fort Worth, Texas.
- Smith, D. B., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87*, 211–219.
- Smith, B., & Robie, C. (2004). The implications of impression management for personality research in organizations. In B. Schneider & B. D. Smith (Eds.), *Personality and organizations* (pp. 111–138). Mahwah, NJ: Lawrence Erlbaum Associates.
- Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9*, 219–242.

- Stanush, P. L. (1996). Factors that influence the susceptibility of self-report inventories to distortion: A meta-analytic investigation. Unpublished doctoral dissertation. College Station, TX: Texas A&M University.
- Templer, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unproctored field conditions. *Computers in Human Behavior, 24*, 1216–1228.
- Tett, R. P., Anderson, M. G., Ho, C., Yang, T. S., Huang, L., & Hanvongse, A. (2006). Seven nested questions about faking on personality tests: An overview and interactionist model of item-level response distortion. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 43–84). Greenwich, CT: Information Age Publishing.
- Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology, 60*, 967–993.
- Tippins, N. P. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 2–10.
- Tippins, N. P., Beaty, J., Drasgow, F., Wade, M. W., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology, 59*, 189–225.
- Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Psychological Science, 5*, 243–262.
- Vasilopoulos, N. L., Cucina, J. M., & McElreath, J. M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology, 85*, 306–322.
- Vasilopoulos, N. L., Reilly, R. R., & Leaman, J. A. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85*, 50–64.
- Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210.
- Waters, L. K. (1965). A note on the “fakability” of forced-choice scales. *Personnel Psychology, 18*, 187–191.
- Whitley, B. E., Jr. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39*, 235–274.
- Wood, J. L., Schmidke, J. M., & Decker, D. L. (2007). Lying on job applications: The effects of job relevance, commission, and human

resource management experience. *Journal of Business Psychology*, 22, 1–9.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, 20, 71–87.

Zickar, M. J., & Robie, C. (1999). Modeling fake good on personality items: An item-level analysis. *Journal of Applied Psychology*, 84, 551–563.

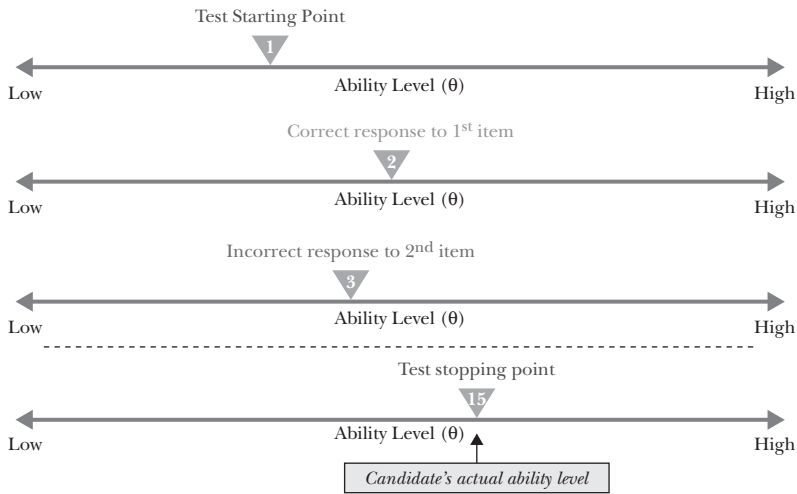
Chapter Five

COMPUTERIZED ADAPTIVE TESTING

Rodney A. McCloy and Robert E. Gibby

Dating from the 1960s, *computerized adaptive tests*, or CATs, have been developed to meet a host of assessment needs. The basic premise behind a CAT is that it provides a test examiner or administrator the ability to individually assess a respondent by selecting and presenting items based on the respondent's ability or trait level (θ). Unlike a respondent's true score in classical test theory (Allen & Yen, 2002; Crocker & Algina, 1986), which is conditional on the test (or set of items) in question, the θ (or ability) score is a characteristic of the respondent and is independent of test content (Lord & Novick, 1968). Typically starting with an item around the middle of the θ distribution for the construct being assessed (an item of moderate difficulty), a CAT chooses and presents an item from a pool of items that has been calibrated against an item response theory (IRT; De Ayala, 2009; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Van der Linden & Hambleton, 1997) model that describes response behavior.¹ The respondent then answers the item, and the CAT estimates a new provisional θ for the respondent based on the information provided by the response to this most recently administered item. Given this new estimated θ , the adaptive algorithm then goes to the item pool and selects the next item to present. All things being equal, the next item to

Figure 5.1. Illustration of the Relation Between a Candidate's Response and Item Selection in a Computerized Adaptive Test



present should be the one that provides the maximal amount of information about the respondent's θ . As an illustration, consider Figure 5.1.

Through the selection and scoring of multiple items, the test adapts item presentation to the respondent based on his or her answers to the previous items, resulting in a test that is neither too easy nor too difficult (although high-ability respondents have been known to report that a CAT seems like the most difficult test they have ever taken—exactly because it does not waste time administering many items that are “too easy” for them). This result is a critical benefit of adaptive tests, as the test administrator is able to learn the most about the respondent when the test questions are at the same level as the respondent's ability level.

Building on the seminal work of Lord (1980; Lord & Novick, 1968), Weiss (1982, 1983, 1985), and others in the areas of item response theory and individually tailored testing, the development of CATs increased significantly in the 1980s and 1990s (see Drasgow & Olson-Buchanan, 1999; Van der Linden & Glas, 2000; Wainer, 1990). In the public sector, the U.S. Department of Defense (DoD) developed a CAT program

around its paper-and-pencil Armed Services Vocational Aptitude Battery (ASVAB) that, with modification, remains in operation today (Sands, Waters, & McBride, 1997). In the private sector, State Farm (Zickar, Overton, Taylor & Harms, 1999), Procter & Gamble, and other companies developed CATs for pre-employment selection of candidates. In the case of Procter and Gamble (P&G), the Computer-Adaptive Reasoning ASDF Test (CARAT), developed in 1993, was in operation for only one year because it became too challenging to wheel laptops from offices to campuses or other test locations (remember that personal computers were not widely available in 1993). For educational testing, the Educational Testing Service in the United States began development of a CAT version of the Graduate Record Examination (GRE) in 1988 and implemented it in 1993 (ETS, 1993). In The Netherlands, the National Institute for Educational Measurement developed two CATs to assess mathematics ability for placement and assessment of achievement among adult students (Verschoor & Straetmans, 2010).

Over time as accessibility of personal computers has increased throughout the world, CATs have become more common in pre-employment selection, credentialing, and educational achievement in content areas ranging from language and music ability to medical knowledge and clinical skills. In addition, the use of the Internet has seen an increase in the development and deployment of CATs that are administered in an unproctored environment, especially for pre-employment selection and language assessment on a global scale. The purpose of this chapter is to provide *practical* considerations for the use, development, deployment, and ongoing maintenance of a CAT based on the literature and through our own development and use of adaptive testing programs for the selection and placement of U.S. Armed Services personnel and for selection of pre-employment candidates for P&G's global hiring process.

Considerations for the Use of Computerized Adaptive Testing

Given the wide range of use of adaptive testing over the last four decades, it makes sense to conclude that the use of CAT offers many advantages to those needing to assess respondents on

some ability or domain of knowledge. It also makes sense that there are unique challenges and concerns that have to be overcome with CATs, especially in industry when used to select and place employees. This section provides an overview of the practical advantages and challenges that surround the use of a CAT.

Advantages of CAT

Dating back to the early development and use of CATs, Green (1983) noted several advantages for their use relative to paper-and-pencil (P&P) tests, including delivery of tests at the appropriate level of difficulty for respondents, improved test security, elimination of physical problems associated with written tests, immediate scoring, improved pretesting of items, easier elimination of faulty items, and the ability to implement a broader array of question, stimulus and response types. Meijer and Nering (1999) added the benefits of shorter tests, enhanced measurement precision, and the ability to test on demand.

Greater Precision

Building on this list, we argue that the primary advantage for use of adaptive testing is that it provides a more accurate estimate of a respondent's theta (ability score) than that provided by traditional tests. By using information about how an item measures the underlying ability or trait, combined with how the respondent has answered the previous items on the test, a CAT is able to present items close to the candidate's true theta level. As the respondent moves through the adaptive test, each answer provides more information on which to base the respondent's score.

Shorter Testing

In this way, an adaptive test is typically able to use fewer questions than would be required by a paper-and-pencil test to achieve the same level of score precision (although this is not always the case; cf. Zickar, Overton, Taylor, & Harms, 1999). A shorter test is often seen as an advantage because it requires less administration time and fewer resources to deliver it. In addition, a shorter test requires less time commitment from the respondent, often

resulting in reduced respondent attrition that could be a problem for longer tests. Finally, a shorter test results in the delivery of fewer items, providing greater security for the item pools that underlie the CAT.

More Difficult to Cheat

Given that respondents have the potential to receive completely different items through adaptive delivery and are unlikely to gain the entire item pool through prior administration of the CAT, it is more difficult to cheat on a CAT as compared to a P&P or other computerized test. This result is even more prevalent for the use of CATs under unproctored testing conditions. Although use of a CAT cannot ensure that the respondent taking the test under unproctored conditions is the same one who shows up to be hired or placed in a job, this capacity to reduce the amount of exposed test content means that CATs provide a more effective unproctored screening method than many other alternatives, especially when combined with a supervised verification test.

Improved Scoring

Because CAT scoring is computerized, there are fewer scoring errors relative to what can be expected from paper-and-pencil test scoring procedures (for example, manual scoring, optical scanners). In addition, CAT ability scores are available as soon as the testing process is complete. Therefore, test results can be immediately provided to key stakeholders (including candidates) who need to make decisions.

Improved Test Security

As Green (1983) noted, test security is greatly enhanced with a CAT compared to delivery of a P&P test, because it is difficult to improve performance on the test by learning only a few items. This benefit results from the typically large number of items in the item pool that can be drawn upon and delivered adaptively to the respondent. CATs also provide unique security benefits as compared to other computerized testing methods (for example, on-the-fly test generation, forms casting). For example, CATs often increase security through item

exposure control algorithms, item pool rotation, and continual item research, calibration, and ongoing refreshment of items. These techniques are uniquely facilitated in a CAT because of its focus at the item level. Given the costs that go into the initial development of the item pool, the ability to manage test security is paramount to the success of an adaptive testing program.

Easier Ongoing Maintenance

Independent of improved security, the ability to upgrade, refresh, and retire items in a CAT's item pool through the delivery of research content to respondents and subsequent calibration of that content is an important advantage of adaptive tests. Once a research program is established, even ongoing item development, research, and calibration can be automated through technology to produce expansive item pools that cover the full range of the trait or ability assessed by the test. This ability of a CAT is especially important for global pre-employment or placement testing programs, because it requires less cost and fewer resources to localize and manage ongoing. In addition, the ability to use differential item functioning (DIF; see Embretson & Reise, 2000) procedures to assess the effectiveness and calibration of a localized CAT is critical to effectively extending tests to other cultures.

Unproctored Internet Testing

Ongoing evolution of the algorithms and technology for providing secure and refreshed CATs has resulted in test developers in the public and private sectors becoming more comfortable in recent years with extending adaptive tests to unproctored Internet delivery (typically when combined with supervised verification testing programs, but not always). Combined with the cost and resource reduction, broader respondent reach, and improved speed and efficiency afforded by 24/7 access to testing, along with several other benefits of unproctored Internet testing (UIT; see Gibby, Ispas, McCloy, & Biga, 2009, and Tippins, 2009), the improved security offered by adaptive delivery of test content has created a proliferation of adaptive tests over the last decade.

Reduced Impact of Candidate Preparation

A unique advantage of CATs (relative to P&P or other computerized assessments) is that they eliminate any benefit for respondents who are more savvy or have more experience with taking tests. Because items and responses work together to determine how the adaptive algorithm moves the candidate through the test, respondents typically view and complete one item at a time. In addition, respondents are not allowed to go back to change answers on previous items or view previously exposed content.² As a result, respondents are not able to look at all available test items, identify easier content or item types that may be distributed across a test battery, and then concentrate their efforts on these items first while saving more time-consuming, difficult items for last. The net result is to provide both a testing experience and a final score that are less dependent on the respondent's skill in taking tests.

Elimination of Issues for Physical Test Delivery

Consistent with delivery of other computerized assessments, an advantage of CAT over P&P tests is that the item responses and scoring keys are stored electronically rather than in a desk drawer or filing cabinet. In addition to providing safer storage, encrypted electronic transmission of score results—often through an applicant tracking system as part of pre-employment testing—is frequently safer than transporting paper copies of the test along with answer sheets and scoring grids, because they are less likely to be stolen, misplaced, left in the back of a taxi, given to a girlfriend, or found blowing all over the runway of a major international airport after the box they were packaged in fell off a baggage cart and broke apart with the impact of the fall (yes, all are from our personal experiences with P&P tests). Also consistent with delivery of other computerized assessments, CATs overcome the problems associated with the physical delivery of paper-and-pencil tests. For example, respondents need not worry about erasing an errant response completely, and test administrators do not end up frustrated because of having to score an answer sheet that the respondent failed to complete appropriately. For each answer, the computer can direct the respondent to answer the test according to its design.

Challenges of CAT

Although computerized adaptive testing offers testing specialists many advantages, there are also many challenges that need to be considered and overcome to ensure efforts to develop, implement, and maintain an ongoing adaptive test program. We present several of these challenges next to help you determine whether CAT might be appropriate for your testing needs.

Costs

As described by Meijer and Nering (1999), the primary concern we have encountered in talking with practitioners considering development of an adaptive test is the cost. Perhaps the first large cost encountered in CAT development is that associated with generating sufficient test content (items). Even though adaptive tests deliver fewer questions to any particular respondent, they require development and calibration of many more items than traditional tests to ensure the full range of the ability or trait being assessed is represented in the underlying item pool (that is, they need to provide adequate coverage of the theta distribution).

Advances in item generation programs that permit automated item cloning have made it easier to develop large quantities of test items, but one still needs a large, diverse pool of “seed” or “parent” items that span the full range of theta for proper CAT item pool development. Because cloned items should have item parameters that are at least similar to those of the seed items, the capacity to clone items will be of limited utility for filling any extant gaps in coverage of the theta distribution. Cloning benefits item exposure concerns more than it addresses item coverage woes. In short, item generation and cloning are two separate issues. You still need to generate sufficient initial content before worrying about cloning.

In addition to developing more items than required by P&P tests, there is also the cost associated with delivering the CAT, both for the computers and the delivery software or engine. Unless computers are readily available to the targeted respondents of the CAT, hardware will need to be purchased and loaded with the adaptive delivery software or connected to the Internet

for online delivery. They will also need to be maintained and kept free of viruses or other issues that could have a deleterious effect on the delivery of the test.

Given that access to computers is improving every day across the globe, the bigger issue around the cost of CATs beyond item development has been access to commercially available adaptive delivery systems or software. Because the pull for CAT has been at the level of test publishers, large corporations, and DoD, off-the-shelf adaptive delivery systems have been hard to find, especially for online, unproctored CAT delivery. Part of the reason for this is that the technology and computer infrastructure were not well established until recently. As access to computers and the Internet has increased, organizations, agencies, and test publishers have increasingly moved to adaptive testing solutions for their pre-employment selection and placement systems. For example, DDI, SHLPreVisor, Aon Hewitt, [Aon Hewitt], and other companies have all developed adaptive testing platforms to deliver their own and clients' CAT content.

Data Analysis Requirements

Another challenge for the development and ongoing maintenance of a CAT is the data analysis requirements for calibrating items against an underlying IRT model. Depending on the model chosen,³ the number of items desired for the final item pool, and the localization requirements for the test, the amount of respondent data required for proper calibration and screening of items can be extensive. Typically, a three-parameter logistic (3PL) model requires a minimum of around one thousand responses per item to yield model convergence and thus well estimated parameters. For the sake of comparison, a two-parameter logistic (2PL) model might do well with approximately half that many responses per item, whereas Rasch and one-parameter logistic (1PL) can sometimes be well estimated with as few as 100 responses per item (see De Ayala, 2009, for a more complete treatment).

For CATs developed against a simpler IRT model (that is, one with fewer parameters to estimate, such as the Rasch model) or focused on a single country, the number of items researched and the number of responses needed to effectively manage DIF analysis should be much less than one thousand responses per item.

Depending on the content and purpose of the test, one option for overcoming such extensive data requirements for item calibration and screening might be to use subject-matter experts (SMEs) to calibrate items against an underlying adaptive model (see Fetzer and Kantrowitz's case study, Chapter 15 in this volume). This method eliminates initial data requirements for calibration while still permitting ongoing automation of item research and calibration using respondent data.

Access to IRT Expertise

No matter what methods are used to calibrate content, however, another challenge of developing and maintaining a CAT is access to a resource specializing in or having deep understanding of IRT. Although courses in IRT are becoming more common in education and psychology graduate programs, few people specialize in the area.

Ongoing Maintenance

As described above, CATs have the advantage of easier ongoing item research and calibration. The challenge, however, is that items have to be developed on a continual basis so they can be fed into the adaptive testing process for research and calibration back against the existing item pool. Therefore, the cost issues described above are relevant not only for the initial development of the CAT, but also for its ongoing use. Money and resources will need to be deployed against item writers who are experts in the test topic or for the development of item generation algorithms to automate development of new test questions.

Another challenge for the ongoing maintenance of a CAT is the assessment of whether items continue to function in the same way as when they were initially calibrated and released for operational use with respondents. Assessment of items whose calibration has drifted over time (that is, the values of their item parameters have changed relative to what they were prior to operational use) will need to be developed and managed with experts who understand the IRT model underlying the CAT. In some cases, items that show drift can be recalibrated based on accumulated respondent data. In other cases, items that exhibit drift may need to be eliminated from the CAT. Items that have

drifted need to be identified, however, or else your estimates of respondents' ability levels (theta scores) will be erroneous, leading to a number of potential problems. The good news is that assessment of existing test items can be incorporated into ongoing item development to produce better calibrated items that can survive under live conditions in the future.

Respondent and Client Perceptions

As described above, the benefits of CAT include producing tests that maximize the delivery of items at a respondent's true theta, the goal being to obtain a shorter, more precise, and unique test across respondents. In practice, however, this benefit can create issues for the successful use of a CAT with respondents. As mentioned earlier, one issue we have encountered is that adaptive tests often seem more difficult than their P&P counterparts because they are adapting to the respondent's theta level. In a pre-employment selection system when the goal is to maintain the interest of the best candidates, it is important that the adaptive test is communicated in a way that effectively manages this perception.

In particular, the delivery of different items for each respondent can create perceptions that the testing process is unfair. Because of the complicated algorithms underlying adaptive delivery, it can be difficult to explain how CATs actually yield more accurate results for respondents than P&P or other computerized tests tend to because the adaptive delivery maximizes the respondent's opportunity to demonstrate his or her true theta level. In addition to managing perceptions among respondents for CAT use, the perceptions of stakeholders that influence or control decisions for use of a CAT have to be managed initially and over time. As with respondents, explaining to stakeholders that the CAT will produce a unique test often raises concerns of disparate treatment and degraded fairness of the assessment process.

Legal Implications

When used to select employees, a CAT falls under the federal *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978) and the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003), among

other best practice guidelines for the use of tests, such as the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). As far as we are aware, no mention of CAT is made in any of these guidelines.⁴ One legal challenge that we could envision arising would be a complaint based on concerns over individuals being evaluated one with another despite having taken different items and thus different tests. Nevertheless, we are unaware of any United States legal or regulatory precedent (Equal Employment Opportunity Commission, Office of Federal Contract Compliance) for the use of CAT, making it unproven in this arena. Through personal experience, however, we can report the acceptance of P&G's cognitive ability CAT for use in European countries with Work Councils.

Given the relative lack of legal and regulatory precedent for the use of CAT, our legal partners have recommended that we place greater effort on documenting and explaining the design and use of the CAT as compared to that typically offered for a traditional test. In particular, the functionality of a CAT needs to be explained, along with its equivalence to more traditional testing procedures—a point well demonstrated in the literature (for example, Moreno & Segall, 1997; Moreno, Wetzel, McBride, & Weiss, 1984). Demonstrating the equivalence of a CAT with more traditional procedures is also important for enabling use of an alternate testing procedure in the event that a respondent requires an accommodation to computerized testing.

Considerations for the Development of a CAT

The preceding list of advantages and challenges is not exhaustive, but it does present many of the important considerations when deciding whether to use a CAT. Ultimately, the decision to use CAT will come down to balancing the pros and cons against the business or practical needs required for testing in your situation. Should the balance tip in favor of the advantages and the decision to move forward with using a CAT, there are additional issues that need to be considered regarding its development. These considerations also apply for the evaluation of previously developed CATs that the reader may be considering adopting or licensing for use in their practice. The good news is that the

development of a CAT resembles standard test development in many ways. In this section, we provide an applied overview of these considerations. Readers interested in the specifics of CAT development are referred to more detailed treatments of the subject (for example, Sands, Waters, & McBride, 1997).

End-User Requirements

Perhaps the primary differentiating factor between developing a CAT and a standard, non-adaptive test is the need to specify the desired end state and requisite end-user requirements before writing a single test item. Put more simply, the question of whether or not the respondent will be able to access and complete the test needs to be determined before development begins. Because a CAT requires use of a computer, considerations need to be made for the costs, processing requirements, screen capabilities, security/encryption, and placement of computers, among other considerations that could be unique to each testing case.

The use of CAT under unproctored conditions typically requires the respondent to provide his or her own computer, reducing the expense of purchasing or transporting computers to administer the tests. It also creates a loss of control over the type of computer and software being used, resulting in myriad additional issues related to web browsers, language packs, and computer locking software, among others, that have to be considered to ensure that the testing experience is equivalent among respondents.

Content Type

The use of CATs has predominantly been in the area of cognitive or mental ability tests, especially for the selection and placement of employees. Given this precedent, the theory and use of CAT with cognitive and mental ability content types is well founded, and guidance on content development and item pool requirements is readily available.

Building on this base, and given the technological improvements of the past decade, new developments in using

non-cognitive and other content types have proliferated in recent years. An example of this work for personality content is provided in Fetzer and Kantrowitz's case study in Chapter 15 of this volume. Depending on the content type chosen for the adaptive test, the development considerations outlined below may be different. Where appropriate, an attempt is made to call out where these differences by content type exist. The good news, however, is that no matter what content type is used, many of the development considerations are the same.

IRT Model

One early decision regards choosing the item response theory (IRT) model to be used in calibrating items against the underlying adaptive algorithms that determine item choice. For multiple-choice items, the models differ in terms of the number of parameters that need to be estimated. Although there are several models available (for example, Van der Linden & Hambleton, 1997), the popular choice for CATs has been the three-parameter logistic (3PL) model. The model gets its name from the (1) mathematical functional form of the curve (logistic) that relates examinee ability (the standing on the latent trait, θ) to the probability of answering the item correctly (the item response function, or IRF) and (2) number of parameters (three) that define the curve's shape. The three parameters in the 3PL model describe the item's difficulty (the b parameter), discrimination power (the a parameter), and estimated probability of guessing the item correctly (the c parameter).

One very important consideration when choosing a model is that, in addition to determining the size of item calibration samples, the adaptive model used will drive the complexity of the scoring algorithm and computing demands of the underlying delivery engine and platform. The 3PL model can be reduced to a 2PL model by dropping or fixing the c parameter and to a 1PL or Rasch model by fixing the a parameter to a single value (some fixed value in the 1PL, a value of 1.0 in the Rasch), resulting in reduced complexity but less precision for estimating the respondent's ability via the CAT. As an example, the CAT engine for a Rasch model, which involves only the difficulty parameter

(all items are seen as being of equal discriminating power, and there is no attempt to estimate a parameter representing chance answering), would have fewer “moving parts.” Therefore, calculation of candidates’ ability levels and item information would be less complex because fewer item parameters would be involved.

In addition to the 3PL, 2PL, 1PL, and Rasch IRT models, other models could be used to develop a CAT. In particular, partial credit, unidimensional pairwise comparison, polytomous, and multidimensional models could be implemented depending on the content type and scoring needs for the test. Nevertheless, the use of multiple-choice items, relative computational simplicity, and enhanced capacity to allow different items to discriminate to different degrees while also modeling chance responding tend to push most CAT developers to the 3PL model (for example, DoD’s CAT-ASVAB program, P&G’s CARAT & Reasoning Screen).

Data Requirements and Item Parameterization

This consideration area involves how one chooses to determine the parameters for the items in the CAT item pool. Typically, item parameters are calculated using statistical software designed for the purpose (for example, Parscale, Multilog). As described above, IRT models can require large samples of respondents for their development. This is especially true when estimating parameters for the 3PL model, where samples of one thousand (and preferably closer to two thousand) or more respondents for each item are recommended to obtain dependable and interpretable estimates of the a , b , and c parameters. Once obtained, however, these IRT parameter estimates are deemed not to depend on the sample from which they came—a key benefit of IRT-based estimation.

Other methods for rationally estimating the IRT parameters of each item constituting the CAT that reduce the amount of data required have also been employed, but they are less well understood and do not provide sample invariance. Fetzner and Kantrowitz (Chapter 15) describe new work that employs subject-matter experts to define parameters of non-cognitive test items.

It is important to emphasize that it is not sufficient for the parameterization (or calibration) sample just to be large—it

should also be representative of the intended candidate population. Regardless of the method chosen, the data requirements need to align with the requirements of the item parameter estimation procedure in a way that provides information on whether or not the test is appropriate for local use. For CATs used across many countries or with respondents from different backgrounds or cultures, the amount of data required will be more extensive. Obtaining a representative candidate sample in both empirical and rational estimation or item parameters ensures effective calibration across demographic segments and is vital to virtually any testing program and essential to global testing applications.

Current CAT systems allow important features with regard to item parameters, including the capacity to assess whether items might be working differently (that is, differential item functioning or DIF) for different important groups of interest (race/ethnicity, gender, age, geographic region, business sector). Such items should be replaced, and today's CAT platforms (with today's computing technology) support real-time DIF assessment. Current systems also allow assessment of the somewhat related topic of item parameter drift (IPD)—the notion that item parameters are no longer fixed at their initial calibration values. This can happen for several reasons, with item compromise/exposure being a major one. Failure to detect items that have experienced parameter drift can lead to incorrect ability estimates for examinees. Identifying IPD will help ensure the integrity of the ability estimates generated by the CAT engine.

The main message for this consideration area is that the effort and data requirements of a CAT are often much greater than those encountered in the development of non-adaptive tests. No matter what method *is* chosen, the data used in parameter estimation provide the statistical basis for use and defensibility of the CAT. Therefore, we strongly advise against cutting corners on data collection and parameter estimation procedures when developing a CAT.

Measurement Precision

A great advantage of developing tests on an IRT foundation is the capacity to structure your test to measure well in the areas

of the ability or trait (θ) for which you require the most precision. For example, a test having a pass/fail cutoff score requires great precision at and immediately around the cut score, with precision elsewhere in the ability range of much less concern. IRT greatly enhances development of this sort of “peaked” test, where discrimination is great at one point (or a few points, should multiple cutoffs be established), because one can empirically determine the amount of information provided by each item at each point along the range of θ . With these data in hand, items can be selected that provide the most information about candidates at the desired level(s) of θ (at the cutoff points).

Within IRT, “information” has a technical meaning, referring to the precision of measurement provided by an item or by a test. The amount of information provided by the item or test varies with the level of ability being assessed (the level of θ). The greater the information provided, the less the error of measurement involved. Item information can also be used to develop tests for use in more general selection settings. When a test is used for a broad range of applicants and top-down selection is desired, a reasonable amount of discrimination throughout the θ range is required as compared to a very high level of discrimination at only a few discrete points. Again, item information data can be used to evaluate the amount of measurement precision obtained throughout the range of θ .

In either case, item information data support the construction of a test with the desired levels of measurement precision in the desired areas of the θ continuum. When building conventional tests, this result allows the test developer to determine the measurement characteristics of any test that is a function of calibrated test items—even before the test has been administered as a unit. In CAT, the information function can be used to identify the appropriate item(s) to consider for administration immediately following an answer from a respondent. The measurement precision of the test can thus be carefully shaped by writing, calibrating, and including new items that provide information in the areas of ability or trait where insufficient levels of precision currently exist.

How to Calculate Information

With a CAT, the item each examinee sees after responding to the previous item(s) is chosen from the item pool because it provides a large amount of information about the examinee's current ability (θ) level. In fact, the quickest way to estimate the examinee's θ is to always administer the item from the pool that provides the most information about the examinee's θ estimate. This approach would get to a respondent's estimated ability most quickly because, if followed, it would mean that each item administered to that particular respondent would be the item that provided the smallest amount of error in estimating the respondent's ability level. This "maximum information" rule, however, is not typically followed (the reason will be discussed in a later paragraph). Nevertheless, the process of determining which item to administer next requires knowing how much information each item in the item pool provides at the respondent's currently estimated θ level. There are two primary ways that item information can be obtained.

One approach involves the establishment of information look-up tables. For this approach, the information provided at each of many levels of θ is calculated for each item in the item pool and then recorded in a look-up table. As the examinee's ability estimate is updated, the CAT item selection algorithm (that is, the "CAT engine") examines the look-up table and selects the item for subsequent administration based on the values of information in the table (and any other item selection data that should be considered). This is the approach used by the Department of Defense's (DoD) CAT-ASVAB program. The method was adopted originally to reduce the amount of computer processing power required to estimate information in real time during test administration.

A second approach to determining item information for the items in the CAT pool involves calculating information in real time. Such "on the fly" calculation has become much more feasible than in the early 1990s, when CAT-ASVAB first appeared, because of increased computational power. With this approach, the CAT engine simply calculates the amount of information provided by each item in the item pool at the examinee's current

estimated theta level. These data are then used (along with any other item selection data that should be considered) to identify the next item for administration to the respondent. The CAT that P&G uses for applicant selection purposes incorporates such a real-time item information calculation. One advantage of this approach is that you get actual item information values rather than approximations (albeit relatively close ones) that look-up tables typically provide.

Starting/Stopping Rules

The general approach for starting an adaptive test is to select an item of approximately average difficulty, with the assumption being that this item will be close to the estimated ability level of the highest proportion of examinees. This practice is linked to the development of CATs for assessing cognitive ability, which is normally distributed with most respondents' scores lying in the middle of the range of theta (ability) for the test. Some deviation from the mean difficulty level is permissible, of course, depending on your respondent base and test score decision criteria. For example, CATs could be designed to administer an item of slightly above-average difficulty if one believes the pool of examinees has been restricted or self-selected for the trait being assessed (often, cognitive ability). Others prefer to begin by administering an item of slightly below-average difficulty, with an eye partly toward providing the examinee with a reasonable probability of answering the first item correctly.

As described above, CATs have the capacity of providing a unique test to each examinee, depending on the size of the item pool, the IRT model selected, the responses of each examinee, and the number of items delivered to the respondent. Regarding stopping rules, adaptive testing presents two primary means for determining the end of the test.

First, one might decide that the examinee has seen enough test items when the person's estimated level of theta is reached with a predetermined level of precision. This is possible in IRT, where each individual has a specific standard error of measurement based on the items delivered and how they were answered by the respondent. With this approach, some examinees might

complete only twelve items, whereas others might need to complete twenty items or more.

Second, one might choose to deliver a fixed number of items to each examinee, similar to delivery of a more traditional “flat” test. The result is that one estimates the ability of examinees with potentially different levels of precision, but the fixed-length approach has the logistical advantage of being relatively predictable in terms of the amount of time that is likely to be required to complete the test—a very important consideration for real-world operational testing. In addition, this approach limits the number of items administered (and thus exposed) during any single test. Examinees whose ability would be estimated most precisely would be those with ability levels nearest the point of maximum measurement precision for the item pool and those who responded in a fashion consistent with the underlying IRT measurement model (that is, answering items correctly that they are expected to answer correctly and vice versa). With the “predetermined precision level” approach, a respondent could choose to respond inconsistently with regard to the measurement model (that is, purposely missing some items that the person had the ability level to answer correctly), thus increasing the length of the test and therefore the number of items exposed.

Perhaps not surprisingly, it appears to be more common to select a fixed test length than a desired level of measurement precision for stopping a CAT. For example, both the Department of Defense’s CAT-ASVAB program and P&G’s CAT employ a fixed test length. Ultimately, the choice of stopping rule should be made based on weighing the levels of precision, standardization, and candidate expectations required of the testing process.

Item Timing

For most CAT applications, a time limit for the test is established to help manage the overall testing experience for the respondent and administrator. The time limit may be enforced either by establishing a total test time (as is done for non-adaptive tests) or by setting the time requirement at the level of the individual item, such that each item must be completed within a set period

of time. For both types of item timing, the time remaining to respond is typically displayed on the computer screen, with warnings to the respondent when time is about to run out.

Depending on how timing is set for the CAT, it is important to note that the amount of time offered could influence the difficulty level of the items in the item pool. For example, giving thirty seconds for an examinee to answer the question could make the item more difficult than if one minute were provided for the response. Therefore, it is highly recommended that data collection for the calibration of a CAT be done with the operational time limits in place so that the estimation of item parameters yields accurate calculation of item information by the adaptive engine.

Movement Between Items

One important difference between CAT and traditional non-adaptive tests involves the ability (or lack thereof) to return to an item and change a response once provided. Because a CAT selects each subsequent item on the basis of the response to the previous item, respondents are typically not permitted to return to an item and change their answers. Once given, an answer to a CAT item is the answer of record. This also has the obvious implication that there is no going back on a test to check over one's work if a respondent completes a CAT prior to reaching the time limit.

Although one could program a CAT to permit examinees to go back to previously answered items, this practice is not common because it increases the exposure of test content. Unless additional controls are placed on the ability to return to a previous point in the test, a respondent could theoretically access all of the items in the pool, creating concern for the security of the CAT's items. It is for this reason that most CATs do not allow candidates to return to previous items.

Quality Assurance

Running quality assurance checks on the resulting CAT engine and testing system is one of the most important steps in the

entire development process, as it provides assurance that the adaptive algorithm and item pool are working together to produce expected results. This process can be a bit involved and more difficult than first anticipated because CAT item presentation is based on each candidate's item responses. Nevertheless, various "stock" response scenarios can be generated and evaluated via computer to determine important characteristics of the CAT, such as the following:

- Is the CAT engine calculating information and candidate ability or trait levels correctly?
- What happens when an examinee answers all items correctly (or incorrectly)?
- What levels of measurement precision are reflected by the current item pool?
- How quickly does the estimation algorithm converge on the candidate's ability estimate when the candidate responds (fails to respond) in accordance with the underlying IRT measurement model?
- How should the item selection algorithm, including any exposure controls that have been put in place, work? For example, how many items should be considered for presentation from the item pool: All of them? Just a subset? If just a subset, how many should be considered at once?

Failure to conduct adequate QA could lead to the complete failure of the CAT program. Without QA checks, one problem that could go undetected is the potential for "wasted" items (those in the item pool that, because of their characteristics and those of the other items in the pool, are never administered). Another difficulty that could go unnoticed is insufficient item exposure control, which could occur because too few items have sufficient discriminating power in one or more ranges of the ability distribution, leading to the need to administer certain discriminating items over and over again. Therefore, it is highly recommended that the CAT programmer, IRT/calibration expert, and test developer (assuming these are not the same person!) work together to develop a QA strategy that answers questions similar to those provided above.

Validation

As with the development of any test, CAT developers will need to establish the validity of their measure. Fortunately, one can validate a CAT the same as one can validate any non-adaptive test. The important point here is to be sure to plan for the validation study and ensure the appropriate participants and study design. For global assessment, representative, global samples across business segments can be critical for ensuring not only appropriate scientific results, but also local buy-in for test use. In addition, delivering the validation study under the same conditions that respondents will experience under live use of the CAT (for example, unproctored, timed) is recommended to get an accurate read on the validity of the test.

Administration Type—Supervised vs. Unproctored Delivery

Although it is difficult to imagine writing the words in this paragraph twenty years ago, one major decision that now lies before any organization wishing to develop an online selection test is whether to administer that measure in a supervised (“proctored”) environment. Despite its somewhat controversial nature (see Tippins, 2009; Tippins, Beaty, Drasgow, Gibson, Pearlman, Segall, & Shepherd, 2006), more organizations have begun to choose unproctored Internet testing (UIT) as the means for delivering their selection or screening tests.

A primary reason for employing UIT is its promise of greatly reduced costs typically affiliated with testing programs. Some of the costs UIT obviates include travel costs for job candidates, proctoring costs (for example, hiring, training, travel), and hardware costs (for example, purchase, distribution, maintenance). Other advantages include increased convenience for test respondents, including the speed with which they may enter the application process (for example, they need not wait for appointments to take the test) and the capacity to take the test at a time and place of their own choosing. This latter advantage leads directly to another advantage for the organization employing UIT: an expanded applicant base, arising from the capacity to obtain test

scores from many individuals who might otherwise not have been able to complete the assessment because of logistical constraints (for example, other commitments at the appointed testing times, inability to travel to the testing location). In addition, some argue that UIT programs enhance an organization's image, demonstrating an interest in providing convenient testing options to job candidates and to embrace cutting-edge technology. Finally, test proctors do not provide foolproof test security, sometimes being unskilled, untrained, and/or unmotivated.

Of course, UIT brings several challenges along with it. Perhaps most salient of these challenges is the potential for applicant cheating. With no one observing the testing process, candidates could well be assisted by friends or other resources, thereby inflating estimates of their true ability. Although warnings not to gain assistance from others in completing the test or to fake responses have been shown to help reduce cheating in the classroom (Kerkvliet & Sigmund, 1999) and on personality tests (Dwight & Donovan, 2003; McFarland, 2003), it is highly recommended that a verification procedure be employed along with any UIT. In particular, the guidelines published by the International Test Commission (2006) urge organizations to verify the ability level of candidates who pass tests offered as a UIT (see Guideline 45, in particular). Such verification testing can take various forms, depending on the policies and general comfort level of the organization in question. Some organizations might prefer a shorter UIT followed by a longer, perhaps augmented, verification test administered under supervised conditions on-site (the model P&G currently uses with its CAT), whereas other organizations might prefer to verify the UIT score with a shorter on-site test.

Bartram (2009) identifies other types of score verification procedures being developed, including data forensic analysis, randomized test construction (RTC) (CAT helps with this by providing applicant-specific test content), and remote authentication, all of which are post-hoc verification procedures. For example, Segall (2001, 2002) has developed a statistical procedure for comparing item responses from unproctored and proctored settings. And finally, it should be noted that cheating can

and does occur in supervised settings, too. Indeed, Bartram has opined that:

“It is paradoxical that the concerns raised over risks of cheating in UIT, despite the fact that cheating is also an issue in proctored assessment, have resulted in the development of technologies, policies and procedures that potentially make online UIT more secure than traditional paper-and-pencil proctored tests.” (p. 13)

Critics of UIT have raised other concerns, as well, including the potential to create a negative image of the organization (for example, “this company does not mind if it employs a few charlatans” or “this company lets a computer interact with me instead of taking the time to engage me personally”) and the contention that UIT constitutes unethical test practice (Pearlman, 2009). Nevertheless, the “cow is out of the barn” with regard to UIT. The question before us seems not whether to consider UIT as a viable testing alternative, but rather how to develop and operate a UIT program in the best possible way.

Ongoing Evaluation and Management of CATs

As with any test, and independent of whether administered under supervised or unproctored conditions, the CAT’s items become exposed each time it is delivered to examinees. In unproctored delivery, the situation is more severe, as it is much easier to capture and disseminate exposed test questions to others. Ultimately, releasing items “into the wild” means that item pools probably need to be refreshed continuously, as compromised items will quickly lose their discriminating power. It is in this respect that CAT provides a great advantage over non-adaptive tests, because adaptive tests require fewer items to hone in on a candidate’s ability or trait level. Fewer items administered means less item exposure and thus reduced item compromise relative to conventional tests.

In addition, item exposure can be monitored and used to determine which item should be presented next to the examinee. This is why the rule of “always administer the item that provides

the maximum amount of information” tends not to be followed in the delivery of CATs in real-world conditions. Instead, item information is balanced with item exposure, so that the same two or three items are not administered to every candidate with a particular estimated ability level. Consider the first item given to each candidate. Because the initial ability estimate is fixed to a particular starting value for every respondent (say, the average level of ability), strict adherence to the rule of maximum item information would mean that every respondent would begin the test by answering the same single item—the item that provides maximal information at the initial estimate of ability. Instead, operational CAT programs tend to identify several items that provide relatively high levels of information at the initial ability estimate and then select an item from among them, thus introducing some variability in item delivery.

The decision of how much weight to give to information as opposed to item exposure is more policy than science, but we note that both the CAT-ASVAB and P&G CAT programs implement item exposure control considerations in their item presentation algorithms. Also, the research literature contains several examples of development of operational exposure control algorithms for CAT (Chang & Ansley, 2003; Georgiadou, Triantafillou, & Economides, 2007; Hetter & Sympson, 1997; Pastor, Dodd, & Chang, 2002; Segall, 2003).

In addition to evaluating exposure through item delivery, CATs also afford the ability to assess whether or not the parameters for an item have changed or “drifted” over time. Item parameter drift algorithms have been developed for several CATs and help test administrators and developers understand whether each item in the test is operating the same way six months or a year after it was initially deployed for use with respondents. The basic premise of such algorithms is similar to that for differential item functioning, where the parameters deployed at launch are compared with estimates calibrated based on the accumulation of examinees’ responses at some point post launch (for example, six months, one year) to determine if the two sets of parameters are equivalent. If they are not, then the newly calibrated parameters can replace the old set, or the items that have drifted can be delivered as research items in the test.

The capacity to administer new items as research-only items and calibrate them to the same scaling as that possessed by the operational test items is a unique feature afforded in several CATs. By incorporating one or more unscored research items along with live items, it is much easier to facilitate research and expansion of the item pool with a CAT than with more traditional procedures (for example, development of alternate test forms).

Applications of CAT in the Real World

The primary goal of this section is to provide you with examples of how the above considerations for use and development of CATs have been incorporated into live CATs used in public and private industry. We begin by discussing the comparatively long history of adaptive test use for the selection and placement of personnel in the military Services. We will then provide an overview of how one multinational organization implemented adaptive testing to select pre-employment candidates across the globe under unproctored conditions.

DoD: CAT-ASVAB

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple aptitude test battery used by all military services (Army, Navy, Air Force, Marine Corps, and Coast Guard) for selection and classification of military applicants into their enlisted accessions programs. The battery comprises ten tests: Paragraph Comprehension, Word Knowledge, Mathematics Knowledge, Arithmetic Reasoning, General Science, Electronics Information, Auto Information, Shop Information, Mechanical Comprehension, and Assembling Objects. The first four tests constitute DoD's enlistment test, the Armed Forces Qualification Test (AFQT).

ASVAB has comprised various tests over the years. For example, two speeded tests (Coding Speed, Numerical Operations) were dropped from CAT-ASVAB in 2002 and replaced with the non-verbal Assembling Objects test (these two tests are still administered in the standard, non-adaptive, paper-and-pencil format). Originally developed as a paper-and-pencil test, DoD began

developing a CAT version of the battery in the late 1970s. DoD's CAT-ASVAB testing program began officially in 1979, although its initial roots date to research conducted by the Marine Corps in 1977 (McBride, 1997). The initial version of CAT-ASVAB was first used operationally twenty years later (1997) and is now administered at all Military Entrance Processing Stations (MEPS) (Pommerich, Segall, & Moreno, 2009). CAT-ASVAB was "the first large-scale adaptive test to be administered in a high-stakes setting" (Segall & Moreno, 1999, p. 35). Today, approximately 200,000 applicants to military service complete the CAT-ASVAB each year.

CAT-ASVAB tests are of fixed length, comprising either eleven or sixteen items, with a total of 145 items across the individual tests. Respondents have 154 minutes to complete the CAT-ASVAB, which is the same time limit as that used with the paper-and-pencil ASVAB. On average, however, testing time runs at about 50 percent of that for the non-adaptive version.

In September of 2008, DoD developed a version of CAT-ASVAB that could be administered on the Internet (iCAT ASVAB). This new version of CAT-ASVAB will allow it to be administered in previously infeasible locations, especially computer-based learning centers belonging to the National Guard and the individual military services. Currently, iCAT ASVAB is administered in a proctored environment, but DoD is researching the possibility of fielding iCAT in various types of unproctored settings.

P&G's CAT: The Reasoning Screen

Working with HumRRO in the early 1990s, P&G was among the first private companies to develop a supervised computer-adaptive test, the CARAT (Computer-Adaptive Reasoning ASDF Test), for use in screening candidates as part of its hiring process in the United States. Building on learning from the development of CAT-ASVAB, the CARAT was developed to replace the paper-and-pencil Problem Solving Test (PST) in use at that time. The PST comprised three cognitive ability subtests: Numerical Reasoning, Paragraph Comprehension, and Data Interpretation. As described above, the CARAT was abandoned within a year of its initial launch because of the difficult logistics involved with

transporting and finding computers on American campuses to administer the test.

Starting in 2004, the CARAT program was revitalized as P&G assessed whether an adaptive test could be delivered online and unproctored as part of a completely new, global, online, and unproctored assessment system for entry-level managerial and office administrator candidates across all P&G functions (Gibby, Ispas, McCloy, & Biga, 2009). Although P&G has delivered its online non-cognitive assessment under unproctored, online conditions since 1999, several key changes had to occur between 1993 and 2004 to warrant development of a new cognitive ability CAT. The first requisite change was the improvement of technology and accessibility to computers across the globe for P&G's candidate pool. The second change was the need to more efficiently manage costs and resources for administering the P&G selection tests. Between 1999 and 2008, P&G saw its candidate volumes grow from 25,000 to more than 500,000 applications per year. A third change was increasing concern over impending talent shortages in certain parts of the world that are predicted to occur within the next decade. P&G considers its ability to scale its testing programs and equate its testing practices and results across the globe to be a key competitive advantage, because it allows the company to find the best talent anywhere on earth and independent of source through their testing programs.

In the face of these changes that were driving the design and needed outcomes of their testing process, P&G determined that an online and unproctored cognitive ability CAT, the Reasoning Screen, was needed to provide incremental prediction of on-the-job and training performance above that already provided by their online, unproctored non-cognitive assessment. Development of the Reasoning Screen began in 2004 based on design considerations for each the following categories.

Candidates

Based on P&G's increasingly international business and focus on developing markets, the Reasoning Screen needed to be delivered globally with a single set of item parameters and scoring for all candidates. Based on candidate reactions research with more than three thousand candidates across the globe, it was determined that they were willing to complete multiple, short, online

tests as long as they understood why they were being asked to complete the tests, were able to access the tests at a time of their choosing, and quickly knew whether or not they were moving forward in the hiring process.

Content

The original design of the Reasoning Screen incorporated the use of three types of cognitive ability content: Figural Reasoning, Logic-Based Reasoning, and Numerical Reasoning, with the latter being the only historical content type previously used by P&G (dating from the late 1950s). These three content domains were chosen based on their ability to incrementally predict cognitive ability according to Carroll's (1993) survey of factor-analytic studies on human cognitive abilities. However, it was quickly determined that only the Figural Reasoning content would be used in the Reasoning Screen.

Ease and Cost of Item Development and Translation

The decision to use only the Figural Reasoning content was partially based on the initial and ongoing costs to develop the hundreds (and eventually thousands) of items that would be needed to manage exposure of the Reasoning Screen under unproctored conditions. It was the translation costs for the two verbal item types (for example, Numerical Reasoning, Logic-Based Reasoning), however, that forced the decision to move forward with the use of only Figural Reasoning items in the CAT. More specifically, P&G determined that the translation costs of these two item sets alone would equal the up-front development and ongoing delivery costs of their entire testing program!

Item Parameterization and Timing

Despite not using the hundreds of items developed for both the Numerical and Logic-Based Reasoning in the adaptive Reasoning Screen, P&G still globally calibrated these items along with the hundreds of Figural Reasoning items that were developed for the cognitive ability testing program. In fact, more than 138,000 candidates completed the research forms used to calibrate the Reasoning Screen empirically using a 3PL IRT model. Prior to this parameterization, however, more than fifty thousand

candidates across the globe completed the research forms for all three item types under untimed conditions. Based on analysis of these data, item-level timing was determined and incorporated into the data collection efforts that provided the final IRT calibration.

Ultimately, the final IRT calibration of all three content types was used to assess cross-cultural and other demographic (for example, gender, U.S. race/ethnicity) differential item functioning (DIF). Any Figural Reasoning items displaying DIF were removed from the item pool for the Reasoning Screen. In addition, any Numerical or Logic-Based Reasoning items displaying DIF were removed from inclusion to our paper-and-pencil verification test, the Reasoning Test, which incorporates all three content types.

Validation

As described above, validation of the Reasoning Screen was similar to validation studies conducted for P&G's more traditional cognitive and non-cognitive tests. The only change in the validation process was to validate the test with incumbents under unproctored, online conditions—a change necessary to approximate the operational validity of the test with candidates who would receive the test under these conditions. Validity results of the test exceeded expectations, with an uncorrected concurrent validity coefficient of $r = .29$ with supervisor-rated performance for the Reasoning Screen, and were consistent across all groups (for example, gender, culture, race/ethnicity, age).

Long-Term Sustainability

In line with the information provided above, the ability to manage the development, research, and exposure of content in the Reasoning Screen more easily through adaptive delivery was a primary factor in choosing to develop and implement a CAT for P&G's online, unproctored cognitive test. As part of the test, P&G delivers research content to every candidate, which allows them to refresh the Reasoning Screen in real time and in a more cost-effective way than provided by paper-and-pencil testing.

Launched in July 2008, the Reasoning Screen delivers fifteen items to each candidate. Test instructions are available in more than twenty languages. To date, more than 200,000 candidates in

more than eighty countries have completed the Reasoning Screen. In addition, thousands of candidates passing the Reasoning Screen have moved forward in the hiring process to complete P&G's supervised, P&P cognitive ability test—the Reasoning Test. This final, supervised cognitive test serves both as an independent hurdle in the hiring process and a way to verify the result of the online, unproctored Reasoning Screen. These tests were developed and validated together, with the scoring on the two tests set with the expectation that 85 percent or more of candidates passing the Reasoning Screen should pass the Reasoning Test. After the first year of using the Reasoning Screen, the convergence between the Reasoning Screen and P&G's verification Reasoning Test has exceeded expectations, with a 90.4 percent convergence across the globe.

Summary

In some areas of psychometrics, technology has arguably outpaced our psychometric expertise. For example, computer technology permits the development and administration of novel item types (drag and drop, point and click, video-based scenarios, high-fidelity simulations) that have the potential to offer hundreds of potentially scorable events, thus challenging our notions of how best to assess reliability and validity of the measures that contain them. With CAT, however, we have an example where technology has finally caught up enough with our psychometric expertise that it is now feasible to avail ourselves of the benefits of adaptive testing first discussed by Lord (1980), Weiss (1982, 1983, 1985), and others. We have argued that CAT offers substantial advantages (greater precision, shorter testing time, less susceptibility to test strategies) but involves laying a firm, expansive foundation (large item pools, large development samples). The use of CAT in both the public and private sector speaks to its advantages.

As is evident from our work to develop adaptive tests for the public and private sectors, we are proponents of using CATs for selection of candidates in industry. For the organizations that we have worked with, the assessment system requirements have aligned well to the adoption of adaptive tests, with costs and

concerns for adaptive testing being outweighed by their benefits. Given that many assessment vendors are moving forward with the development of adaptive delivery engines and assessments, it appears that other organizations are also seeing the benefits of adaptive testing.

Over the next decade, the true test of adaptive testing in industry will lie with candidates' willingness to complete these tests and with the results of government audits and legal challenge for adaptive tests. In particular, the growth of UIT presents both great promise and formidable questions to our field. Nevertheless, we believe that CAT provides many test users with one of the best ways to assess the potential of their examinees.

Notes

1. Each IRT model is defined by a number of item parameters. For example, a three-parameter logistic (3PL) model comprises three item parameters: b (the item difficulty or "location" parameter), a (the item discrimination or "slope" parameter), and c (the lower asymptote or "pseudo-guessing" parameter). Item calibration is the process of estimating these parameters from data generated by the administration of the items to a large sample of respondents.
2. Backtracking is disallowed because it would dismantle the effective functioning of the adaptive algorithm.
3. Many IRT models are available (see Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). For multiple-choice items, the choice typically comes down to either a three-parameter logistic model (3PL) or a one-parameter logistic model (1PL) or Rasch model. Although there is a two-parameter logistic (2PL) model, it tends not to garner as much attention when multiple-choice items are used.
4. The *Standards* are undergoing revision as of this writing. Perhaps the new version will address CAT directly.

References

- Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

- Bartram, D. (2009). The International Test Commission guidelines on computer-based and internet-delivered testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2, 11–13.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Chang, S. H., & Ansley, T. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 40, 1, 71–103.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it* (pp. 90–126). Mahwah, NJ: Lawrence Erlbaum Associates.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich.
- De Ayala, R.J. (2009). *Theory and practice of item response theory*. New York: Guilford.
- Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, 43(166), 38295–38309.
- Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, 16(1), 1–23.
- Educational Testing Service. (1993). *The GRE computer adaptive testing program (CAT): Integrating convenience, assessment, and technology*. Princeton, NJ: Educational Testing Service.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5(8).
- Gibby, R. E., Ispas, D., McCloy, R. A., & Biga, A. (2009). Moving beyond the challenges to make unproctored internet testing a reality. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2, 64–68.
- Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 69–80). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.
- Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.),

- Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.
- International Test Commission. (2006). International guidelines on computer-based and Internet-delivered testing. *International Journal of Testing*, 6, 143–172.
- Kerkvliet, J. R., & Sigmund, C. (1999). Can we control cheating in the classroom? *Journal of Economic Education*, 30(4), 331–343.
- Lord, F. M., (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.
- McBride, J. R. (1997). Research antecedents of applied adaptive testing. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 47–57). Washington, DC: American Psychological Association.
- McFarland, L. A. (2003). Warning against faking on a personality test: Effects on applicant reactions and personality test scores. *International Journal of Selection and Assessment*, 11(4), 265–276.
- Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23, 187–194.
- Moreno, K. E., & Segall, D. O. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169–174). Washington, DC: American Psychological Association.
- Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement*, 8, 155–163.
- Pastor, D. A., Dodd, B. G., & Chang, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, 26, 147–163.
- Pearlman, K. (2009). Unproctored internet testing: Practical, legal, and ethical concerns. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2, 14–19.
- Pommerich, M., Segall, D. O., & Moreno, K. E. (2009). The nine lives of CAT-ASVAB: Innovations and revelations. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved June 26, 2010, from www.psych.umn.edu/psylabs/CATCentral/

- Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- Segall, D. O. (2001, April). Measuring test compromise in high-stakes computerized adaptive testing: A verification testing approach. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, Washington.
- Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics*, 27(2), 163–179.
- Segall, D. O. (2003, April). An adaptive exposure control algorithm for computerized adaptive testing using a sharing item response theory model. Paper presented at the Annual meeting of the National Council on Measurement in Education, Chicago, Illinois.
- Segall, D. O., & Moreno, K. E. (1999). Development of the CAT-ASVAB. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 35–65). Mahwah, NJ: Lawrence Erlbaum Associates.
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
- Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2, 2–10.
- Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, 59(1), 189–225.
- Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Norwell, MA: Kluwer.
- Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
- Verschoor, A. J., & Straetmans, G.J.J.M. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 137–149), New York: Springer.
- Wainer, H. (1990). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report No. 74–5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Models Program.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473–492.

- Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*, 774–789.
- Zickar, M. J., Overton, R. C., Taylor, L. R., & Harms, H. J. (1999). The development of a computerized adaptive selection system for computer programmers in a financial services company. In F. Drasgow & J. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 7–33). Mahwah, NJ: Lawrence Erlbaum Associates.

Chapter Six

APPLICANT REACTIONS TO TECHNOLOGY- BASED SELECTION

What We Know So Far

Talya N. Bauer, Donald M. Truxillo,
Kyle Mack, and Ana B. Costa

The use of high-tech selection procedures continues to increase as companies strive to innovate their recruiting strategies and streamline their application and selection processes, in an effort to minimize personnel time, reduce overall time-to-hire, and optimize the number and caliber of eligible candidates for any given position. By foregoing the traditional newspaper classifieds in exchange for the ease and accessibility of the Internet to recruit and select a wider range of potential candidates (Koong, Liu, & Williams, 2002; Peters, 2001; Piturro, 2000), organizations can capitalize on the 60 percent increase in Internet job searches that has taken place in the United States within the past decades (Horrigan & Rainie, 2002). The cost of attracting employees online is estimated to be as low as \$900 per hire, which differs significantly from the traditional estimates of \$8,000 to \$10,000 per hire via the traditional recruiting methods (Cober, Brown, Blumental, Doverspike, & Levy, 2000).

Additionally, by implementing the use of online applications, it has been reported that organizations can cut more than eleven days off the typical forty-three-day hiring cycle (Cappelli, 2001).

As Lievens and Harris (2003) note, "Internet recruitment has, in certain ways at least, significantly changed the way in which the entire staffing process is conducted and understood" (p. 132). In addition, it is clear from these statistics that the utilization of high-tech recruitment and selection procedures is becoming more prevalent. These metrics show the changing ways in which individuals search for, apply for, and acquire information about jobs in today's world, but they also suggest that organizations must join the digital bandwagon to adjust to these growing trends. Beyond simply modifying existing recruitment and application practices, organizations have increasingly implemented high-tech assessment procedures. For example, one study reported that online personality testing was popular in approximately 20 percent of companies surveyed, with one-fifth of the respondents planning to implement additional online testing in the future (Piotrowski & Armstrong, 2006). In order to decide whether or not to implement high-tech selection procedures, however, organizations must evaluate the associated costs and benefits of each procedure and make critical decisions about whether to use a proctored versus an unproctored test and whether the costs of developing complex multimedia tests (such as situational judgment tests; SJTs) justify the benefits. A recent article reported that two-thirds of employers who use standardized tests in their selection procedures are relying on some form of unproctored Internet testing, up from 31 percent in 2005 (Beaty, Dawson, Fallaw, & Kantrowitz, 2009), even though researchers continue to disagree on utility of these measures (Tippins, 2009).

An additional consideration is how job applicants perceive these high-tech selection procedures. Although high-tech procedures are becoming more and more commonplace and appear to be approaching the norm in the testing arena, research on how applicants perceive and react to these high-tech assessments lags behind practice. Research has often indicated that applicants can have unfavorable reactions to traditionally administered tests, such as cognitive ability and personality tests, particularly

when compared to other selection methods, such as interviews and work samples (Hausknecht, Day, & Thomas, 2004). However, it is still unclear if and how those reactions differ based on the testing medium.

Applicant reactions to these selection procedures can have effects on applicants' self-perceptions and on their intentions and behaviors directed toward the organization. In support of this, empirical evidence shows that reactions to traditional selection systems impact applicants' intentions to accept potential job offers (for example, Gilliland, 1993; Macan, Avedon, Paese, & Smith, 1994; Truxillo, Bauer, Campion, & Paronto, 2002) and their willingness to recommend an organization to other job applicants (for example, Gilliland, 1993; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). And although it has not been empirically tested, it has been suggested that such reactions may affect applicants' subsequent job performance as employees (for example, Gilliland, 1994; Hausknecht, Day, & Thomas, 2004; Hunthausen, 2000).

Because selection continues to become more technologically advanced and increasingly draws on automated and computerized approaches, the goal of this chapter is to summarize what we know regarding applicant reactions to high-tech selection to date and what issues need to be considered further in the future. The guidelines currently available for testing procedures include the *Standards for Educational and Psychological Testing* (AERA, APA, CME, 1999), the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003), and the *International Guidelines on Computer-Based and Internet Delivered Testing* (Lievens, 2006). While these documents are relevant for developing testing principles and protocols, these guidelines must continue to evolve to address issues related to applicant reactions arising from the growing array of high-tech procedures available. Similarly, research on applicant reactions to high-tech procedures is relatively scant. And as a result, many of the topics discussed in this chapter are extrapolated from our work on applicant reactions to selection in general, findings from parallel areas of research, or based on our own experiences.

In this chapter we will briefly discuss some of the more high-tech procedures that are available and more widely used

by employers including online selection and recruitment (for example, Sylva & Mol, 2009). We will also provide a brief discussion of Gilliland's (1993) model of applicant reactions and use it to organize our discussion of the effects of high-tech procedures on reactions. Additionally, for each of Gilliland's ten procedural justice rules, we will address challenges, benefits, and recommendations for practice. We will also discuss how recommendations for practice may be extrapolated from other applicant reactions research and discuss new ideas generated from applicant reactions research.

Available Technologies

A variety of selection procedures are in use (see Gatewood, Feild, & Barrick, 2007, for more on this). For this chapter, we will focus on five "high-tech" media (many of which are included in chapters in this book). These include web-based management simulations, video-based assessments, Internet voice response (IVR) and web-based screenings, telephone role plays, and computerized adaptive testing (CAT).

We define *web-based management simulations* as work samples of various levels of fidelity that are presented via the web rather than in person. *Video-based assessments* are work samples (or situational judgment tests; SJTs) that are presented in video format rather than by paper and pencil. Response options may or may not be presented as well as applicant open-ended response. *Interactive voice response* (IVR) and web-based screenings are used to assess candidate minimum qualifications or may be used to differentiate among qualified candidates such as is done with biodata rather than using traditional paper surveys and application forms. *Telephone role plays* are synchronous role-play exercises, but they remove the need to be co-located. They also may be considered less media-rich, as they do not provide additional cues that would exist in a face-to-face interaction. *Computerized adaptive testing* (CAT) is utilized instead of a traditional paper-and-pencil test with items of varying difficulty. CAT can adapt to the ability level of the test-taker, thus requiring fewer questions than a traditional test in determining the candidate's actual ability level.

Applicant reaction issues associated with using high-tech selection methods include factors identified in older, traditional applicant reactions models such as fairness (for example, Anderson, 2003; Gilliland, 1993). Further, high-tech selection introduces a new range of issues that may affect applicant reactions, such as unproctored Internet testing (for example, Bartram, 2000; Tippins, 2009) and applicants' privacy concerns regarding information that can be easily or mistakenly disseminated (for example, Bartram, 2000; Harris, Van Hove, & Lievens, 2003). The next sections focus on Gilliland's (1993) fairness model of applicant reactions to selection and how it applies to high-tech selection.

Applicant Reactions Research and Models

Applicant reactions became an important area of research in the early 1990s. In an early empirical study, Smither and colleagues (1993) focused primarily on procedural and distributive justice perceptions and linked perceptions to test characteristics such as predictive face validity. Arvey, Strickland, Drauden, and Martin (1990) also looked at applicant perceptions but focused primarily on the effect of perceptions on applicant motivation. In addition to these empirical studies, a number of models of applicant reactions were developed in the early 1990s (for example, Arvey & Sackett, 1993; Schuler, 1993). However, by far the dominant framework for investigating applicant reactions is Gilliland's (1993) model based on organizational justice theory.

Gilliland's Model of Applicant Reactions

There are several existing reviews of applicant reactions research and theory (for example, Anderson, Lievens, van Dam, & Ryan, 2004; Hausknecht, Day, & Thomas, 2004; Ryan & Ployhart, 2000; Truxillo & Bauer, in press) and several discussions of Gilliland's applicant reactions framework (for example, Bauer, Truxillo, Sanchez, Craig, Ferrara, & Campion, 2001; Ryan & Ployhart, 2000; Steiner & Gilliland, 2001), so a lengthy treatment of the topic is outside the scope of this chapter. Gilliland's model builds on previous organizational justice research and focuses on

procedural justice (Leventhal, 1980; Thibaut & Walker, 1975), which concerns the fairness of the procedures used to make decisions; distributive justice (Adams, 1965; Cohen, 1987), which concerns the fairness of the distribution of outcomes based on equity and social comparisons, and interactional justice (Bies & Moag, 1986), which concerns the quality of the interpersonal treatment received during the selection procedure.

At the heart of the model are procedural justice rules specific to applicant reactions adapted from work by Leventhal (1980). These ten applicant reactions rules are further subdivided into three components. *Formal characteristics of the selection system* is comprised of four justice rules: *job relatedness*, *opportunity to perform*, *reconsideration opportunity*, and *consistency*; *explanations offered during the selection process* is comprised of *feedback*, *selection information*, and *honesty*; and *interpersonal treatment* is comprised of *interpersonal effectiveness*, *two-way communication*, and *propriety of questions*. Note that, while distributive justice is hypothesized to have a major effect on applicant reactions and behavior (Gilliland, 1993), and this has been consistently born out in the research (Ryan & Ployhart, 2000), the fairness of the selection process itself can affect or color how applicants perceive the selection process.

In one examination of Gilliland's (1993) model, Bauer, Truxillo, Sanchez, Craig, Ferrara, and Campion (2001) developed the Selection Procedural Justice Scale (SPJS) over a series of studies. They performed exploratory and confirmatory factor analyses to test the factor structure of the ten rules organized into three higher order components as Gilliland had suggested. They found evidence of two higher order factors, rather than the three components suggested by Gilliland: a structural factor, which involves the specifics of the actual process itself, and a social factor, which involves communication with and treatment of job applicants. Table 6.1 provides a brief explanation of each of the procedural justice rules that will be used in our discussion as well as a summary of recommendations.

Gilliland (1993) notes that applicant reactions are important to study because of the effect that they may have on both individuals undergoing the selection procedures as well as on the organization, as a result of the outcomes from those reactions.

Table 6.1. Procedural Justice Rules, Descriptions, and Recommendations for High-Tech Selection

<i>Procedural Justice Rules</i>	<i>Rule Descriptions</i>	<i>Recommendations</i>
Job Relatedness	Extent to which a test either appears to measure the content of the job or appears to be a valid predictor of job performance.	<p>Make sure the investment in job-related assessment will work long term and not become dated quickly.</p> <p>Create SJTs with fidelity and realism to enhance job relatedness.</p>
Opportunity to Perform	Having adequate opportunity to demonstrate one's knowledge, skills, and abilities in the testing situation.	Consider how workers with less computer experience may react to high-tech selection, especially if the job does not require complicated technology use.
Consistency	Uniformity of content across test sittings, in scoring, and in the interpretation of scores. Assurance that decision-making procedures are consistent across people and over time.	We recommend that organizations work to ensure the highest level of consistency possible. This will avoid concerns regarding testing fairness on this dimension.

Feedback	Providing applicants with informative and timely feedback on aspects of the decision-making process.	Providing as much feedback as possible is a plus in terms of applicant reactions. While it may not be possible or legally advisable to give specific performance, at least feedback regarding receipt of the submission, whether the applicant is still a candidate, or other aspects of the process should be given to applicants to enhance perceptions of adequate feedback and respect.
Selection Information and Explanations	The provision of justification for a selection decision and/or procedure.	Research shows that providing information about the quality and purpose of selection tasks helps applicants feel more confident in the process and enhances reactions. Building these in is advisable.
Propriety of Questions	The appropriateness of the questions asked during recruitment or the selection procedure.	Be sure to carefully go over content to ensure that all information gathered is necessary and proper.
Interpersonal Effectiveness	The degree to which applicants feel they are treated with warmth and respect by the test administrator.	Having smiling avatars, pleasant automatically generated messages, and appealing graphics can help applicants feel treated well on this dimension. Communications, even if computer-generated, should treat applicants with respect.

(Continued)

Table 6.1. (Continued)

<i>Procedural Justice Rules</i>	<i>Rule Descriptions</i>	<i>Recommendations</i>
Two-Way Communication	The interpersonal interaction between applicant and test administrator that allows applicants the opportunity to give their views or have their views considered in the selection process.	Having contact information available as well as online support or chat functions can help applicants at least understand how to communicate if something goes wrong during their application.
Reconsideration Opportunity	The opportunity for applicants to review their scores, or to be eligible for reconsideration if they are not selected.	Making the process clear to applicants regarding reconsideration opportunities can help make them feel more positive about the process.
Honesty	The importance of honesty and truthfulness when communicating with applicants, in particular, when either candidness or deception would likely be particularly salient in the selection procedure.	Check to be sure that all information available to applicants is updated and accurate. Schedule regular “check-ins” to view websites as applicants see them to ensure the content is up-to-date.

Research has found that characteristics of a selection system can affect such things as applicants' perceptions of fairness (for example, Hausknecht, Day, & Thomas, 2004; Macan, Avedon, Paese, & Smith, 1994). But in addition, these fairness reactions affect important outcomes such as self-perceptions (for example, Bauer, Maertz, Dolen, & Campion, 1998), perceptions of the organization (for example, Hausknecht, Day, & Thomas, 2004; Macan, Avedon, Paese, & Smith, 1994), as well as behaviors like self-selection from the process (for example, Ryan, Sacco, McFarland, & Kriska, 2000). Applicant fairness reactions may also affect intentions to litigate (for example, Bauer, Truxillo, Sanchez, Craig, Ferrara, & Campion, 2001), a finding of central interest to employers.

As noted earlier, outcome fairness or outcome favorability (whether the applicant passes the test or gets the job) is one of the key determinants of applicant reactions (Ryan & Ployhart, 2000); a great deal of negative outcome favorability is inherent in selection systems, as the goal is to eliminate applicants who do not fit the job. However, employers may be able to improve applicant reactions, regardless of the outcome an applicant receives, by focusing on these characteristics of the process as described by Gilliland (1993). Thus, much of this chapter will focus on the impact of technology on these procedural justice rules and how they, in turn, may affect additional outcomes. We propose that reactions to different technologies used to deliver selection procedures can be best understood based on the effect of the technology, whether positive or negative, on each of the rules on this expanded list. The next section of the chapter will discuss the benefits and challenges of using high-tech assessment procedures and their potential impact on each of the procedural rules. We conclude each discussion with recommendations for practice designed to minimize negative reactions and maximize positive reactions. Because empirical research specifically designed to assess the effects of high-tech media on applicant reactions remains scant, with some notable exceptions discussed below, much of our discussion must remain speculative.

Job Relatedness

Job relatedness refers to the extent to which a test either appears to measure the content of the job or is perceived to be a valid predictor of job performance. It is based on the accuracy rule, which suggests that decisions should be made using accurate information (Gilliland, 1993). Job relatedness was originally considered by Gilliland to be the most important procedural rule in determining applicant perceptions of fairness, and is one of the most frequently studied determinants of applicant reactions in general (Ryan & Ployhart, 2000; Schleicher, Venkataramani, Morgeson, & Campion, 2006). As noted earlier, Bauer, Truxillo, Sanchez, Craig, Ferrara, and Campion (2001) proposed that, based on factor analytic evidence, this rule should be split into the domains of job relatedness–content, which captures how much the content of the test appears to be related to the job, and job relatedness–predictive, which captures how well the test appears to predict job performance. These two additional factors echo previous research (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993) suggesting that job relatedness is split into perceptions related to face validity and perceived predictive validity. While job relatedness appears primarily to be an issue of test content rather than test medium, technology may enhance or inhibit the psychological and physical fidelity of a test, making it more or less acceptable to applicants.

While assessment content most likely remains the prime determinant of job relatedness reactions, some high-tech delivery platforms may have secondary effects because they alter or enhance test content and, importantly, psychological and physical fidelity. For example, the use of video or animation in work samples and situational judgment tests affects both the format of the test as well as test content and can make it a closer simulation to the actual job tasks and the knowledge, skills, and abilities (KSAs) needed to perform the job. Of the five technology media highlighted above, we see reactions in the job relatedness domain being affected mostly by the use of video, computer animation, and avatars in simulations and situational judgment tests (SJT) and by the use of the telephone and other technologies when conducting role plays and interviews.

Potential Benefits

High-tech assessment strategies such as video-based SJTs, and to a lesser extent SJTs that employ computer animation, have the potential to present far more realistic situations to applicants, thus potentially enhancing both content and predictive job relatedness perceptions over written items. However, the few studies that have empirically examined applicant reactions to video-based versus written SJTs have come to somewhat conflicting conclusions. Chan and colleagues (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Chan & Schmitt, 1997) and Richman-Hirsch, Olson-Buchanan, and Drasgow (2000) found that participants showed more positive reactions and had higher perceptions of face validity for video-based SJTs in laboratory settings. In contrast, Lievens and Sackett (2007) found that perceptions of face validity did not differ across the two formats in an actual selection context. Thus, it is difficult to draw firm conclusions about this potential benefit from the empirical literature. In these cases, it is also difficult to separate the form of the SJT (written or video-based) from the content of the SJT. Perhaps these researchers came to different conclusions because the content of the SJT under study was more or less conducive to delivery via multi-media.

Potential Challenges

Hausknecht, Day, and Thomas (2004) provide meta-analytic evidence indicating that applicants like interviews and work samples tests above all other types of selection procedures, and they speculate that the favorability of these selection tools may be due to the close relationship between the content of the selection tool and the duties of the job. While it seems intuitive that work sample tests provide content that appears job related, it is less clear that the favorability of interviews is due to their job relatedness, and may instead reflect the social nature of the interview. If this is indeed the case, then conducting interviews over the telephone or using video conferencing may lessen the favorability of the interview in the eyes of applicants. Other research suggests that the medium

used to conduct the interview may affect the content of the interview, with possible effects on perceptions of job relatedness. For example, one concern regarding job relatedness is the finding that telephone interviewers tended to ask more closed-ended questions than face-to-face recruiters (Silvester & Anderson, 2003).

Recommendations for Practice

While available research does not definitively show that video- and avatar-based SJTs lead to higher job relatedness perceptions, these technologies do not appear to reduce them. However, the potential of these technologies to enhance applicant reactions comes at a cost. Video-based and multimedia SJTs are far more expensive to produce than their written counterparts. Furthermore, the content of video-based SJTs may become stale rather quickly as clothing styles and work environments change, necessitating frequent re-shoots of the material. Avatar-based SJTs, which use computer animation instead of video, may help to mitigate these costs, as the “clothing” on the avatars can be changed rather easily without affecting the animation or the content of the SJT. Decisions regarding whether or not to use SJT formats that provide more realism should thus be made by weighing the practical costs and benefits of the technologies outside the realm of applicant reactions.

Consideration should also be given to the effect of the technology on the content of the selection instrument, especially when the medium used to deliver the selection procedure is linked to tasks performed on the job. For example, selection instruments that are delivered online or via computer may appear more job related to applicants for jobs such as software programmer who expect to use computers in the jobs to which they are applying, simply because the medium seems job related. In contrast, applicants to positions that require a lot of human interaction, such as sales, account management, or customer service, may view face-to-face interviews as more job related.

Opportunity to Perform

In the context of selection, Gilliland (1993) defines opportunity to perform (OTP) as having adequate opportunity to demonstrate one's knowledge, skills and abilities in the testing situation. OTP is based on the voice rule (Thibaut & Walker, 1975), which suggests that people should have the opportunity to express themselves prior to a decision. Although OTP has been given relatively little attention in the applicant reactions literature, a recent study illustrated that OTP is one of the most important procedural rules in the applicant reactions context because it provides a cognitive pathway by which applicants can justify poor performance to themselves (Schleicher, Venkataramani, Morgeson, & Campion, 2006). In other words, applicants may believe that they did not receive a favorable outcome because they were not provided with enough opportunity to show their skills and abilities. According to Schleicher and colleagues, OTP is thus particularly salient after applicants receive a selection decision, because it provides a convenient target for self-serving bias. Of the five high-tech methods discussed above, we see reactions in the domain of OTP as being primarily influenced by the use of the Internet and other computer related technologies to deliver test content, because these factors may influence the testing experience and alter opportunity to perform for some applicants. Internet testing refers simply to test content that is disseminated through the Internet. While the term includes both Internet-based tests administered on location or otherwise in a proctored environment, much of the debate has centered on unproctored Internet testing (UIT), in which applicants respond to test content on their home computers or in a similarly unproctored environment.

Potential Benefits

While it is difficult to pinpoint the effect of UIT on OTP perceptions, some initial research indicates that applicants generally prefer UIT to traditional paper-and-pencil test administrations, and we can speculate that their preference may be due to enhanced

OTP perceptions. For example, Gibby, Ispas, McCloy, and Biga (2009) reported that applicants preferred UIT to traditional testing methods because it created greater flexibility. Similarly, Beaty and colleagues (2009) reported that 93 percent of applicants felt that the location they chose to take the assessment allowed them to perform their best, and 88 percent were satisfied with their overall testing experience.

Potential Challenges

Tippins, Beaty, Drasgow, Gibson, Pearlman, Segall, and Shepherd (2006) also note that the use of UIT includes notable disadvantages, including issues regarding test security, the identity of candidates, and the increased opportunities for some applicants to cheat. A higher prevalence of cheating and the inability to confirm the identity of candidates in UIT has the potential to affect test validity, although empirical research that directly addresses this issue is scant. Kaminski and Hemingway (2009) actually found evidence that the validity of an unproctored test did not differ significantly from a proctored one, although the content of the two assessments was not identical. However, there is some indication that these issues likely affect applicant reactions, especially in the domains of OTP and consistency of administration. We will address the effect of UIT on OTP in this section and reserve the discussion of consistency for a later section.

Two main forms of cheating can occur when UIT is used: the applicant may use answer keys or other materials to aid in test-taking, or the applicants may actually have other people take the test for them. While both of these behaviors may be problematic in any unproctored setting, the ease with which answer keys can be disseminated on the Internet and the difficulty in verifying identity make these issues particularly important when the Internet is used. In fact, it is fairly easy to find answer keys to widely used tests within a few searches on the Internet or on popular social networking sites such as Facebook, although the quality of such “answer keys” is debatable. Taken together, cheating on the part of some applicants is likely to affect honest applicants’ perceived opportunity to

perform, in that dishonest applicants are provided with greater opportunity to look their best. However, while it is clear that cheating might affect actual opportunity to perform, it is less clear that it will affect applicant *perceptions* of their opportunity to perform.

Other issues such as variability in the speed of Internet connections and the applicant's overall familiarity with computers and the Internet can also affect OTP for Internet-based tests. Internet connection speed is especially important on power tests, in which applicants have a set amount of time to complete the test or in which speed is a criterion (Tippins et al., 2006). Differences in familiarity with computers among applicants will also impact applicants' opportunity to perform because those who are familiar with computers and the Internet will have a clear advantage over those who are not. Finally, some applicants may not have Internet connections at all, which limits their ability to access the assessment, although this issue may be considered less in terms of "opportunity to perform" and more in terms of "opportunity to apply."

Finally, we believe that OTP is a key issue for CAT. Applicants may not feel that they have the same control in a CAT test, as opposed to a paper-and-pencil test where they can see all of the questions and work on different parts of the test as they wish. In this way, CATs may be viewed as providing less OTP to applicants, although there has been no research on this issue to date.

Recommendations for Practice

Based on the available research, it is difficult to make specific recommendations pertaining to applicant OTP when using high-tech selection procedures. In terms of UIT, we can cautiously suggest that applicants' flexibility to choose their own optimum environment may outweigh the challenges associated with cheating and test security, although more research is needed before making confident recommendations, and we only make this recommendation for non-cognitive tests. We would also like to see more research comparing the perceptions of OTP among older and younger workers and between those who are more familiar and less familiar with technology in general.

Consistency

In the context of selection systems, consistency refers to the uniformity of content across test sittings, in scoring, and in the interpretation of scores (Arvey & Sackett, 1993) and is based on a procedural justice rule assuring that decision-making procedures are consistent across people and over time (Gilliland, 1993). When considering technology, consistency may also be interpreted as the consistency of interpersonal treatment, of access (for example, Internet access and experience), of time given, and of testing environment.

Potential Benefits

Some high-tech assessment strategies have the potential to increase consistency in the applicant experience. For example, research suggests that applicant screening using IVR technology was rated lower than applicant screening involving face-to-face interactions on the dimensions of interpersonal treatment and two-way communication, but the two were not significantly different on other applicant reactions variables (Bauer, Truxillo, Paronto, Weekley, & Campion, 2004). Moreover, it is possible that IVR screening and other forms of automated interviewing may increase both actual and perceived consistency across applicants because each applicant receives the same set of questions delivered in the same fashion, although we are not aware of any research that has directly examined this issue. Similarly, Internet testing, whether proctored or unproctored, does have the potential to increase the consistency of administration because the test items can be presented and test-takers' responses can be captured in an identical fashion. This is especially the case when administered via a corporate intranet where the testing environment can be more tightly controlled (Lievens & Harris, 2003). Finally, applicants may see high-tech assessment as more consistent than paper-and-pencil tests administered by HR staff, which are thus more subject to human error.

Potential Challenges

Although there are several practical advantages to using high-tech assessment strategies such as UIT, the nature of the medium

virtually guarantees that the testing experience will be varied across individuals. Internet-based tests, taken at home or in other unproctored settings such as libraries, Internet cafes, or coffee shops will have much more variation in the testing environment than a paper-and-pencil test, which is almost universally given under strict conditions, although these differences may not be obvious to applicants. These elements may affect applicant performance on selection instruments, and they are also likely to affect fairness reactions in the domain of consistency. Elements of the medium such as Internet connection speed, computer speed, memory, software and hardware configuration, and monitor size and resolution can all affect the testing experience leading to inconsistent administration. However, consistency may also be affected by other elements of the test environment such as the temperature of the room, lighting, and the presence or absence of distractions. Additionally, in order to limit the potential for cheating, many tests that are administered over the Internet provide a subset of items drawn from a bank of possible items, which decreases the consistency of the item content.

Another potential challenge to consistency comes in the form of computer adaptive testing (CAT). Although CAT may improve the testing experience by substantially reducing the length of an assessment and the time needed to complete it, the reduction in length comes at a cost to consistency of administration because each applicant sees a different set of items. In order to reduce length, CAT-based tests start with an item of average difficulty and then adapt subsequent questions to the ability level of the applicant based on his or her performance on the previous question. The result is that top performers will receive questions that are consistently difficult, while low performers will see questions that are consistently easier, which amounts to an inconsistent, albeit psychometrically equivalent, test administration. Again, however, whether these differences in consistency would be obvious to the average applicant remains an empirical question.

Recommendations for Practice

There is an ongoing debate regarding when and how UIT should be used as a selection or screening strategy, if at all.

In spite of its potential drawbacks, UIT is appropriate and cost-effective for low-stakes testing situations, but may not be acceptable for selection situations except perhaps for non-cognitive tests. It is likely not appropriate when there are security risks to test content or for cognitive testing, and some sort of follow-up assessment in a proctored environment is necessary (Tippins, et al., 2006; Tippins, 2009). However, it is clear that UIT will have profound effects on the consistency of administration. What is less clear is whether applicants who are generally familiar with the Internet will perceive the inconsistencies as unfair. Research, albeit preliminary, indicates that most applicants felt satisfied with the testing experience and felt able to perform their best (Beaty, Dawson, Fallaw, & Kantrowitz, 2009). Based on the general research consensus, we can thus tentatively conclude that UIT may be preferred by applicants over traditional paper-and-pencil tests. However, UIT may result in inconsistent consideration of scores, such that the scores of honest and dishonest test-takers are viewed similarly. Thus, each organization must carefully weigh this potential benefit against the drawbacks and decide for themselves if and how to implement UIT.

Feedback

This procedural justice rule refers to providing applicants with informative and timely feedback on aspects of the decision making process (Gilliland, 1993). In the selection context, professional standards recommend that feedback of some sort be given to applicants (SIOP, 2003). Feedback in a selection context may pertain to applicants' performance on specific selection measures or the selection decision itself. Empirical research has demonstrated that feedback is particularly important as a practical consideration because it represents a relatively inexpensive and straightforward method of improving overall reactions to the selection system (Lievens, DeCorte, & Brysse, 2003). Strangely, in spite of the obvious logical beneficial effects of feedback on reactions, applicants are often given almost no feedback regarding their performance in many situations, frequently not even an indication that they were not selected. It is

also our observation that this practice may have become even more widespread with the advent of high-tech assessments, perhaps because of the ease with which tests are administered to large number of applicants.

Potential Benefits

Providing timely and accurate feedback provides important benefits to both the organization and the candidate. Because feedback can be processed quickly using online databases and delivered automatically using email and other web-based technologies, high-tech assessments have the potential of notifying applicants of their selection outcome almost immediately. For individuals who are selected, this helps reduce time to hire and lessens the chance that a desirable applicant will take a job elsewhere while waiting for selection feedback. In that same spirit, providing timely and informative feedback to applicants who are rejected also allows them the opportunity to look for work elsewhere rather than waiting in limbo for an unfavorable response from an organization. Providing feedback is also highly related to the concept of social validity delineated by Schuler (1993) and the idea of treating applicants with dignity and respect, important to interpersonal dimensions of selection fairness (for example, Gilliland, 1993). Waung and Brice (2007) noted that more negative perceptions of the organization may result when applicants do not receive feedback. Anseel and Lievens (2009) expanded on this idea, finding that feedback regarding performance on individual tests was associated with test attitudes and later test performance.

Potential Challenges

Because providing timely and informative feedback is a low cost and easily implemented intervention to improve applicant reactions, we see relatively few potential challenges or drawbacks to this aspect of high-tech selection systems. The primary hindrance may be the relatively large number of applicants who are screened in high-tech selection systems.

Recommendations for Practice

Unlike so many situations in which participants go through the necessary steps to apply for a position and even interview with a particular company only to never hear back (good or bad) on their status or overall outcome, high-tech procedures lend themselves to providing feedback almost instantaneously. For example, due to the enhanced availability of applicant data when using high-tech procedures, feedback can be processed and disseminated quickly using automated processes such that applicants who take online computer adaptive tests for screening purposes could be notified of their performance and their selection status immediately after taking the test. Rather than waiting for HR personnel to tally up results, analyze data, and contact individuals, applicants can be notified using form letters and emails as soon as the selection decision is made. Moreover, depending on the testing format and specific selection procedure, the feedback can be objective and quantitative and may include justifications for the selection decision and validity information, which have also been shown to increase overall applicant reactions. Emails thanking applicants for their time and notifying them of the selection decision them can also automatically be generated and sent.

We see providing timely and informative feedback as a relatively easy and inexpensive method for increasing applicant perceptions of fairness and decreasing time to hire, and it is one of the chief advantages of using automated high-tech selection procedures that utilize the Internet and other computer technology. We recommend that organizations that are already using web- or computer-based testing implement some form of automated feedback, and that those who are not either make the switch or implement back-end systems to automate this important task. The challenge for employers is to provide sufficient feedback for applicants so that they feel respected and fairly treated, while at the same time not providing overly detailed feedback that may compromise test security or provide information that could in some way bolster legal action. In short, high-tech selection procedures make timely feedback to applicants feasible, which it may not have been in less automated contexts.

Selection Information/Explanations

The selection information procedural justice rule refers to the provision of justification or explanation for a selection decision and/or procedure (Gilliland, 1993). Gilliland further stated that perceptions of fairness are likely to be influenced by information on the validity of the selection process, information related to the scoring of the test, and the way in which scores are used in decision making. More often than not, however, applicants are not even provided with feedback regarding the selection decision when they are not hired, let alone specific information about the test itself. Nevertheless, explanations appear to increase applicants' perceptions of selection procedure fairness, for both selected and rejected applicants (Gilliland, Groth, Baker, Dew, Polly, & Langdon, 2001; Ployhart, Ryan, & Bennett, 1999; Truxillo, Bauer, Campion, & Paronto, 2002; Truxillo, Bodner, Bertolino, Bauer, & Yonce, 2009). The 2009 Truxillo study also noted in their meta-analysis that providing feedback to applicants has the added benefit of increasing applicants' perceptions of organizational attractiveness, and that feedback explanations were more likely to affect applicants' perceptions of fairness to a personality test than to a cognitive ability test. Meta-analytic evidence on explanations in a range of organizational contexts besides selection (Shaw, Wild, & Colquitt, 2003) has found that excuse explanations (shifting blame from the organization) were more effective than simple justifications for using a process, although this difference was not confirmed in the 2009 meta-analysis on providing explanations to applicants.

Potential Benefits

One of the critical benefits of high-tech procedures is that applicants can easily access information about the selection procedure, testing format, scoring process, and how hiring decisions will be made prior to applying for the position. For example, even as far back as in 2001, Cappelli found that ninety-five out of one hundred companies sampled had career websites and that 86 percent of corporate websites featured a link to the career site on the home page of the main site. The

career sites for the remainder of the companies could be found two to three clicks away. Along those lines, Truxillo and colleagues (2002) suggested that a practical and inexpensive way to improve applicant reactions was simply by providing applicants with adequate information prior to or during the selection process. Thus, rather than relying on a trained proctor or employee to relay relevant information to applicants via the phone, mailings, or in person, all explanations and information can be pooled together and consistently communicated the same way every time. Additionally, explanations have been found to have a positive impact on test performance and test-taking motivation on cognitive ability tests (Truxillo et al., 2009), suggesting that explanations may have the added benefit to employers of improving the utility of assessments, regardless of testing medium.

Potential Challenges

Organizations may be reluctant to provide explanations about their procedures to applicants; some employers see explanations as somehow providing applicants with the “ammunition” they need to sue. However, for validated selection procedures, providing such information has few downsides. It may in fact prevent negative reactions among applicants and thus may also prevent negative actions by applicants.

Recommendations for Practice

Not surprisingly, research has shown that providing explanations affects important applicant reactions and behaviors, such as applicant fairness perceptions, test-taking performance and motivation, and organizational attractiveness (for example, Truxillo et al., 2009), regardless of the type of explanation (excuse or justification). Thus, we recommend that employers provide explanations to applicants about the selection process, especially applicants receiving negative outcomes. We further recommend that these explanations be made treating applicants in ways that they feel respects their time, effort, and experience with the selection process. More research is needed to determine

the precise type or types of explanations that are most effective and when it is best to provide explanations.

Propriety of Questions

The propriety of questions procedural justice rule refers to the appropriateness of the questions asked during recruitment or the selection procedure. Perceptions of question propriety may be influenced by inappropriate questions and prejudicial statements, and they may also be decreased when applicants feel that their privacy has been invaded (Gilliland, 1993). In that sense, the propriety of questions issue is no greater among online applicants than among those who apply in person. However, we also believe that this rule is related to more recent concerns about the security and privacy of applicant data. Thus, we link the question propriety rule to more holistic privacy concerns in the context of Internet testing and online screening (Bauer, Truxillo, Tucker, Weathers, Bertolino, & Erdogan, 2006), which we see as important to the current discussion.

Potential Benefits

Although we see concerns over privacy as a potentially important drawback to using Internet testing and screening procedures, there are some potential benefits to employing high-tech procedures where propriety of questions is concerned. First, high-tech test administration helps ensure that standardized and appropriate test content is consistently administered to all applicants and facilitates the removal of offending items from the instrument if concerns with the item are found. Along those same lines, by using structured interview techniques, employers minimize the possibility that interviewers and raters will ask inappropriate or illegal questions.

Potential Challenges

One of the key challenges to high-tech solutions is the overall perception that personal data may not be safe because they are collected and stored on the Internet, especially by individuals

who are not comfortable or familiar with computers and automated systems. Along these lines, Bauer, Truxillo, Tucker, Weathers, Bertolino, and Erdogan (2006) found that information privacy concerns were significantly and negatively related to fairness perceptions, organizational attractiveness, and other important variables in the context of an online screening. It thus appears that question propriety, and by extension the overall propriety of the testing experience, is very important for online testing, especially for those applicants with higher pre-existing concerns about privacy.

Recommendations for Practice

Because privacy concerns are particularly salient when the Internet is used as a testing or screening mechanism, organizations which either currently use or are planning on using the Internet as part of the selection process should pay particularly close attention to privacy and security matters and should look for ways to enhance candidate perceptions of security.

Non-Technology-Specific Procedural Justice Rules

In addition to these six rules, Gilliland (1993) proposes four additional procedural justice rules that may be important to applicant reactions: interpersonal treatment, two-way communication, reconsideration opportunity, and honesty. We combine our discussion of interpersonal treatment and two-way communication, which both focus on interpersonal interaction and are less relevant to high-tech selection. We then briefly discuss reconsideration opportunity and honesty, which we consider to be more functions of HR and corporate policies than of the test medium *per se*.

Interpersonal Treatment and Two-Way Communication

Gilliland (1993) originally described treatment at the test site as the interpersonal effectiveness of the test administrator, referring

to the degree to which applicants feel they are treated with warmth and respect by the test administrator. However, the rule can be expanded to include the overall treatment of the candidate at the test site or during the testing process (Bauer, Truxillo, Paronto, Weekley, & Campion, 2004). Research has shown that the strongest predictor of impressions of a company and expectations regarding job offers and acceptance of those job offers is the warmth and thoughtfulness of an interviewer (Liden & Parsons, 1986; Coyle, White, & Rauschenberger, 1978).

The two-way communication procedural justice rule refers to the interpersonal interaction between applicant and test administrator that allows applicants the opportunity to give their views or have their views considered in the selection process (Gilliland, 1993). Previous research has shown that applicants for a high-status job expressed more anger and resentment toward computerized and paper-and-pencil interviewing than toward traditional face-to-face interviewing (Martin & Nagao, 1998), although this may be a function of the level of the job and the fact that high-tech assessment was less common at that time.

Depending on the testing format or selection procedure (for example, cognitive ability testing), human interaction may not be necessary, and may even detract from the overall goal. As such, it may be advantageous for some tests to be administered using automated procedures. However, any selection system that reduces or eliminates human interaction may affect applicant perceptions of interpersonal effectiveness, and we thus see high tech procedures as generally detrimental to interpersonal treatment perceptions. For instance, by placing an online assessment, IVR screening, or computer assisted interview at the beginning of the process, organizations may appear to be distant, impersonal, and lacking of that 'human element' to applicants. Not surprisingly, research has shown that applicants rate IVR screening techniques lower in terms of interpersonal treatment than face-to-face interviewing (Bauer, Truxillo, Paronto, Weekley, & Campion, 2004). It is possible, however, that receiving feedback via an email, text message, computer printout, or automated message rather than from a

real person may actually be less threatening to some individuals. Additionally, for individuals in the high-tech industries, such as computer science, information technologies, and management information systems, high-tech administrations and feedback may actually be advantageous to an organization—giving it the appearance of being technologically savvy and ahead of the curve.

Reconsideration Opportunity

In the context of selection systems, reconsideration opportunity refers to the opportunity for applicants to review their scores or to put themselves up for reconsideration if they are not selected the first time (Gilliland, 1993). Elements of the selection process that affect this rule appear to be more of an issue of organizational policy rather than the presence or absence of technology in the selection system, and are thus outside the scope of this chapter. It should be noted, however, that high-tech procedures may help facilitate reconsideration opportunity for those organizations that want to provide it. For example, the speed with which computerized assessments can be scored and the ability to store applicant information in a database would help facilitate reconsideration of applicants who were previously rejected if the hiring needs of the organizations change.

Honesty

The honesty procedural justice rule refers to the importance of honesty and truthfulness when communicating with applicants, and in particular, in instances when either candor or deception would likely be particularly salient in the selection procedure (Gilliland, 1993). It is said to be an important component of applicant reactions, distinct from honesty inherent in other forms of explanations and information provided to applicants (that is, selection information, feedback). As such, we see this rule as more related to organizational policies and not inherent in any particular assessment medium.

Future Directions for Research and Practice

Throughout this paper, we've described the ways in which a key applicant reactions model (Gilliland, 1993) relates to high-tech selection issues. Where possible, we've included empirical applicant reactions research that may be most relevant and provided recommendations for practice. What our review may best illustrate is that there has been relatively little empirical work on applicant reactions to high-tech assessment and that recommendations for practice must often be gleaned from work in other contexts. Given the profound changes in recent years with regard to the way that selection procedures work, the time is ripe for additional applicant reactions research on high-tech selection that can specifically guide practice.

We encourage practitioners and applicant reactions researchers to keep up-to-date on research on web design and website attractiveness (for example, Cober, Brown, Keeping, & Levy, 2004; Dineen, Ling, Ash, & Del Vecchio, 2007). In short, to better guide practice applicant reactions research needs to catch up to the changes that technology changes that have occurred in recruitment and selection.

Another fruitful avenue of research is to further explore the interaction between attributions and high technology selection. For example, the work of Ployhart and colleagues (Ployhart, Ehrhart, & Hayes, 2005; Ployhart & Harold, 2004) may yield important insights into applicant reactions in the high-tech arena regarding attributions that applicants make. And finally, research into the implications for test-taking motivation may also lead to helpful information about applicant reactions and high-tech selection in terms of effort and outcomes (for example, Arvey, Strickland, Drauden, & Martin, 1990; Sanchez, Truxillo, & Bauer, 2000; Truxillo, Bauer, & Sanchez, 2001).

Conclusion

In conclusion, when designing and implementing high-tech selection procedures, the applicant reaction implications should be taken into account. The automation made possible by advances in technology have the potential to maximize applicant reactions

through “high-tech, high-touch” procedures or to alienate users. Taking time up-front to analyze systems from this applicant-focused perspective can help researchers and practitioners make the most of the technologies available to them.

References

- Adams, J. S. (1965). Inequity in social exchange. *Advances in Experimental Social Psychology*, *1*, 267–299.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Anderson, N. (2003). Applicant and recruiter reactions to new technology in selection: A critical review and agenda for future research. *International Journal of Selection and Assessment*, *11*, 121–136.
- Anderson, N., Lievens, F., van Dam, K., & Ryan, A. M. (2004). Future perspectives on employee selection: Key directions for future research and practice. *Applied Psychology: An International Review*, *53*, 487–501.
- Anseel, F., & Lievens, F. (2009). The mediating role of feedback acceptance in the relationship between feedback and attitudinal and performance outcomes. *International Journal of Selection and Assessment*, *17*, 362–376.
- Arvey, R. D., & Sackett, P. R. (1993). Fairness in selection: Current developments and perspectives. In N. Schmitt & W. Borman (Eds.), *Personnel selection* (pp. 171–202). San Francisco: Jossey-Bass.
- Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*, 695–716.
- Bartram, D. (2000). Internet recruitment and selection: Kissing frogs to find princes. *International Journal of Selection and Assessment*, *8*, 261–274.
- Bauer, T. N., Maertz, C. P., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology*, *83*, 892–903.
- Bauer, T. N., Truxillo, D. M., Paronto, M. E., Weekley, J. A., & Campion, M. A. (2004). Applicant reactions to different selection technology: Face-to-face, interactive voice response, and computer-assisted

- telephone screening interviews. *International Journal of Selection and Assessment*, 12, 135–148.
- Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, 54, 387–419.
- Bauer, T. N., Truxillo, D. M., Tucker, J. S., Weathers, V., Bertolino, M., & Erdogan, B. (2006). Selection in the information age: The impact of privacy concerns and computer experience on applicant reactions. *Journal of Management*, 32, 601–621.
- Beaty, J. C., Dawson, C. R., Fallaw, S. S., & Kantrowitz, T. M. (2009). Recovering the scientist–practitioner model: How I/Os should respond to unproctored internet testing. *Industrial and Organizational Psychologist*, 2, 38–53.
- Bies, R. J., & Moag, J. S. (1986). Interactional justice: Communication criteria of fairness. In R. J. Lewicki, B. H. Sheppard, & B. H. Bazerman (Eds.), *Research on negotiation in organizations* (Vol. 1, pp. 43–45). Greenwich, CT: JAI Press.
- Cappelli, P. (2001). Making the most of online recruiting. *Harvard Business Review*, 79, 139–146.
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82, 143–159.
- Chan, D., Schmitt, N., DeShon, R. P., Clause, C., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, 82, 300–310.
- Cober, R. T., Brown, D. J., Blumental, A. J., Doverspike, D., & Levy, P. (2000). The quest for the qualified job surfer: it's time the public sector catches the wave. *Public Personnel Management*, 29, 1–18.
- Cober, R. T., Brown, D. J., Keeping, L. M., & Levy, P. E. (2004). Recruitment on the net: How do organizational web site characteristics influence applicant attraction? *Journal of Management*, 30, 623–646.
- Cohen, R. L. (1987). Distributive justice: Theory and research. *Social Justice Research*, 1, 19–40.
- Coyle, B. W., White, J. K., & Rauschenberger, J. (1978). Background, needs, job perceptions, and job satisfaction: A causal model. *Personnel Psychology*, 31, 889–901.

- Dineen, B. R., Ling, J., Ash, S. R., & Del Vecchio, D. (2007). Aesthetic properties and message customization: Navigating the dark side of recruitment. *Journal of Applied Psychology, 92*, 356–372.
- Gatewood, R., Feild, H. S., & Barrick, M. (2007). *Human resource selection* (6th ed.). South-Western College Publishing.
- Gibby, R. E., Ispas, D., McCloy, R. A., & Biga, A. (2009). Moving beyond the challenges to make unproctored internet testing a reality. *Industrial and Organizational Psychologist, 2*, 38–53.
- Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694–734.
- Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79*, 691–701.
- Gilliland, S. W., Groth, M., Baker, R. C., Dew, A. F., Polly, L. M., & Langdon, J. C. (2001). Improving applicants' reactions to rejection letters: An application of fairness theory. *Personnel Psychology, 54*, 669–703.
- Harris, M. M., Van Hove, G., & Lievens, F. (2003). Privacy attitudes toward internet-based selection systems: A cross-cultural comparison. *International Journal of Selection and Assessment, 11*, 230–236.
- Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639–683.
- Horrigan, J. B., & Rainie, L. (2002). *Getting serious online*. Washington, DC: Retrieved Pew Internet and American Life Project. November 24, 2009, from www.pewinternet.org/reports/toc.asp?Report=55.
- Hunthausen, J. M. (2000). Predictors of task and contextual performance: Frame-of-reference effects and applicant reaction effects on selection system validity. Unpublished doctoral dissertation. Portland State University.
- Kaminski, K. A., & Hemingway, M. A. (2009). To proctor or not to proctor? Balancing business needs with validity in online assessment. *Industrial and Organizational Psychology, 2*, 24–26.
- Koong, K. S., Liu, L. C., & Williams, D. (2002). An identification of internet job board attributes. *Human Systems Management, 21*, 129–135.
- Leventhal, G. S. (1980). What should be done with equity theory? New approaches to the study of fairness in social relationship. In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27–55). New York: Plenum.

- Liden, R. C., & Parsons, C. K. (1986). A field study of job applicant interview perceptions, alternate opportunities, and demographic characteristics. *Personnel Psychology*, *39*, 109–122.
- Lievens, F. (2006). The ITC guidelines on computer-based and internet-delivered testing: Where do we go from here? *International Journal of Testing*, *6*, 189–194.
- Lievens, F., De Corte, W., & Brysse, K. (2003). Applicant perceptions of selection procedures: The role of selection information, belief in tests, and comparative anxiety. *International Journal of Selection and Assessment*, *11*, 67–77.
- Lievens, F., & Harris, M. (2003). Research on internet recruiting and testing: Current status and future directions. In C. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 18, pp. 131–165). Chichester, England: John Wiley & Sons.
- Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology*, *92*, 1043–1055.
- Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, *47*, 715–738.
- Martin, C. L., & Nagao, D. H. (1998). Some effects of computerized interviewing on job applicant responses. *Journal of Applied Psychology*, *74*, 72–80.
- Peters, K. (2001). Five keys to effective e-cruiting. *Ivey Business Journal*, *65*, 8–10.
- Piotrowski, C., & Armstrong, T. (2006). Current recruitment and selection practices: A national survey of Fortune 1,000 firms. *North American Journal of Psychology*, *8*, 489–496.
- Piturro, M. (2000). The power of e-cruiting. *Management Review*, *89*, 33–37.
- Ployhart, R. E., Ehrhart, K. H., & Hayes, S. C. (2005). Using attributions to understand the effects of explanations on applicant reactions: Are reactions consistent with the covariation principle? *Journal of Applied Social Psychology*, *35*(2), 259–296.
- Ployhart, R. E., & Harold, C. M. (2004). The applicant attribution-reaction theory (AART): An integrative theory of applicant attributional processing. *International Journal of Selection and Assessment*, *12*, 84–98.
- Ployhart, R. E., Ryan, A. M., & Bennett, M. (1999). Explanations for selection decisions: applicants' reactions to informational and sensitivity features of explanations. *Journal of Applied Psychology*, *84*, 87–106.

- Richman-Hirsch, W. L., Olson-Buchanan, J., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880–887.
- Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26*, 565–606.
- Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, S. D. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology, 85*, 163–179.
- Sanchez, R. J., Truxillo, D. M., & Bauer, T. N. (2000). Development and examination of an expectancy based measure of test-taking motivation. *Journal of Applied Psychology, 85*, 739–750.
- Schleicher, D. J., Venkataramani, Y., Morgeson, F. P., & Campion, M. A. (2006). So you didn't get the job. Now what do you think? Examining opportunity-to-perform fairness perceptions. *Personnel Psychology, 59*, 559–590.
- Schuler, H. (1993). Social validity of selection situations: A concept and some empirical results. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 11–26). Mahwah, NJ: Lawrence Erlbaum Associates.
- Shaw, J. C., Wild, E., & Colquitt, J. A. (2003). To justify or excuse? A meta-analytic review of the effects of explanations. *Journal of Applied Psychology, 88*, 444–458.
- Silvester, J., & Anderson, N. R. (2003). Technology and discourse: A comparison of face-to-face and telephone employment interviews. *International Journal of Selection and Assessment, 11*, 206–214.
- Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*, 49–76.
- Society for Industrial and Organizational Psychology, Inc. (SIOP). (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
- Steiner, D. D., & Gilliland, S. W. (2001). Procedural justice in personnel selection: International and cross-cultural perspectives. *International Journal of Selection and Assessment, 9*, 124–137.
- Sylva, H., & Mol, S. T. (2009). E-recruitment: A study into applicant perceptions of an online application system. *International Journal of Selection and Assessment, 17*, 311–323.
- Thibaut, J. W., & Walker, L. (1975). *Procedural justice: a psychological analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2, 2–10.
- Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, 59, 189–225.
- Truxillo, D. M., & Bauer, T. N. (in press). Applicant reactions to selection. In S. Zedeck, H. Aguinis, W. Cascio, M. Gelfand, K. Leung, S. Parker, & J. Zhou (Eds.), *APA handbook of industrial and organizational psychology* (Vol. 2). Washington, DC: APA Press.
- Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology*, 87, 1020–1031.
- Truxillo, D. M., Bauer, T. N., & Sanchez, R. J. (2001). Multiple dimensions of procedural justice: Longitudinal effects on selection system fairness and test-taking self-efficacy. *International Journal of Selection and Assessment*, 9, 336–349.
- Truxillo, D. M., Bodner, T. E., Bertolino, M., Bauer, T. N., & Yonce, C. A. (2009). Effects of explanation on application reactions: A meta-analytic review. *International Journal of Selection and Assessment*, 17, 346–361.
- Waung, M., & Brice, T. (2007). The effects of acceptance/rejection status, status notification, and organizational obligation fulfillment on applicant intentions. *Journal of Applied Social Psychology*, 37, 2048–2071.

Chapter Seven

INTERNATIONAL ISSUES, STANDARDS, AND GUIDELINES

Dave Bartram

In the relatively recent past, it was possible to think of individual countries as ‘closed systems’. Changes could be made in terms of best practice, procedures, and laws affecting the use of tests in one country without there being any real impact on practice or law in other countries. People tended to confine their practice of assessment to one country and test suppliers tended to operate within local markets—adapting prices and supply conditions to the local professional and commercial environment.

This has changed. Assessment is an international business. Many test publishers are international organizations, selling their tests in a large number of countries. Many of the organizations using tests for selection and assessment at work are multinationals. In each country test suppliers and test users are likely to find differences in practices related to access, user qualification, and legal constraints on test use.

Not only are organizations becoming more global in their outlook, but so too are individuals. The increased mobility of people and the opening of access to information provided by the Internet have radically changed the nature of the environment in which we operate.

Despite globalization, there remain large variations between countries in the standards they adopt and the provision they make

for training in test use. There are differences in the regulation of access to test materials and the statutory controls and powers of local psychological associations, test commissions and other professional bodies (Bartram & Coyne, 1998a, b; Muñiz, Prieto, Almeida, & Bartram, 1999; Muñiz, Bartram, Evers, Boben, Matesic, Glabeke, Fernández-Hermida, & Zeal, 2001). Because of this, there have been a number of initiatives to establish international agreement on what constitutes good practice, what constitute “good” tests and what the criteria should be for qualifying people as test users.

In this chapter, we look at the impact the emergence of online testing has had on assessment practices on an international basis. We will review the attempts that have been made to define good practice in this area and the complexities of differences in law and custom around the world that affect practice. The chapter will also address the complex issue of using comparable measures to assess and compare people from diverse countries and cultures who use different languages (or use the same languages differently).

The Globalization of Assessment Through Technology

Bartram (2008a) has described how, prior to the invention of the printing press, the written word was controlled by a small number of skilled experts; literary “guardians” who managed the flow of knowledge and information to others. The ability to mass produce books changed that forever. Within a few decades of the invention of moveable type, marked by the publication of Gutenberg’s bible in 1455, the mass production of books meant that anyone could have access to this information first-hand. The impact of this was profound, although slow to spread, as it relied on people developing the literacy skills necessary to take advantage of the newfound access to books. Up until the development of the World Wide Web (Berners-Lee, 1999), employment testing had remained in the Gutenberg era of print. While printing evolved over the centuries following Gutenberg, it remained the case that paper-and-pencil tests were physical materials that required costly printing, warehousing, and shipping and the distribution of key intellectual property to end-users (such as scoring keys, norm tables, and so on). The web created as big a revolution in how information flows as Gutenberg had done over half a millennium earlier.

Now the impact of such changes are felt much more quickly. In the last century it took radio about thirty-five years to develop a world-wide audience of fifty million people; for television, this time shrank to thirteen years. For the web, it took a mere four years. While there was computer-based testing before the web, so too there were books before Gutenberg, but the stand-alone computer was not an effective delivery mechanism. Prior to the web, computers provided an advantage for materials that could not be presented in paper format and arguably provided advantages in terms of standardization of presentation, but their relatively high costs and inaccessibility reduced the impact of computerization on test logistics. Prior to the advent of web-based distribution, the computer had made very little inroads on the volume of paper-and-pencil testing.

The main impact the use of the Internet has had in testing has been to enable assessment programs to run multinationally. Related to this has been the development of remote unproctored forms of assessment and the development of new ways of managing the risks associated with 'assessment at a distance', especially in high-stakes situations. The use of the Internet for assessment raises many other issues, such as the impact of remote assessment on applicant reactions (Anderson, 2003), implications for the design of robust systems, the use of complex test forms and online simulations, to name but a few. The most rapid expansion in the use of online testing has been in the area of recruitment. Here there has been an insatiable demand for fast, reliable, and valid assessment at the front end of the recruitment funnel. For post-hire applications, we have seen the mushrooming of personnel appraisal and personal development planning systems (360-degree feedback) as the logistical advantages of managing distributed assessment over the Internet are realized. While some of these issues will be touched on here, for a more extensive treatment the interested reader is referred to Bartram (2006, 2008a) and to Bartram and Hambleton (2006).

Growth of the Web

The impact of the hypertext web interface in the mid-1990s was only one of the enablers for the growth of the Internet. The other was the availability of low-cost hardware with user-friendly operating systems and applications. Since 1995, the web audience

has grown to over 1,600 million people (as of 31 March 2009: figures from Miniwatts Marketing Group) or around 25 percent of the world's population. However, the pattern of distribution of usage is very uneven. The Nordic countries and other northern European countries have usage levels around 80 to 90 percent of the population, while many of the eastern European countries are around 15 to 20 percent. Rates of growth in usage (from 2000 to 2008) vary considerably: Areas like the Middle East have experienced a 1,296 percent increase in that period, with usage in 2008 at 23.3 percent of the population. Africa has seen a growth rate of 1,100 percent, with usage rising to 5.6 percent of population. Over the same time, European usage has grown by 274 percent to 48.9 percent and North America by just 133 percent to 74.4 percent. It is clear that the areas of the world where usage is relatively low are growing fastest and showing signs of catching up with the more developed areas. We are now seeing ceiling effects as countries like Iceland (with the world's highest level of usage at 90 percent of population) having very little room for further growth. It may be a few years before Africa reaches these levels.

These differences in levels of penetration and differences in growth rates make forecasting difficult. However, it is clear that the Internet is now a basic part of the infrastructure of most of the developed world and the developing world is catching up fast. Infrastructure development is now focused more on increasing bandwidth and speed of operation. Nevertheless, we still find that employment assessment programs in South Africa cannot rely wholly on Internet delivery as the infrastructure is not sufficiently embedded. A further constraint on online delivery of assessment is that many of the current systems were developed in North America or Europe and were designed to work with the Roman alphabet of single byte characters. A major investment is needed in the software required to deliver tests in language that use double byte characters (such as Chinese). For these and a range of other reasons, an alternative to technology-based delivery (such as the Gutenberg solution of printed paper) will still be needed for some time to come in multinational assessment programs that encompass the developing as well as the developed nations. However, even in these cases, technology can assist through automation of scoring and online data aggregation.

Impact of the Online Revolution on Testing

An obvious impact of the Internet is that tests and documents can be downloaded directly to users. This means that the web can be used as a complete commercial solution for test publishers. There is no longer any need for printing and production, warehousing, and postal delivery services. More significant for testing, however, is the shift in locus of control provided by the web from the “client side” to the “server side.” For paper-and-pencil testing, publishers had to provide users with test items, scoring keys, and interpretation algorithms. As these were released into the public domain, the danger of compromise and security breaches was high. Test users can (and do) pass these materials on to people who are not authorized to use them. All the test data also reside with the user. The process of developing norms, checking the performance of test items, and carrying out validation studies is dependent upon costly procedures for recovering data from users.

For the Internet that situation is reversed. The data and the intellectual property reside on the publisher’s server. The users have access only to those parts of the process that they need. In formal terms, the test user takes on the role of a “data controller,” while the publisher has the role of “data processor.” This provides distributors with potentially better levels of control over intellectual property and over the management of personal data. Both of these issues are addressed in more detail later.

International Guidelines for Technology-Based Assessment

As the globalization of testing has increased, so the need for some international agreement on good practice has also increased. The International Test Commission (ITC) was established in 1976. However, in the past two decades it has assumed a much greater importance than it had in its early days. Since the early 1990s it has taken an increasingly important leadership role in the development of international guidelines relating to various aspects of testing (see Oakland, 2006, for a more detailed history of the ITC). The ITC is an association of national psychological associations, test commissions, publishers, and other organizations and individuals

committed to promoting effective testing and assessment policies and to the proper development, evaluation, and uses of educational and psychological instruments. The ITC is responsible for the *International Journal of Testing* and publishes a regular newsletter, *Testing International* (available from the ITC website: www.intestcom.org). It also holds biennial international conferences on issues of concern to testing within a global context.

Of direct relevance to the present chapter is the work the ITC has done on the development of international guidelines on computer-based testing and testing on the Internet. In 2001, the ITC Council initiated a project on developing guidelines on good practice for computer-based and internet-based testing (see Coyne & Bartram, 2006, for details of the development). The aim was not to “invent” new guidelines but to draw together common themes that run through existing guidelines, codes of practice, standards, research papers, and other sources and to create a coherent structure within which these guidelines can be used and understood. Furthermore, the aim was to focus on the development of guidelines specific to computer-based and Internet-based testing, not to reiterate good practice issues in testing in general (these are covered in the ITC’s Guidelines on Test Use). Clearly, any form of testing and assessment should conform to good practice issues, regardless of the method of presentation.

Contributions to the guidelines were made by psychological and educational testing specialists, including test designers, test developers, test publishers, and test users drawn from a number of countries. Many of these contributions came through the ITC Conference held in Winchester, UK, in 2002 (subsequently published in Bartram & Hambleton, 2006). The Guidelines were completed in 2005 (ITC, 2006). They incorporated material from the report of the APA Task Force on Testing on the Internet (Naglieri, Dragow, Schmit, Handler, Prifitera, Margolis, & Valesquez, 2004) as well as building on other relevant guidelines, such as the ATP Guidelines on Computer-Based Testing (ATP, 2002).

The ITC guidelines address four main issues:

1. Technology—ensuring that the technical aspects of CBT/Internet testing are considered, especially in relation to the hardware and software required to run the testing.

2. Quality—ensuring and assuring the quality of testing and test materials and ensuring good practice throughout the testing process.
3. Control—controlling the delivery of tests, test-taker authentication, and prior practice.
4. Security—security of the testing materials, privacy, data protection, and confidentiality.

Each of these is considered from three perspectives in terms of the responsibilities of:

1. The test developer
2. The test publisher
3. The test user

In addition, a guide for test-takers has been prepared.

A key feature of the Guidelines is the differentiation of four different modes of test administration:

1. Open mode—where there is no direct human supervision of the assessment session. Internet-based tests without any requirement for registration can be considered an example of this mode of administration.
2. Controlled mode—where the test is administered remotely and made available only to known test-takers. On the Internet tests, such tests require test-takers to obtain a logon username and password. These often are designed to operate on a one-time-only basis.
3. Supervised/proctored mode—where there is a level of direct human supervision over test-taking conditions. For Internet testing this requires an administrator to log-in a candidate and confirm that the test had been properly administered and completed. Much traditional paper-and-pencil testing falls into this category: administration conditions can be quite variable (numbers of candidates can vary, the comfort and distraction levels in the venue can vary, and so on) but they have in common the presence of someone whose role it is to supervise the session.
4. Managed mode—where there is both a high level of human supervision and also strict control over the test-taking environment.

In CBT testing this is normally achieved by the use of dedicated testing centers, where there is a high level of control over access, security, the test administration environment, the quality and technical specifications of the test equipment, and the qualification of test administration proctors and other staff.

The main impact of the Internet on testing practice within the work and organizational sphere has been on the adoption of “controlled” mode for large-scale assessment in employment testing. While this mode is also referred to as UIT (unproctored Internet testing; Tippins, 2009) it is important to distinguish between open and controlled modes, both of which are unproctored. Controlled mode has introduced a whole range of new challenges for test designers around ensuring security of test content, protection against cheating, and methods of candidate authentication (Bartram, 2008a).

Cultural Factors Affecting Assessment and the Choice of Norm Reference Groups

International assessment raises the whole question of how to define the characteristics of the groups of people from whom applicants may be drawn. It is not sufficient to identify them simply as belonging to a country or to assessment as being either national or international. Unfortunately, it is more complex than that. For the purpose of defining comparison groups for score interpretation, norm groups of people can be thought of in terms of four main sets of variables (Roe, personal communication, 2008):

1. Endogenous person variables (biological characteristics such as gender, age, race).
2. Exogenous person variables (environmental characteristics such as educational level and type, job level and type, organization worked for, industrial sector, labor market, language).
3. Situational examination and format variables (paper and pencil or computer, online or offline, proctored or unproctored) setting and “stakes” (pre-screening, selection, development, research).

4. Temporal variables (generational differences relating to when norm data was gathered and over what time span, the currency of norms in terms of the timescale over which scores might be expected to change).

Any “norm” group can be defined in terms of people sampled across a range of the above factors. Culture, and hence culture-specific norms, can be defined as a set of exogenous variables relating to shared values, cognitions, knowledge, standards or cultural norms, and language. In practical terms, “culture” matters for assessment purposes when it is related to people for whom within-group variability on relevant constructs is relatively small compared to variability on the same constructs between them and other groups. Defined in this way it is clear that culture can vary within as well as between countries, as too can language. Traditionally, we have regarded within-country comparisons as being associated with a single language version of a tests with a “country norm,” while between-country comparisons were more problematic. For example, Bartram (2000) presented the following example:

“An Italian job applicant is assessed at a test centre in France using an English language test. The test was developed in Australia by an international test publisher, but is running from an ISP located in Germany. The testing is being carried out for a Dutch-based subsidiary of a U.S. multinational. The position the person is applying for is as a manager in the Dutch company’s Tokyo office. The report on the test results, which are held on the multinational’s intranet server in the U.S., are sent to the applicant’s potential line-manager in Japan having first been interpreted by the company’s out-sourced HR consultancy in Belgium.”

However, we should not lose sight of the fact that in most cases similar levels of complexity may apply to within-country assessments. Most countries these days are a mix of peoples from differing linguistic and cultural backgrounds. The issue of how best to make comparisons between people requires some guidelines that are not based on outmoded notions of countries as homogeneous collections of people who are distinct from the people found in other countries.

International Guidelines on Test Adaptation

Test adaptation refers to the process of modifying a test such that it operates in an equivalent way for two or more groups. Often this is identified with the process of linguistic translation, but adaptation is more than that. Just as much adaptation may be needed to use a test developed in the USA in the UK as would be needed to use one developed in the UK in France. The ITC Guidelines on Test Adaptation (Hambleton, 2005) were the first guidelines developed by the ITC and have had a major influence on setting standards for test adaptation. The guidelines focus on the qualities required of good tests. They were developed by a thirteen-person committee representing a number of international organizations. The objective was to produce a detailed set of guidelines for adapting psychological and educational tests for use in various different linguistic and cultural contexts (Van de Vijver & Hambleton, 1996). This is an area of growing importance as tests become used in more and more countries, and as tests developed in one country are used in another. The ITC Adaptation Guidelines apply wherever tests are moved from one cultural setting to another—regardless of whether there is a need for translation. Hambleton (1994) describes the project in detail and outlines the twenty-two guidelines that have emerged from it. The current version of these guidelines (Hambleton, 2005) fall into four main categories: the cultural context, the technicalities of instrument development and adaptation, test administration, and documentation and interpretation. Over the past few years, work has been carried out on the first major update and revision to these guidelines. This is scheduled to be published during 2011.

The new version of the guidelines will take account of a wider range of development models. The original version of the guidelines was focused on ability and achievement testing and on well-established tests in one language being adapted for use in another. While this model is still relevant, there are a range of new models that have been developed with international use in mind. That is, models which focus on the development of tests for use in multiple countries and cultures from the outset and not on the adapting of existing tests. For example, Bartram

(2008b) described the development of new item content through parallel item writing simultaneously in three different countries and languages (Chinese, German, and English). Items were reviewed and then translated into the other two languages for trialing. At the end of the process, one was left with a pool of items that had been originated in a diverse set of cultures and that had been calibrated for use in all of them. Experience has tended to show that adaptation problems increase with the increases in cultural distance. By originating test content in culturally diverse environments, the likelihood of future issues arising during adaptation into new target languages is reduced. It might be thought that such a process would reduce the “common” pool of item content to the lowest common denominator. In fact the opposite was the case. Items originated in China worked well in English and increased the variety of English item content (and vice versa for items originated in the UK for the variety of Chinese items). Test development models like this provide ways of increasing the chances of getting a broad coverage of the constructs being assessed and ensuring that the content will be able to be adapted for use in yet other languages and cultures.

Guidance on the Development and Use of Local vs. Global Norms

It is still standard practice to norm tests using national samples—generally with some acknowledgment to ethnic mix demographics but often with no analysis of the size of effects associated with cultural demographics. However, the notion of “national culture” and the identification, in testing, of norms with nationally defined standardization samples or, more commonly, aggregations of user norms, is highly problematic. *The unit of analysis* and the level of aggregation of data should not be defined in terms of some arbitrary political construct (like a nation), unless it can be shown that this corresponds to a single culture or homogeneous group. Definition of the unit of analysis should be tied to the operational definition of culture and the basic notion of relative homogeneity within and heterogeneity between groups. As noted above, culture only matters for assessment purposes when

it is related to some effect or impact on scores that is a group level effect and that is large enough to result in misinterpretation of individual level scores.

The key question to answer in choosing a norm reference group was clearly stated by (Cronbach, 1990) as “Does the norm group consist of the sort of persons with whom [the candidate] should be compared?” (p. 127). Cronbach makes clear that this does not even entail comparing people to others from their own demographic group. In answering Cronbach’s question, the test user also needs to consider whether the focus should be on a broad or a narrow comparison. The broader the comparison group, the greater the degree of aggregation required across situational, temporal, endogenous, or exogenous population variables. A norm group could be narrowly defined, for example, as sampling twenty-five-year-old, white, native-English-speaking males who are graduates in biological science subjects and who were born between 1980 and 1982, or broadly defined, for example, as sampling adult working people tested between 1950 and the present day and working in Europe. Commonly norms represent aggregations of samples from specific populations, but generally are aggregated within rather than across countries. Bartram (2008c) describes how organizations might choose between the use of national and international norms in the evaluation of people who are to work in international settings. Such guidance recognizes the need to know both the possible impact on the inferences one might draw from using one or the other type of norm (that is, by how much a given standard score would change) and consequences of that impact for decisions that might be made (for example, whether or not to hire someone for an international placement).

The use of aggregation across countries (subject to the use of appropriate country weightings) has the potential benefit of reducing the problem of country-related sample biases where these are present. It also does not conceal effects of cultural differences, which are hidden by using culture-specific norm groups. It does, though, have the potential negative effect, where there are real language biases, of treating apparent country differences as “real” rather than being due to language translation bias.

There is no space in the present chapter to go into details of procedures for aggregation, but some guidelines have been suggested by Bartram (2008c). These guidelines describe the formation of norm groups by the aggregation of suitably weighted populations or user norms for a specific purpose. Key to this process is checking that the constructs measured by the test are invariant across samples. It makes no sense to aggregate data from two scales that happen to have the same name but actually measure different things. For complex multi-dimensional instruments, construct equivalence can be checked by seeing whether scale variances and scale inter-correlations are invariant across samples. Where such instruments have well-fitting factor models in the source version, confirmatory factor analysis also can be used to check the equivalence of the structure of the adaptation of the instrument.

When more than one language has been used in questionnaire administration or a single language of administration has been used but the candidates are from different linguistic and cultural backgrounds, this should be taken into account during the interpretation process. This can be achieved through working with people who have expertise in testing and assessment and who are familiar with the culture and the local cultural meaning of behaviors. Without this it is very easy to misinterpret behavior or fail to appreciate the underlying potential because the person is conforming to unfamiliar business or social practices.

Comparing all candidates' scores to the same (multinational, multicultural or multilingual) norm will accentuate the differences due to the cultural behavior patterns of their backgrounds—even though these may be moderated by experience of the different environment. Because the scores are influenced by these arbitrary cultural differences in behavior, the measurement of the level of the underlying trait will suffer. On the other hand, using only the individual language norms for each candidate will reflect the underlying trait levels without relating to any differences between cultural norms that may be relevant.

In international contexts, inferences from scores should take into account both factors. In interpreting the score, it is important to know whether it shows, say, a moderate or extreme tendency to behave in a particular manner in general (relative to the “home”

country norm) and how this would seem in a different context (multinational norm). Comparison of each of the local language norms with an aggregated multinational norm will show where the average profiles diverge. This information should be available to the interpreter, either through the norm information for the different groups, or through qualitative data on which a particular pair of countries or norms differs.

In summary, for international assessments it is recommended that construct invariance across groups should be established before any between-group comparisons are carried. Once this has been done, both local national and relevant multinational aggregations of national norms can be used and areas where these give rise to differences in normed scores can be highlighted and considered by users in the light of what is known about possible sample, translation, or cultural effects.

Case Study: Assessing Construct Equivalence: Comparing Scales Between and Within Countries

One of the key issues in making comparisons between people from different groups, whether be gender, age, culture, language, or whatever, is that the constructs one is using to compare people need to be the same. It is not necessary for people in different groups to score the same on a given construct; indeed, it may be an appropriate reflection of the characteristics that distinguish group membership that they do not. A strong test of the equivalence of a set of constructs is that the correlations between them are the same for both groups of interest. For large numbers of constructs this becomes a very demanding test.

OPQ32 is a personality inventory with thirty-two trait scales (SHL, 2006). Research was carried out on the forced-choice format version of this test to examine the equivalence of the thirty-two trait constructs across and within countries. The forced-choice version of the instrument (OPQ32i) presents

(Continued)

Case Study: Assessing Construct Equivalence: Comparing Scales Between and Within Countries (*Continued*)

candidates with sets of four statements, drawn from four different scales, and candidates have to choose which statement is “most like” them and which is “least like” them. Conventional scoring of this form produces ipsative scale scores. That is, a fixed number of points are allocated between the thirty-two scales according to the choices made. The consequence of this is that one degree of freedom is lost, as the scores on any set of thirty-one of the scales will wholly determine the score on the thirty-second. When using structural equation modeling (SEM) to compare correlation matrices of ipsative scales it is necessary to remove one scale. Recently, Brown and Bartram (2008; SHL, 2009) developed a multidimensional IRT model for scoring forced-choice data that produces normative scale scores (by adopting a different approach to scoring that finds the best fitting set of thirty-two scale score for the set of all possible pairs of choices made by the candidate).

Results of the first study looked at data from 74,244 working adults from nineteen countries: twelve European countries, U.S., South Africa, Australia, China, Hong Kong, India, and New Zealand. Country sample sizes ranged from 861 to 8,222, with an average of 3,713. There were fourteen different language versions: six UK English countries; U.S. English; twelve samples of different European languages; and one sample of “simplified” Chinese. Item level data were available on eleven of the European countries’ data, and normative IRT scores were computed for those. For each country, the correlation matrix of the thirty-two scales (for the normative scale comparisons) or thirty-one scales (for the ipsative scale comparisons) were compared with the UK English data. An exceptionally good fit was found for all English and European languages (Ipsative: median CFI = 0.982 [min 0.960], median RMSEA = 0.019 [max 0.028]—includes U.S., S Africa, and Australia. IRT Normative: median CFI = 0.989 [min 0.982], median RMSEA = 0.024 [max 0.029]—Europe only.) For the Chinese version (simplified Chinese) the test identified a slight

misfit in the model: CFI = 0.945, RMSEA = 0.033. Exploration of the cause of this showed that the constraints violated related to correlations between the scales: Rule Following and Conventional ($r = 0.67$ for the Chinese data but $r = 0.45$ for the English version); and Forward Thinking and Achieving ($r = 0.36$ for the Chinese data but $r = 0.17$ for the English version).

The second study was carried on 32,020 South African candidates in various industry sectors to assess construct invariance as well as scale mean differences between ethnic and first-language groups on the OPQ32. OPQ32i was administered in English in all cases. All candidates were proficient in English to at least Grade 12. The sample was 52.10 percent females and the mean age was 30.67 years ($SD = 8.23$). 47.60 percent were African, 13.50 percent coloured, 9.60 percent Indian, and 29.10 percent white. In terms of education, 37.39 percent were Grade 12, 16.316 percent had higher certificates, 30.99 percent degrees, and 15.31 percent post-graduate degrees. First-language was known for 25,094 of the candidates: 25.90 percent Afrikaans, 27.10 percent English 2.10 percent Venda, 2.20 percent Tsonga, 21.90 percent Nguni (Zulu, Xhosa, Swati, and Ndebele), and 20.80 percent Sotho (North Sotho, South Sotho, and Tswana).

The OPQ32i was scored both conventionally (as ipsative scale scores) and using the multidimensional IRT model to recover latent normative scores. Comparison of the covariance structures of the samples was carried out using SEM with EQS on both ipsative [$k = 31$] and the normative IRT latent trait scale scores [$k = 32$]. Both produce very similar results. The results from the normative scores were as follows.

For comparison purposes, the white group was treated as the "source" group and the other groups as "targets" (this is relatively arbitrary, as comparisons will be symmetrical for other pairings of groups). For comparisons between the white group ($N = 9,318$) and each of the other three ethnic groups the results were:

1. For African ($N = 15,255$) CFI = .972, RMSEA = 0.035

(Continued)

Case Study: Assessing Construct Equivalence: Comparing Scales Between and Within Countries (*Continued*)

2. For Coloured (N = 4,308) CFI = 0.992, RMSEA = 0.020
3. For Indian (N = 3,083) CFI = 0.997, RMSEA = 0.012

For comparisons between the English-first language group (N = 6,793) and each of the other five first-language groups the results were:

1. For Africans (N = 6,494) CFI = .998, RMSEA = 0.011
2. For Nguni (N = 5,488) CFI = 0.978, RMSEA = 0.031
3. For Sotho (N = 5,232) CFI = 0.976, RMSEA = 0.032
4. For Tsonga (N = 555) CFI = 0.991, RMSEA = 0.021
5. For Venda (N = 532) CFI = 0.990, RMSEA = 0.022

While all show exceptionally good levels of fit, it is interesting to note that the RMSEAs, while small, are larger for Nguni and Sotho than for others. This could be due to some random elements in the data caused by people having to operate in their second language, or it could reflect more systematic cultural differences. In the latter case, one would expect a high level of fit between Nguni and Sotho. This was indeed the case, with the CFI = 0.999 and the RMSEA = 0.006.

In conclusion, forced-choice item formats appear to be very robust in terms of construct equivalence across countries, and the present data also shows high levels of equivalence between first-language and ethnic groups within South Africa. It seems that this format does control for potential culture-related systematic sources of response bias associated with Likert rating scales. Together with IRT scoring models, this creates the possibility of bias resistant formats with normative scaling.

Nevertheless we also found differences between and within countries in terms of average scale scores. These effects are relatively small when compared with other demographics (for example, gender, managerial position) and the effect sizes are generally not of substantive significance in terms of individual profile interpretation.

In summary, before any instrument is used across languages or cultures it is necessary to establish at the very least construct equivalence. There should also be evidence to support metric equivalence. That is not only when the constructs are the same but the scale metrics should be comparable in terms of intervals and distribution shapes. Van de Vijver and Leung (1997) go on to define scalar or score equivalence as providing a basis for comparing scale means across cultures. The IRT modeling approach described above provides this level of equivalence for the underlying theta scale scores—which can then be “normed” either against culture-specific groups or cross-cultural groups. Guidance on testing levels of equivalence can be found in van de Vijver and Leung (1997).

The key to establishing scalar equivalence is in the elimination of method bias and freedom from biased items (items exhibited high levels of differential item functioning), which might act to raise or lower mean scale score in one group due to bias effects rather than trait-level differences.

International and National Legal and Professional Standards and Guidelines

In addition to the ITC Guidelines discussed above, there have been a number of other international developments relating to good practice in global test use. The ITC International Guidelines on Test Use (ITC, 2001) focus on the competence of the test user. The Test Use Guidelines project was started in 1995 (see Bartram, 2001, for details). The aim was to provide a common international framework from which specific national standards, codes of practice, qualifications, user registration criteria, etc could be developed to meet local needs. As with other ITC Guidelines, the intention was not to “invent” new guidelines, but to draw together the common threads that run through existing guidelines, codes of practice, standards, and other relevant documents and to create a coherent structure within which they can be understood and used.

The competencies defined by the guidelines were to be specified in terms of assessable performance criteria, with general outline specifications of the evidence that people would need for

documentation of competence as test users. These competences needed to cover such issues as:

- Professional and ethical standards in testing,
- Rights of the test candidate and other parties involved in the testing process,
- Choice and evaluation of alternative tests,
- Test administration, scoring, and interpretation, and
- Report writing and feedback.

The Guidelines in Test Use project received backing from national psychological associations around the world and from international bodies such as the European Association of Psychological Assessment (EAPA) and the European Federation of Psychologists' Associations (EFPA). The Guidelines were also endorsed by many European and U.S. test publishers. Copies of the full Guidelines were published in the first edition of the ITC's *International Journal of Testing* (ITC, 2001) and are now available in fourteen different languages from the ITC website (www.intestcom.org).

Standards and guidelines need to cover tests, the people who use tests, and the testing process. EFPA, whose membership includes thirty-five European countries, including Russia, have addressed the first two of these through the development of European Test Review Criteria and Test User Standards. The EFPA Test Review Criteria were developed by combining the best features of the British, Dutch, and Spanish test review procedures into a single document. The review criteria and supporting documentation are available from the EFPA website: www.efpa.eu. The British Psychological Society (BPS) and the Norwegian Psychological Association have both adopted these standards as the basis for all their reviews and for test registration and certification procedures. Other countries' psychological associations (including Sweden, Denmark, and Russia) are planning to follow suit.

EFPA in conjunction with the European Association of Work and Organizational Psychologists (EAWOP) has also developed a European set of standards defining test user competence. The ITC Guidelines on Test Use were used as the framework for these standards. The EFPA standards are now being used as the basis

for a major review in the UK of the BPS standards for test use and also form the basis for work in Sweden, Norway, Spain, the Netherlands, and Denmark on the development of competence-based test user certification procedures. Plans are now under way to establish a European system for the accreditation of national test user certification schemes that meet the EFPA standards.

All these guidelines attempt to define standards of good practice that can apply across countries. However, we have also to recognize the complex differences in law and practice between countries. For example, there are quite strict laws governing data privacy that relate to test scores in Europe, while other countries have much laxer laws. The European Union recognizes this by setting requirements that prevent test providers in Europe from storing candidate data outside of the EU unless special provisions have been made to ensure their safety (so-called “safe harbor” arrangements). Organizations operating in the global environment would be well advised to adopt those standards that will provide them with access to all countries, rather than trying to operate with minimal standards that might create problems for them in some parts of the world.

While there are considerable safeguards in place concerning the protection of personal data, a problem that arises for psychological assessment is the associated right that individuals have for access to their data “in a meaningful form.” This can place a considerable burden on those responsible for the data (the so-called “data controllers”) and pose issues for what data are stored and how they are stored. For example, if test data are to be retained for scientific research purposes, or norm generation, it must be possible to anonymize the data. System designs also need to take account of the fact that for some countries it is necessary to encrypt data for transmission, and in most cases it is a requirement that personal data be encrypted for storage. Test management systems need to consider such issues when devising the data models they will use.

The right of an individual to have access to and control over the fate of his or her personal data has led to some countries interpreting this as included the right to have access to the original items and scoring procedures. In some cases it has been a matter for debate within courts or tribunals to find a reasonable balance

between the rights of the individual and the need to protect the security of test materials and the intellectual property of others.

We also see marked variations in the positions adopted by psychological associations as to who should or should not be given access to test materials. In Finland it is argued that only psychologists should use psychological tests, while those in HR roles can use competency-based assessments (the distinction is a subtle one, but they would define a psychological test as one that measured psychological constructs, like personality traits, while a competency-based instrument focused on work-competency constructs). Next door in Sweden and Norway, the BPS model is followed: anyone who can demonstrate his or her competence (whether psychologist or not) can use tests. In Germany, the national standards institute (DIN) has been used as the mechanism to develop a national standard for assessment at work (DIN 33430). This has associated with it training courses to certify people.

The DIN process also has associated review procedures for quality assuring assessment methods and procedures. In Norway, Det Norske Veritas (DNV) use ISO normed procedures to certify tests according to the test criteria set by EFPA. The same criteria are used in the UK, but the procedures are different. The British Psychological Society provides both a *registration* option (which identifies whether an instrument meets the minimum psychometric requirements to be called a test) and a *review* option, which provides a detailed evaluation against the EFPA criteria. The Netherlands' Psychological Association also publishes regular reviews of tests, but bases these on a slightly different set of criteria.

In countries like South Africa, positive discrimination and quota systems in favor of under-represented black majority candidates is enshrined in employment law. In most other countries such discrimination would be illegal. Also South Africa has a very strong legal position forbidding the use, including administration, of tests by non-psychologists and requiring tests to be registered.

Protection of Intellectual Property

The issue of protecting intellectual property (IP) has been mentioned briefly. In practice, the advent of online testing has both made this easier and more complex. With online testing, it

becomes possible to ensure that key IP is not distributed (scoring algorithms, report generation rules, item banks, etc.). All that needs to be distributed are the items for candidates to respond to and the reports of the assessment.

Item generation and test generation technologies have made it less and less worthwhile for people to harvest ability test items with a view to making money from training others to “pass the test.” However, it is recognized that copyright law is less strictly observed in some countries (such as China and some of the Eastern European countries) than others.

There are methods that have proved effective in managing item content on the web. These include the use of “web patrols,” regular systematic searches of the web to find item content. Such patrols can find content that is being advertised illegitimately and steps can then be taken through ISPs to shut these down. In most cases, there is no need to have recourse to law. Companies like eBay have built in procedures (Vero) to help copyright holders deal with breaches.

In the end, the price one has to pay for the greater access provided by the Internet is the need for greater vigilance.

ISO Standards Relating to Technology-Enhanced Assessment

ISO is the International Organization for Standardization. It is the world’s largest developer and publisher of international standards. It has a central secretariat in Geneva, Switzerland, that coordinates a network of standards institutes of 162 countries. These include the British Standards Institute (BSI) in the UK, Deutsches Institut für Normung (DIN) in Germany, and the American National Standards Institute (ANSI) in the United States. ISO’s aim is to facilitate the development of consensus between countries on solutions that meet both the needs of business and the broader needs of society.

Recently, ISO has become involved in standards that relate to computer-based assessment. Valenti, Cucchiarelli, and Pantì (2002) reviewed use of ISO 9126 as a basis for computer-based assessment system evaluation. ISO 9126 is a standard for information

technology–software quality characteristics and sub-characteristics. The standard focuses on functionality; usability; reliability; efficiency; portability, and maintainability. Valenti, Cucchiarelli, and Panti (2002) base their review around the first three of these.

The British Standards Institute (BSI) published a standard (BS7988) in 2002: A Code of Practice for the use of information technology for the delivery of assessments. The standard relates to the use of information technology to deliver assessments to candidates and to record and score their responses. The scope is defined in terms of three dimensions—the types of assessment to which it applies, the stages of the assessment “life cycle” to which it applies, and the standard’s focus on specifically IT aspects. This standard has now been incorporated into an ISO standard: Information technology—a code of practice for the use of information technology (IT) in the delivery of assessments. [Educational] (ISO/IEC 23988: 2007). ISO23988 is designed to provide a means of:

- Showing that the delivery and scoring of the assessment are fair and do not disadvantage some groups of candidates, for example, those who are not IT literate;
- Showing that a summative assessment has been conducted under secure conditions and, through identify verification, checking that this the authentic work of the candidate;
- Showing that the validity of the assessment is not compromised by IT delivery; providing evidence of the security of the assessment, which can be presented to regulatory and funding organizations (including regulatory bodies in education and training, in industry, or in financial services);
- Establishing a consistent approach to the regulations for delivery, which should be of benefit to assessment centers that deal with more than one assessment distributor; and
- Giving an assurance of quality to purchasers of “off-the-shelf” assessment software.

It gives recommendations on the use of IT to deliver assessments to candidates and to record and score their responses. The scope does not include many areas of occupational and health–related assessment. While it includes “assessments of knowledge, understanding, and skills” (achievement tests), it excludes

“psychological tests of aptitude and personality.” However, much of what it contains can be generalized to psychological testing.

Most recently, work has been underway on ISO 10667 (Psychological Assessment Services). This has as its starting point the German work on DIN 33430 and is developing it into an international standard for the use of assessments in work and organizational settings. This has been a truly international collaboration, with input from most European countries, the U.S., China, and Africa. ISO 10667 is a service delivery standard. It sets out what constitutes good practice for both service providers delivering an assessment service and the clients who are the consumers of such services. As such it encompasses both issues of user competence and issues relating to the quality of methods and procedures used in assessment. It is possible that this standard will provide in future the framework within which other more specific sets of standards and guidelines can sit.

Conclusions

As the world of testing and assessment has become increasingly global, there has been a parallel development of standards and guidelines that attempt to consider cross-national issues. ISO, EFPA, and the ITC have all contributed to these developments, and all build on the work done at national levels.

This chapter has highlighted the complexities associated with working in a global environment. However, the key to success in managing this complexity is the normal mix of good science and professional practice. Good science provides the tools for ensuring our instruments are fit for use in a multinational, multicultural, and multilingual environment; good practice provides the basis for balancing the needs of organizations to make comparisons between people with the limitations we know to exist in all forms of objective measurement: science never provides all the answers, but it does provide a basis on which sound judgments can be made. Linking science and practice together is policy. Global organizational policies need to recognize the requirements legal constraints (for example, data privacy laws) impose on practice and need to make clear what is and what is

not acceptable practice in the use of assessments when operating in a wide range of different countries.

While setting standards is important to help ensure good practice, there are also major psychometric issues to address. We have seen that key to the use of tests across varying groups is the need to establish construct equivalence. Then one needs to ensure that score differences between groups are not due to some form of systematic bias associated with the test design or response mode, but reflect genuine differences in trait levels. When these conditions are satisfied, comparisons can be made across groups. Organizations make such comparisons daily, often on the basis of poor or biased information. The technologies we have developed for technology-enhanced assessment promise to make such comparisons fairer and more objective.

References

- Anderson, N. (2003). Applicant and recruiter reactions to new technology in selection: A critical review and agenda for future research. *International Journal of Selection and Assessment*, 11(2–3), 121–136.
- Association of Test Publishers (ATP). (2002). *Guidelines for computer-based testing*. www.testpublishers.org/documents.htm.
- Bartram, D. (2000). Internet recruitment and selection: Kissing frogs to find princes. *International Journal of Selection and Assessment*, 8(4), 261–274.
- Bartram, D. (2001). The development of international guidelines on test use: The International Test Commission project. *International Journal of Testing*, 1, 33–53.
- Bartram, D. (2006). Computer-based testing and the internet. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (Chapter 18, pp. 399–418). Oxford, UK: Blackwell.
- Bartram, D. (2008a). The advantages and disadvantages of online testing. In S. Cartwright & C. L. Cooper (Eds.), *The Oxford handbook of personnel psychology* (Chapter 10, pp. 234–260). Oxford: Oxford University Press.
- Bartram, D. (2008b). Extending the ITC guidelines—A case study: Parallel item creation in three languages. Paper presented in invited symposium convened by R. Hambleton on the ITC Guidelines and methodology for adapting educational and psychological tests. *XXIX International Congress of Psychology*, Berlin.

- Bartram, D. (2008c). Global norms: Towards some guidelines for aggregating personality norms across countries. *International Journal of Testing*, 8, 315–333.
- Bartram, D., & Coyne, I. (1998a). The ITC/EFPPA survey of testing and test use within Europe. In *Proceedings of the British Psychological Society's Occupational Psychology Conference* (pp. 197–201). Leicester, UK: British Psychological Society.
- Bartram, D., & Coyne, I. (1998b). Variations in national patterns of testing and test use. *European Journal of Psychological Assessment*, 14, 249–260.
- Bartram, D., & Hambleton, R. K. (2006). *Computer-based testing and the internet*. Chichester, UK: John Wiley & Sons.
- Berners-Lee, T. (1999). *Weaving the web: The original design and ultimate destiny of the world wide web by its inventor*. New York: HarperCollins.
- Brown, A., & Bartram, D. (2008). IRT model for recovering latent traits from ipsative ratings. Poster presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco.
- Coyne, I., & Bartram, D. (2006). Design and development of the ITC guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, 6, 133–142.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, Harper & Row.
- Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229–244.
- Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting tests in multiple languages and cultures. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Lawrence Erlbaum Associates.
- ITC. (2001). International guidelines on test use. *International Journal of Testing*, 1, 95–114.
- ITC. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, 6, 143–171.
- Muñiz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal, and Latin American countries. *European Journal of Psychological Assessment*, 15, 151–157.
- Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J. R., & Zaal, J. N. (2001). Testing practices in European countries. *European Journal of Psychological Assessment*, 17, 201–211.

- Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004, February). Psychological testing on the internet: New problems, old issues. *American Psychologist*.
- Oakland, T. (2006). The International Test Commission and its role in advancing measurement practices and international guidelines. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the internet* (pp. 1–11). Chichester, UK: John Wiley & Sons.
- SHL. (2006). *The OPQ32 technical manual*. Thames Ditton, UK: Author.
- SHL. (2009). *Supplement to the OPQ32 Technical Manual*. Thames Ditton, UK: Author.
- Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology*, 2(1), 2–10.
- Van de Vijver, F., & Hambleton, R. (1996). Translating tests: Some practical guidelines. *European Psychologist*, 1, 89–99.
- Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.
- Valenti, S., Cucchiarelli, A., & Panti, M. (2002). Computer-based assessment systems evaluation via the ISO9126 quality model. *Journal of Information Technology Education*, 1(3), 157–175.

Important Websites

- The International Test Commission: www.intestcom.org
- The International Organization for Standardization: www.iso.org
- The European Federation of Psychologists Associations: www.efpa.eu
- The American Psychological Association: www.apa.org
- The British Psychological Society's Psychological Testing Centre: www.psychtesting.org.uk

Section Two

Case Studies of Technology- Enhanced Assessments

Chapter Eight

WEB-BASED MANAGEMENT SIMULATIONS

Technology-Enhanced Assessment
for Executive-Level Selection and
Development

Terri McNelly, Brian J. Ruggeberg, and
Carrol Ray Hall, Jr.

This chapter presents a case study detailing the development of customized virtual assessment centers for three different executive-level positions within Darden Restaurants, Inc. Through subsidiaries, Darden owns and operates 1,800 Red Lobster, Olive Garden, LongHorn Steakhouse, The Capital Grille, Bahama Breeze, and Seasons 52 restaurants in North America, employs approximately 180,000 people, and serves four hundred million meals annually.

Organizational/Political Landscape

To achieve ambitious revenue growth goals, Darden realized it would need to grow exponentially, expanding through the acquisition or creation of new restaurant concepts and through organic

growth (higher guest counts). To support this overarching strategic company objective, Darden's Talent Management team was asked to ensure HR systems supported the recruitment, hiring, training, and development of the talent required. At the time, Darden operated as a highly matrixed and geographically dispersed organization, with each restaurant brand (in Darden language, "restaurant concept") having its own policies, procedures, systems, and processes, including HR (for example, Red Lobster's performance management system was different from Olive Garden's performance management system). The only enterprise-wide assessment in place in 2005 was a traditional assessment center for operations general managers (restaurant managers). While that assessment center was generally accepted by Darden Operations across the concepts, both the participants and the managers who served as assessors found the tools and processes to be time-consuming and administratively burdensome. In particular, there was concern regarding the number of managers who needed to be taken out of restaurants to serve as assessors as well as the amount of time these managers were out of their restaurants. Focus groups revealed a strong desire to shorten the overall duration of the assessment process and to reduce or even eliminate the use of internal assessors.

The organizational context, assessment history, and focus group feedback were all taken into careful consideration as Darden looked to design and implement a standardized assessment process for more senior leadership roles in the organization—the director, officer, and senior officer positions. Specifically, Darden sought a new assessment process that was:

- Customized for the Darden culture,
- Similar in process but distinct in content for the three levels,
- Cost-effective to deliver while still having a high-touch "feel" for candidates,
- Administratively efficient for both administrators and participants, and
- Able to be used for external selection, internal promotion, and internal development.

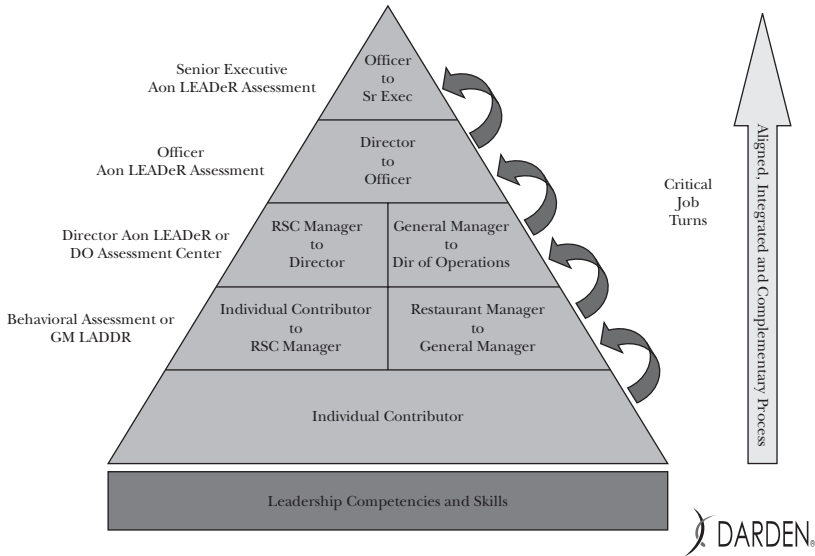
As a first step, an exhaustive analysis of those key jobs (director, officer, and senior officer) was undertaken. Successful incumbents

were interviewed to discover the skills required for effective performance across the various levels of the organization, within specific functions of the organization, and across the centralized corporate headquarters. Focus groups were conducted with key strategic stakeholders to learn what skills would be required in the future for Darden to meet its aggressive growth and expansion goals. Finally, surveys were conducted to define the skills, competencies, and language that would make up the new leadership competency models and success profiles.

Throughout these activities, a consistent message was communicated around the need to create a common language for the entire organization and to create standardized assessment processes to support growth and allow for easier movement between restaurant concepts. However, as would be expected in a change of this magnitude, there was initial resistance and skepticism, largely based on the belief or perception that the restaurant concepts were too different to employ standardized processes (for example, “Red Lobster operates different from Olive Garden so how can we develop people the same way?”). In addition, there was an inertia-based reluctance to give up existing, well-established assessment processes that were unique to each restaurant concept. Accordingly, the Talent Assessment team approached the initiative as a large-scale organizational change effort and recognized that each focus group, each interview, and each request for input was a chance to communicate the message and create champions for a more enterprise-wide approach.

The leadership competency models created from the job analysis work then served as the foundation for designing a comprehensive and aligned assessment architecture (Figure 8.1) that was applied across all the critical jobs within the organization.

With five operational brands at that time, and multiple functions within the corporate headquarters, the Talent Assessment team undertook a strategy that included leaders from a broad array of functions being involved in the change management process to introduce the competency models and the assessment architecture associated with those models. This approach built support and buy-in for incorporating the new leadership competency framework and language into the broader organization.

Figure 8.1. Assessing Critical Job Turns

Virtual Assessment Center Solution

As discussed earlier, there were several concerns from operators about the existing assessment center (for example, time-consuming and administratively burdensome on internal assessors). The idea of implementing a traditional assessment center using professional third-party assessors, while addressing concerns about the use of internal resources, was quickly dismissed due to the relatively high projected overall costs and the time commitment required of candidates/participants. As an alternative to the traditional assessment center, Darden considered two remote delivery alternatives. One alternative reviewed was implementing a telephone-based assessment program. In this solution, assessment materials resemble those used in a traditional assessment center. Participants access a static set of materials on the web from their own locations, read and prepare for the different exercises, and interact via the telephone with trained third-party assessors (see Chapter 10 in this book). This solution, while addressing issues of efficiency and cost, did not have the face validity or high-touch “feel” that Darden was seeking, especially for

higher-level positions. Ultimately, Darden selected a hybrid model that leverages technology to deliver a more dynamic experience and that effectively addresses the challenges, constraints, and objectives outlined above. More precisely, Aon's integrated leadership assessment platform (LEADeR) and a series of off-the-shelf, day-in-the-life, web-delivered business simulations, were leveraged to create unique virtual assessment centers for each leadership level.

The virtual assessment centers implemented at Darden combine the depth of the assessment center methodology (International Task Force on Assessment Center Guidelines, 2009) with the ease and cost-effectiveness of conducting the assessments remotely and individually. More specifically, to meet all of the defining characteristics of an assessment center as outlined by the International Task Force on Assessment Center Guidelines, we:

- Conducted a systematic job analysis effort to create competency models;
- Defined specific behaviors observable in the assessment that are then categorized into skills, which are in turn categorized into broader competencies;
- Created a skill-by-assessment technique matrix showing the linkage of measures to job-relevant competencies;
- Used multiple assessments techniques in each virtual assessment center, including cognitive ability measures (Aon's Business Reasoning Test and the Thurstone Measure of Mental Alertness), personality instruments (the 16PF), a behavioral interview, a web-enabled business simulation consisting of eight unique challenges (that include live, telephone role-play exercises) embedded in a program of varying lengths (three hours for director level, three and a half hours for officer level, and four hours for senior-officer level), a strategic presentation, and a debrief interview;
- Leveraged technology to deliver the job-related simulation (the day-in-the-life business simulation);
- Used multiple assessors to observe and evaluate candidates;
- Used only trained and certified assessors;
- Recorded behavior through notes from interviews and role plays and captured behavior from the web-based simulation (for example, emails written, decisions made, action

- plans submitted), with all behavior scored using behaviorally anchored rating scales; and
- Integrated data to create a report providing feedback at the competency level, the skill level, and the behavioral level across all assessment activities.

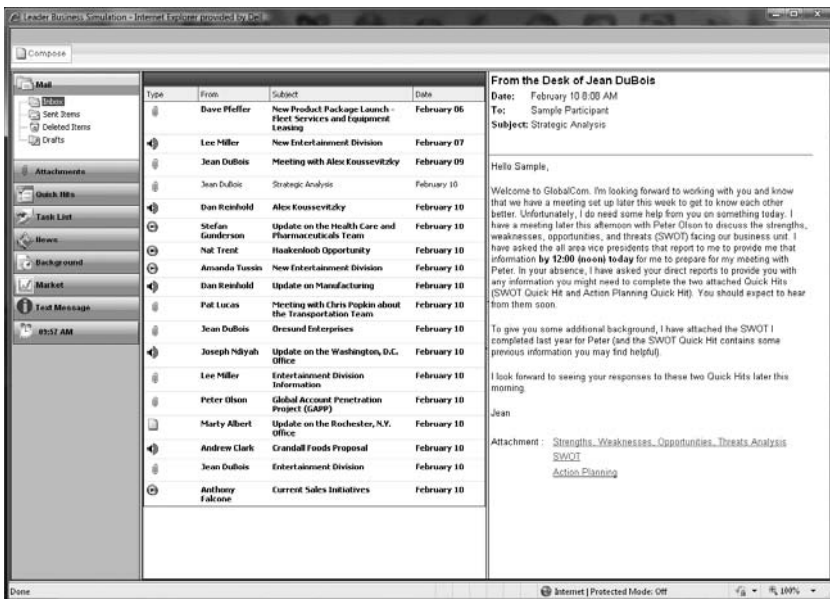
The Darden Talent Assessment team worked with Aon to map the newly identified leadership competencies to the assessment centers. Specifically, the behaviors assessed in each of the assessment components were compared to the competency and skill definitions in Darden's new competency model. This mapping also highlighted which skills and competencies were measured multiple times across activities and which skills needed additional measurement in order to develop a robust assessment of each skill and competency.

The business simulation for each assessment center was then customized to measure the Darden-specific skills and competencies and to reflect the complexity of the challenges for that specific level of the organization. For example, in the officer-level business simulation, the participant takes on the role of a division vice president for a nationwide convenience store chain. The participant encounters several challenges in the simulation that require him or her to:

- Conduct a SWOT analysis and complete an action plan describing current state and key issues associated with meeting personal, people, and business objectives;
- Lead a business planning initiative with franchise owners;
- Coach a direct report on curbing turnover and implementing key initiatives in his/her area (live telephone role play);
- Recommend sites for new store expansion;
- Influence others to invest in a new inventory control initiative;
- Handle a public relations issue regarding potentially contaminated milk products from a local vendor (live telephone role play);
- Resolve a conflict with a vendor regarding potential contract violations and safety/security issues (live telephone role play); and
- Communicate effectively with managers on key issues and risks facing the team (live telephone role play)

The participant receives email, voice mail, video mail, and text messages throughout the simulation related to these challenges, as in a real-world managerial environment (see Figure 8.2). Based on this information, participants are then asked to make decisions, provide rationale for decisions made, and in some cases, interact via telephone with various people within and outside of the organization. The dynamism of the technology-delivered business simulation enhances the richness of the participant experience by providing a more realistic 21st century “day-in-the-life” experience relative to traditional brick-and-mortar assessment centers. First, participants receive and send information via a computer, whereas in most brick-and-mortar assessment centers participants receive information on paper and must reply on paper. Second, the participant receives text messages and email to follow up on specific decisions made during the simulation, allowing for assessors to gain insight on rationale and how the participant prioritized information when making decisions. Third, throughout the simulation, information relevant to the challenges arrives non-sequentially, more closely mirroring

Figure 8.2. Screenshot of Business Simulation



real life than the traditional assessment center wherein the participant receives information in organized packets relevant to a specific exercise. For example, in one of the Darden assessments the participant receives information related to an upcoming coaching call with a direct report interspersed with information related to a new product launch, updates on status for different client teams, and information from HR about succession planning.

Third-party Ph.D.-level coaches serve as assessors for these assessments and are certified in the assessment process. A two-and-a-half-day certification program provides training on observing and evaluating candidate behavior in each of the assessment components, including conducting and evaluating role plays, behavioral interviews, interpreting personality measures, and cognitive ability tests. Assessors must pass tests and behavioral sample assessments throughout the training program in order to become certified. In addition, to prepare to deliver the Darden-specific assessments, certified assessors participate in a four-hour orientation and training on the customized assessment components and on Darden's structure and organizational culture. Assessors new to the Darden programs are shadowed by an experienced assessor to provide calibration and support throughout the assessment process.

The technology-driven platform allows for measurement of skills that would normally not be measured in a typical brick-and-mortar assessment center. Specifically, the system collects information throughout the simulation on when the participant received information, opened emails, listened to voice mails or video mails, received and responded to text messages, submitted information, used a "task list," scheduled meetings on a calendar, and sent or replied to emails. This allows for the measurement of skills such as personal productivity, multitasking, planning and organizing, and prioritizing.

Further, the technology-driven assessment center platform accommodates mini-situational judgment tests (called "Quick Hits") tied to the specific challenges embedded in the simulation. Quick Hits can be closed-ended (multiple-choice or forced-rank) or open-ended, and ask participants to make decisions about the most suitable tactics to address the situations they encounter

throughout the simulation. The system scores the closed-ended Quick Hits and allows the assessor to download the open-ended Quick Hits to score against behaviorally anchored ratings scales.

While the system is able to track and score some skills, the system-level scoring accounts for at most 10 to 15 percent of the scores ultimately reported in the Darden assessments; observations and evaluations from the various assessors comprise the bulk of the scoring. The system-level scoring is connected to a report-writing application, which is a technology-enabled platform that helps the assessors to review information, provide ratings based on behaviors observed during the exercises, and integrate information from all assessment components. The report-writing application provides the assessor with a systematic, consistent process for efficiently creating a detailed feedback report for each participant.

Implementation and Maintenance

Using a web-enabled tool presented a new medium for administering assessments at Darden. While there were no significant technological challenges per se, there were initial challenges to the organization's IT capacity in implementing the virtual assessment centers. Darden was working on a large number of information technology projects at that time, and the web-based business simulation delivery platform was on a long queue of other technology initiatives and priorities awaiting IT review. Reviewing security measures, firewalls, server capacity, and other technical parameters required a significant commitment of time and human capital. Internally, to justify this investment in a robust assessment architecture that leverages state-of-the-art technology, the Talent Assessment team emphasized the relevance of the assessment platform to Darden's strategic growth objective.

Initially, to create an assessment site, two file rooms were converted within Darden's human resources department to simulate an office setting. Both offices were outfitted with computers and telephones to accommodate the web-enabled assessment. Internal and external candidates, as well as assessors, all traveled

to the talent assessment headquarters in Orlando, Florida, for these assessments. The IT department helped to ensure no security or firewall issues would interrupt the online assessment experience. Smaller, but no less important, issues were reviewed to ensure an assessment would flow seamlessly once begun—print directories, dedicated telephone lines, password protection, and access to a printer. Computers were purchased specifically to be dedicated to these assessments with the appropriate hardware (RAM, sound card) and software (latest versions of Internet Explorer, JavaScript, Macromedia Flash, Windows Media Player) for ease of administration.

As an additional step in the broader change initiative, four high-potential leaders were initially asked to pilot the new assessment for their personal development and to provide feedback prior to the enterprise-wide launch. This pilot approach not only allowed for fine-tuning of the overall process (enhanced pre-assessment communications, refined report content, schedule adjustments) but also helped to create “assessment ambassadors” within the organization who championed the new process and helped gain buy-in.

The virtual assessment centers are now firmly entrenched in the talent management culture at Darden. Coming full circle, this application of technology-enabled assessments has allowed Darden and its strategic partners to convert the original general manager assessment center into a virtual assessment center at its new corporate headquarters in Orlando, Florida. What once were two small converted file rooms is now a new and dedicated 1,500-square-foot assessment facility encompassing nine assessment offices, two large conference rooms, and a computer control room that houses a dedicated server and other essential IT equipment. Darden is now embarking on the use of digital recording technology, web-based simulation role plays, and scoring software that allows assessment results to be digitally scored at any time and place by trained assessors. The ROI to Darden is expected to yield upwards of \$500,000 annually, gained from the direct reduction of costs from travel expenses, including large blocks of hotel rooms and conference space, as well as indirect costs through the reduction of time spent away from the job for internal candidates. More importantly, leaders are being assessed against those job-specific leadership competencies that are

expected to propel Darden's growth and allow it to fulfill many of its people- and values-based objectives.

Success Metrics and Insights

In order to support the overall value and impact of the leadership assessments, short- and long-term success metrics were established by which return on investment (ROI) could be evaluated. Participant feedback was gathered from the initial implementation and continues to be collected with each assessment conducted. In particular, participants are asked to complete a questionnaire regarding the entire assessment process. Responses provide insight into the perceived value and impact of the program as well as providing valuable information for continuous process improvement. Questions focus on participant perceptions of (1) scheduling and logistics, (2) quality and clarity of pre-assessment materials and communications, (3) on-site facilities and amenities, (4) simulation role-play exercises, (5) career directions interview, (6) cognitive ability tests, (7) professionalism and friendliness of third-party assessors, (8) the clarity and simplicity of the technology (the web-based simulation), and (9) the total assessment process/experience.

In addition to the ratings on the questionnaire, participants are asked for overall comments and recommendations for changes and improvements. Furthermore, participant feedback is sought from participants in the follow-up feedback and coaching sessions. This type of participant input has been a critical guide to identifying and prioritizing process improvements. For example, initial feedback from participants indicated that the communication about the assessment process was not clear. After revising the communication, participant feedback became overwhelmingly positive (100 percent of respondents rated the "quality and clarity of pre-assessment materials and information" as excellent or very good). Another example relating to the assessment itself indicated that some participants were not able to see the value or applicability of the cognitive ability measures to their overall assessment. To address this feedback, we created a brief orientation for each assessment component to transparently communicate the purpose. Current feedback shows that 97

percent of the participants rated the applicability of each component as “good” or better.

In addition to participant feedback, quarterly review meetings were established during which the project team from Darden and the consulting partner together review and discuss assessment results and trends. The data reviewed in these meetings include hires/promotions by level, average competency ratings across levels and broken out by internal and external candidates, observations and insights of assessors (areas of consistently strong performance, areas of consistently weaker performance, process improvement suggestions), and comparisons of average competency ratings to external norms and internal Darden standards.

It is worth noting that the external norms were established to provide a comparative benchmark of competency and skill performance of individuals at similar levels going through similar assessments in other organizations. Darden standards were established for each of the skills and competencies measured in the assessment on the basis of focus group feedback related to the expected level of performance at that level. Specifically, a group of nine executives representing all of the restaurant concepts were asked to provide ratings indicating the ideal level of proficiency for each of the skills and competencies measured in the assessments across each level of the organization. These standards were established as an interim organizational benchmark and will eventually be replaced by internal Darden norms based on actual data once large enough normative samples are obtained. After the norms and standards were established, the assessment report was customized to graphically display the external norm and Darden standard for comparison purposes.

Additional metrics became feasible after a reasonable number of participants had gone through the assessments. Specifically, analyses were conducted to determine the relationship between hire and promotion decisions and assessment results. Results showed that hire and promotion decisions tend to be positively related to overall assessment results ($r = .37$, $p < .001$, $N = 151$), such that those receiving higher recommendation scores are more likely to be hired/promoted. *Note, though, that this is not to be considered an uncontaminated predictive*

validity coefficient. Because assessment results are made available to Darden leaders, the correlation suggests that Darden leaders factor the assessment results into hire and promotion decisions. The results also suggest or support that the assessment process successfully identifies individuals whom Darden leaders believe exhibit job and organization culture fit (suggesting support for the validity of the process).

Continuous Process Improvements

The review meetings have also focused on process and logistical issues with an eye toward process improvement, cost reduction, and/or enhanced ROI. Among the process improvements identified and implemented on the basis of these reviews was the tailoring of assessment content and evaluation criteria to the retail and hospitality industry, which further enhanced the face validity of the assessments. This customization was done for each of the three levels of assessment and was driven by feedback from participants. The ability to make efficient and cost-effective changes to the assessments was significantly aided by the technology platform, which allowed for modular and centrally implemented edits and modifications that avoided a complete redesign of the program.

In an effort to sufficiently manage assessment costs while maintaining the quality of assessment and the output, a key process change was implemented in late 2008. In particular, Darden began taking advantage of the true virtual assessment approach offered by the web-based platform and telephone interactions, migrating from the use of on-site lead assessors to a completely remote process. This change provided immediate savings through expense reduction as well as assessor time reduction, while also increasing the flexibility of scheduling because assessor travel arrangements were no longer a consideration. The participants are still brought into the on-site assessment facility and an on-site facilitator greets them and provides the positive, high-touch interaction important to Darden's culture. However, the lead assessor and role players interact with the candidate via telephone, while the web-based business simulation provides the high fidelity day-in-the-life experience.

Lessons Learned

One critical lesson applied here was the value of involving IT professionals from both the vendor and the client organizations at the very start of the process to ensure all potential technology challenges are considered before initiating the project. We would argue that it is not enough to bring IT in at the start of the project; rather it is necessary to involve IT from both sides in a mutual due diligence process at the earliest planning stages to identify issues that could impact implementation time, delivery capabilities, and cost. Furthermore, it is important that the IT teams continuously work together and collaborate on implementation; it should not be the responsibility of just one team.

The importance of treating the implementation of new leadership assessment programs as an organization change initiative cannot be overstated. Implementing a new process of this nature without the necessary senior leadership support, organizational buy-in thorough communication, two-way feedback mechanisms, and other best practice change management steps will all but doom the initiative to failure.

Finally, in introducing technology-enhanced approaches in place of more traditional and well-accepted assessment methods, a gradual approach may reduce resistance to change. Here, the assessment exercises were developed for a virtual technology platform, but initially delivered in a single physical location, with both participants and assessors on-site. In this initial stage, then, only some of the benefits of moving to a technology platform were realized. Once Darden developed trust in the new assessment processes, the Talent Assessment team was able to move to a process whereby the assessors interacted with the participant remotely, on a virtual basis, yielding much greater utility from the technology-enhanced design. Organizations should consider such iterative or phased approaches when introducing new assessments in the face of potential resistance to change.

Reference

- International Task Force on Assessment Center Guidelines. (2009). *International Journal of Selection and Assessment*, 17(3), 243–253.

Chapter Nine

BRIDGING THE DIGITAL DIVIDE ACROSS A GLOBAL BUSINESS

Development of a Technology-Enabled
Selection System for Low-Literacy
Applicants

Adam Malamut, David L. Van Rooy,
and Victoria A. Davis

The Changing Role of Human Resources

In 2004, Marriott International, Inc., embarked on a strategic human resources transformation initiative (HRT). The objective of HRT is to create greater efficiency and standardization of “transactional” HR activities at our hotels (for example, benefits enrollment, payroll and compensation processing, recruiting and applicant processing) and enable HR professionals to shift their time and focus to more strategic and “transformational” activities that have direct impact on the performance of the business (for example, training, performance management, leadership development).

At the foundation of this organizational change was a large-scale HR outsourcing (HRO) partnership with a third-party

provider. Responsibility for many of the transactional activities shifted from in-house HR professionals to the provider and its large network of service centers and/or became technology-enabled to be more efficiently managed by HR staff at the hotels. A key aspect of HRT was to use a self-service model to transfer many responsibilities to the line managers themselves.

With improved HR efficiencies provided by HRO service center support and technology, the role of the HR staff shifted to more transformational activities such as serving as the HRT initiative change management lead, training and development, and perhaps most importantly, serving as the expert HR technologist to line managers. Line managers were now responsible, with assistance from new self-service HR technology, for many of the transactional personnel activities that were previously managed solely by HR staff (for example, compensation processing, performance appraisal routing and completion, applicant/new hire processing). Consequently, in order for HRT to be well-received within the business, the self-service technology had to be well-designed and easy to use in order to prevent disruption to the day-to-day work of the line managers who lead the operations of the hotels. Moreover, HR personnel had to become technology experts to make sure line managers quickly adopted and became proficient with the new tools.

The Hourly (Non-Management) Staffing Challenge

One of the most dramatic examples of an HRT program and change facing hotels was the transition of non-management (hereafter referred to as “hourly”) staffing responsibilities to line managers and the introduction of a technology-enhanced talent acquisition system. Hourly jobs (for example, housekeeping, front desk agents, and maintenance workers) represent the backbone of hotel operations and 85 percent of Marriott’s 200,000-person workforce around the globe.

Hourly talent acquisition is considered one of the pillars (*Right People in Right Jobs*) of Marriott’s service strategy, and for over eighty years Marriott’s service culture has been the company’s greatest competitive advantage. However, there are a number of

challenges associated with staffing the hourly workforce at hotels around the world:

- *Rigorous staffing process:* Marriott uses a multi-phased process (recruiting and sourcing, application completion, evaluation, hiring, and record keeping) to broaden a qualified applicant talent pool and process high volumes of applicants efficiently.
- *Employee turnover:* Hourly jobs in the hospitality industry tend to have high turnover relative to other industries. Even though Marriott's turnover is low relative to industry benchmarks, it remains a focus of HR.
- *War for talent:* It is difficult to find and retain talent that can deliver on service standards across the globe. The scope of sourcing must be far-reaching to find talent, and employment offerings must continue to set the company apart from competitors in order to obtain and retain key talent.
- *Massive volume of open jobs and applicants:* Marriott processes approximately two million hourly applications per year across almost seventy countries.
- *Spotlight on consistency, information protection, and record keeping:* Hourly staffing processes and record keeping are under growing scrutiny by government agencies and labor/work councils in the United States and abroad.
- *Literacy challenges:* Hourly applicant populations tend to be of lower socioeconomic status and varied in education and primary language. Many have limited familiarity with computers and a considerable number are illiterate in their native language. Marriott's incumbent population speaks over forty-five languages, and illiteracy rates are likely comparable to worldwide estimates of 20 percent.
- *Sustainability of company culture and service standards:* Marriott's *Spirit to Serve* culture and *Guarantee of Fair Treatment* policy to all employees are key competitive advantages and require the company to ensure the talent acquisition process is fair and consistent across all hotels around the world.

Addressing these complex challenges through a labor-intensive staffing solution was counter to the objectives of the HRT initiative and the changing role of the human resources function. Moreover,

the hourly staffing program had to be scalable and sustainable to meet the needs of a growing global business. From 2010 to 2013, Marriott is expected to open more than 250 new hotels, which equates to approximately 40,000 new hourly jobs and 1.4 million new applicants above the current two million applicants the company receives on an annual basis (forecast subject to change based on evolving business conditions). The company determined that an efficient technology-enhanced staffing solution was the right investment, and necessary to secure line manager buy-in to HRT and achieve the company's talent and service objectives, desired staffing process efficiencies, and risk-mitigation objectives.

Technology-Enhanced Staffing Solution

To address the company's global hourly staffing challenges, Marriott developed and implemented a fully integrated and technology-enabled hourly staffing system. The system, referred to as "Hourly eHiring," is comprised of four critical elements: an online Applicant Tracking System (ATS), Web-Based Assessments, Structured Interviews, and Fully Integrated HR Systems.

Applicant Tracking System (ATS)

Marriott, in partnership with a third-party IT provider, developed a system containing the following integrated components: job posting system for managers to post jobs to various career portals, online application blank, and a manager desktop application to track the status of applicants and process hires. The posting system allows hiring managers to quickly post open jobs to a wide array of applicant sourcing sites while allowing applicants to quickly search the posted jobs. After finding a job of interest, applicants can apply immediately by clicking "apply" next to the open position; this action links the applicants back to the ATS, where they complete an online application blank and assessment.

Marriott also recognizes that not all applicants, particularly those seeking lower-skilled hourly jobs, have access to or familiarity with computers to search for jobs online. Therefore, the company assists in sourcing applicants by advertising portals to our

open jobs through partnerships with community-based organizations (where many applicants go to apply online). If accommodations are needed, applicants are instructed to (or may do so on their own volition) come into hotels, where they are brought to a private assessment room and are provided assistance in logging on to the personal computer or kiosk to complete the application. When opening hotels, Marriott also conducts job fairs during which applicants are funneled to conference rooms with computer terminals.

Once the application blank and assessment are completed, the applicant's data is instantly sent to the tracking and hiring module of the system, which can be accessed by the hiring manager from his or her computer desktop. At this point the hiring manager can view the applicant's profile and decide whether to proceed to an interview or remove the person from consideration.

In sum, the ATS-enabled process has several key benefits:

- Enables hiring managers to quickly increase the breadth and flow of applicants by posting jobs to a variety of career sites on the web. In fact, at the onset of the ATS implementation, hotels in the U.S. averaged a 150 percent applicant flow increase.
- Reduces paperwork and data processing by managers, since the application process is completed online and self-reported by applicants. This was critical to gain buy-in from line managers who now had to manage the staffing process without significant HR involvement.
- Automates applicant record keeping, improving data consistency and compliance.
- Pools applicants to increase access to other hiring managers looking for talent.

Web-Based Assessment

Perhaps the most compelling benefit of the ATS-enabled solution is the substantial increase of applicant flow and the sourcing of potential talent. However, higher applicant flow can increase the burden on hiring managers to sort through applications and find the most qualified applicants to interview. Additionally,

applications for lower-skilled hourly jobs often contain sparse work history information useful for ranking candidates for interview. Given the massive amount of flow expected through this system, relying on an interviewing process alone to rank all minimally qualified candidates would have proved administratively burdensome.

To address this challenge, online employment assessments (specific to job families) were added to the process to further screen and prioritize candidates for face-to-face interviews. During the assessment conceptualization phase, however, HR professionals and hiring managers expressed concern about using a long and complex computer-based application process for job groups with a higher preponderance of computer and language literacy challenges, specifically, *heart-of-house* jobs (lower guest contact jobs such as housekeepers, kitchen workers, and groundskeepers).

Given the business benefits of the eHiring system from an HRT standpoint, the company made a further investment to create online assessments, measure job family-specific competencies, and link the assessments directly into the ATS and online application process. For the purposes of this chapter, we focus on one of the eight hourly assessments (the Heart-of-House assessment) to illustrate technological advances to address computer and language literacy challenges.

Key Features of the Heart-of-House Assessment

Content

- The skill domain captured on the assessment is aligned to the critical competencies gleaned from a robust job analysis study (detail orientation, learning ability, neatness, dependability, and interpersonal skills).
- Three item types are used to measure these competences: applied learning items measure both learning ability and detail orientation (Figure 9.1), biodata items measure the personality dimensions of dependability and interpersonal skills (Figure 9.2), and presentation items measure disposition toward neatness (Figure 9.3).

- Text is made available to applicants in twenty-two languages.
- An interviewer certification program and job-specific behavioral interview guides (available in twenty-two languages) are utilized. Robust interview protocol allows for greater talent evaluation and provides an in-person verification of assessment results.

Literacy Support

- Text is used sparingly and kept at an eighth grade or lower reading level.
- An optional audio feature streams voice-over recordings of the text in all available languages.
- Visual test stimuli (pictures of guestrooms, grounds, kitchens, etc.) minimize the use of text and overall information processing load.
- One assessment item is presented per screen, eliminating the need to scroll and use web page navigation buttons.
- Assessment length is kept as short as possible (90 percent of applicants complete the new assessment in under twenty-five minutes).
- General assessment and web page navigation instructions are provided at the start of the assessment and also made available via audio.

Technology


- Assessment scores apply to broad families of jobs and remain on record for twelve months, thus reducing the need for applicants to retake assessments within a given year.
- Security technology protects assessment content and response integrity: (1) the system randomizes the order in which assessment items are presented to each applicant, (2) the system contains alternate forms for each assessment, which are presented to applicants in random order, (3) time limits exist for completing the assessments, and (4) opportunities to log in and out of the assessment are limited.
- Simplified assessment website navigation and keyboard design, including “one key” navigation and response option selection that reduces need to use the mouse.

Figure 9.1. Example Applied Learning Assessment Item. Applicants read or listen to the voice-over recordings of the instructions on the computer (Screen 1) and proceed to a question on the next page (Screen 2). The system does not allow applicants to return to the instruction screen once they proceed to the questions.

Screen 1

Applied Learning Instructions

Click here to turn audio off | ⏮ ⏪ ⏩ ⏭




This picture shows how things in the hotel room must be arranged.

1. Place three large pillows in front of the headboard.
2. Place three small pillows in front of the large pillows.
3. Place a blanket across the bottom of the bed.
4. Place a glass upside down on top of a round napkin on the right bedside table.

Screen 2

Click here to turn audio off | ⏮ ⏪ ⏩ ⏭



How many mistakes were made?

0

1

2

3

Figure 9.2. Example Biodata Item Measuring Interpersonal Skill. Applicants read or listen to the voice-over recordings and select their preferred response.

Click here to turn audio off | ⏮ ⏪ ⏩ ⏭ Click here to hide timer | Time Remaining: 48:52

38) How often would others say that you made customers happy?

1 More often than most other people

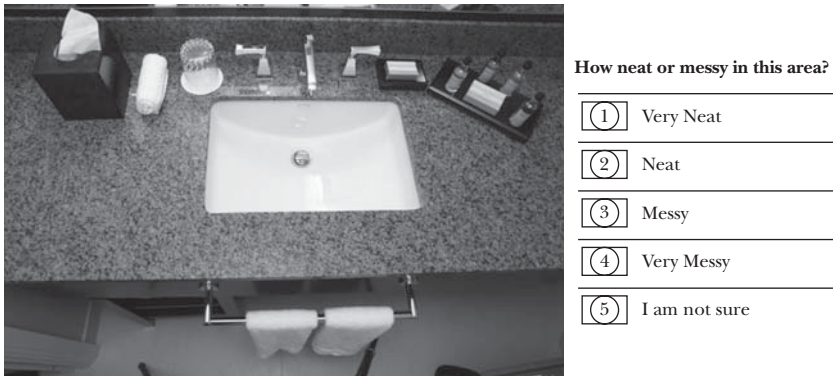
2 About as often as most other people

3 Not quite as often as most other people

4 I have not worked with customers

Working on Question 38 of 45

Figure 9.3. Sample Presentation Item Measuring Disposition Toward Neatness. Applicants read or listen to the voice-over recordings and select their preferred response.



Figures 9.1 through 9.3 illustrate some of the enhancements made to the assessment (for example, multiple item types capturing a broad performance domain, visual stimuli, audio feature, reduced and simplified text).

Assessment Development and Validation

Job Analysis

The foundation of the selection and assessment program started with a large-scale job analysis supported by a third-party vendor. Over one hundred job experts participated in focus groups to draft work activities and competencies for approximately one thousand job titles. This information was used to create online job analyses validation surveys. Job experts (5,610 managers) participated in the survey with an 82 percent completion rate. The data were used to organize the one thousand job titles into a more manageable set of three hundred job classifications based on substantial overlap of work activity and competency importance to the job (as rated by job experts). Eight super job families were also identified by examining competency overlap. Assessments were developed for each of these families.

Assessment Development

Based on the enterprise-wide job analysis, online assessments were developed by Marriott's Assessment Solutions group in concert with a third-party firm with validation expertise. To ensure item validity, hundreds of Marriott in-house subject-matter experts (SMEs) from around the world were utilized. Critical competencies identified during the job analysis became the foci of the assessment scales. Various design steps included SME focus groups (for example, critical incidents and assessment review), piloting of assessment items, as well as cultural adaptation and translation of items from English into twenty-one additional languages.

Cultural Review and Adaptation of Assessment Content

After creating the initial item pool, a representative sample of assessment items, including instructional text, alternate item phrasings, and response options, were provided to SMEs located within the various regions in which Marriott operates. SMEs were selected based on their bilingual skills (that is, target language and English), diverse cultural knowledge, and familiarity with the type of work for which the assessments would support in the selection of new employees. The intent of the cross-cultural review was to explore any potential for cultural irrelevance, differing interpretations, challenging topics, and literacy concerns. SMEs were asked to consider the following as they reviewed items:

- Are the instructions for responding to the item clear?
- Does the item make sense? Does the item work in your local culture/language?
- Do the item response options make sense? Are the response options appropriate for your local culture/language?
- Will the intended meaning of the item and/or response options translate effectively into the target language?
- If the meaning of the item and/or response options will not translate effectively into the target language, what might be the equivalent situation or description in the target culture?
- How might we better phrase the item and/or response options so that the meanings are clear?

Findings revealed that an overwhelming majority of the assessment items could be accurately interpreted across cultures, regions, and languages. Issues addressed as part of item revisions related to:

- Use of colloquialisms (for example, How often have you made things happen that you believed in, no matter what the odds?).
- Terms that when translated could alter the level of interpretation (for example, “argument” was changed to “disagreement” because “argument” can be interpreted more strongly, and in some cultures violently, than actually intended by the assessment item).
- Familiarity of terms in relation to cultural customs (for example, reference to using a “teddy bear” as a gift as opposed to simply a “stuffed animal”).

As a result, assessment items were updated and piloted to ensure cultural neutrality, job relatedness, and appropriateness of reading level (lowered as much as possible) before entering into translation.

Translation and Audio Production of Assessment Content

Throughout the process, Marriott partnered with a third-party global language firm with expertise in translation and localization efforts. Marriott’s Assessment Solutions group collaborated with the firm’s project management, desktop publishing, and technology teams to coordinate in-country translators and centralized quality assurance linguistic leads on the firm’s side.

After items were translated and reviewed by the firm’s linguistic leads, an in-depth quality review of all instructional text and assessment items was again performed by over two hundred Marriott SMEs to ensure accurate translation, minimize preferential phrasings, and identify any potential challenges. Item corrections and updates were then submitted to the firm’s translation team for incorporation into the text.

Based on a need to have fully audio-enabled assessments, the next step in the process was production of the audio files. Narrators were initially auditioned and narrowed down by the firm, who then provided Marriott with a few top voice samples for

each of the twenty-two languages. A sample of Marriott SMEs then selected the eventual narrator in each of the twenty-two languages. Considerations included clarity, tone, speed, dialect, and cultural norms (for example, gender preference in some cultures).

Last, translated text and audio files were loaded into the online assessment system and the same group of in-country, on-property translation SMEs logged into the system as “applicants.” They then performed an “in-context” quality assurance review to make certain that not only were the text and audio accurate, but that the correct character structure (for example, application of Chinese Simplified versus Chinese Traditional in the appropriate text fields) and formatting (for example, right to left justification for Arabic versus left to right for English) for the screens were implemented.

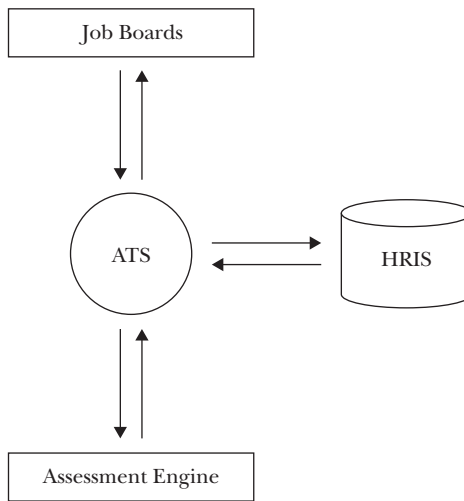
Criterion Validation

A concurrent criterion-validation study was performed for each of the assessments. A representative sample of 882 managers of the incumbent population was asked to participate in the study. Managers were asked to participate in a rater training session (designed to improve rating accuracy) and rate a sample of their subordinates utilizing an online performance evaluation tool (820 complied for a 93 percent participation rate). The content of the performance evaluation was derived based on the results of the job analysis. The managers rated 3,488 incumbents, who were subsequently invited to complete the validation version of the assessments. Incumbents at the hotels were organized by local HR staff to complete the assessments on a computer located in the HR or training office. Incumbents completed two versions of the assessment over a one-hour period in order to validate alternate forms of each (3,170 incumbents completed the assessments for a 91 percent participation rate). All of the assessments yielded validity coefficients of .30 or greater (uncorrected).

Fully Integrated HR Systems

To achieve the desired staffing process efficiencies, it was critical to create interfaces between the component systems of the overarching Hourly eHiring System (Figure 9.4).

Figure 9.4. Diagram of Hourly eHiring System Components Integration



Marriott utilizes an applicant tracking system (ATS) via an application service provider arrangement with a third-party HR technology company. The ATS is configured to have a bidirectional interface via a web service (a software system designed to support interoperable machine-to-machine interaction over a network) to the online assessment engine of another third-party vendor. Data between these systems are seamlessly exchanged. For example, the ATS passes applicant data captured via the application blank (for example, applicant ID, job ID, etc.) to the assessment vendor in order to determine which assessment is appropriate for the job and whether the applicant has completed that assessment within the past twelve months. If the applicant has completed the assessment within the past twelve months, the assessment score on record passes back to the ATS, and the hiring manager's computer desktop application tracks the status of applicants and processes hires. If an assessment score is not on record, the system connects the applicant to the appropriate assessment for the job in which he or she is applying. All of this happens instantaneously to ensure a smooth applicant experience.

The ATS, which contains Marriott's career website and descriptions of open jobs, also connects to a multitude of public job boards. This configuration enables hiring managers to quickly post open jobs to a wide array of sourcing websites to expand the potential talent pool. The manager simply logs into the system, which automatically recognizes the manager and his or her hotel based on the ID entered at login using single sign-on functionality. The hiring manager completes a quick online form about the job (for example, job code and title, location, pay, business context of hotel, etc.) and the ATS instantaneously uploads a predefined job summary and posts a link on the Marriott career site and public job boards. The applicant scanning the job board clicks on the link and is immediately brought to the ATS to read the summary and apply for the job.

The ATS was also configured to have a bidirectional interface with the company's human resources information system (HRIS). This interface automatically pulls the hiring manager's ID and hotel information from the HRIS database to facilitate the single sign-on functionality. Single sign-on functionality is also utilized for internal applicants; that is, internal applicants apply to jobs via an internal career site and are automatically recognized after entering their IDs. Once an applicant enters his or her ID, it is sent to the HRIS to pull employee information over to the application blank on the ATS. The ATS-HRIS interface also exports applicant data (from the application blank and assessment) to the HRIS database for seamless and automatic record keeping, reporting and hiring purposes.

System Implementation Process and Challenges

The culmination of three years of work—from job analysis, assessment validation, systems design, and training—was the implementation of the Global Hourly eHiring System to approximately 1,400 properties, located in nearly seventy countries, using twenty-two languages, and affecting all businesses/brands from the luxury tier to select service brands and Marriott's timeshare business. The process of planning and implementing this system required the coordination of a complex network of internal stakeholders

who had to learn about and become proficient on the new system, external providers who had to help build the system, market operations leadership who had to champion the strategy, and internal information technology specialists who had to make sure the system was developed to work within the company's IT architecture and security protocol.

Key Stakeholders

The needs of many key stakeholders had to be met and consistently exceeded in order to achieve buy-in to the new system and ensure initiative success:

- *Business Sponsors:* Senior executives expected the system to achieve desired business benefits (improved operational efficiency and hourly talent to deliver guest service) and to be implemented on time and within budget.
- *Property General Managers and Market Operations Leaders:* These individuals lead a hotel or the regions/markets in which hotels reside. They required the system to achieve the communicated business benefits, but also required that the implementation have minimal disruption on the day-to-day operations of the hotels.
- *Property-Based HR Professionals:* HR professionals expected the system to deliver better talent and be easy enough to teach hiring managers how to conduct hourly staffing on their own.
- *Hiring Managers:* Hiring managers expected the system to streamline the staffing process, deliver better talent, and be easy enough to execute hourly staffing on their own.
- *Hourly Applicants:* The applicants deserved a consistent, relevant, and professional staffing process that they could complete in an efficient and user-friendly manner.

Centers of Expertise (COE)

The project team included more than fifty members from various COEs. Successful delivery of the Hourly eHiring System (on time, on budget, and with acceptance) required the efficient orchestration of multiple COEs:

- *HR Program Management Office (PMO)*: The HR PMO was the lynchpin for project team coordination. Key responsibilities included establishing the overarching project plan, project management and coordination of all team members (internal and external) against delivery dates, and third-party vendor contract management.
- *Talent Acquisition*: The talent acquisition department was responsible for overarching sourcing and recruiting strategies, ensuring staffing systems were aligned to the strategy and that internal and external recruiters and hiring managers were utilizing the systems appropriately to meet the talent acquisition needs of the company.
- *Assessment Solutions*: This group was responsible for the global assessment strategy, job analysis, and assessment development and validation. Additionally, the assessment team, in partnership with a third-party firm with validation expertise, was responsible for ongoing evaluation of assessment validity, fairness, utility analyses, and subsequent assessment reconfiguration based on analyses.
- *HR Communication and Change Management*: This group was a vital member given the revolutionary change the system brought to the day-to-day operations of our hotels and the imperative to make all stakeholders committed believers in the effort. Responsibilities for this COE included development of the communication strategy, which included a multitude of scheduled messages to all stakeholders explaining the purpose of the initiative, the business case, and the WIFM (What's in it for me?). These messages came in a variety of formats (for example, in-person market leadership meetings, interactive webcasts, newsletters, video communications to all employees featuring senior executives explaining the business case, on-property general manager and HR leader "stand-up" meetings, etc.). Additionally, this COE was responsible for creating training associated with this significant change effort. Training programs on how to use the new system were developed in a variety of formats (interactive webcasts, audio streaming, interactive PowerPoint decks, short job aids explaining process steps utilizing screen grabs from the system, etc.).

- *HR Systems and Information Technology:* This group was responsible for ensuring the Hourly eHiring System was designed in a manner consistent with the process steps outlined in the talent acquisition and selection strategies, ensuring performance and usability of the systems for hiring managers and applicants, leading all systems testing before go-live, and ensuring all system components adhered to the company's IT architecture and data security protocol.
- *Third-Party Vendors:* Four vendors supported this effort. The ATS provider was responsible for ensuring that the content in the ATS was designed to business specifications and interfaced appropriately to the assessment provider and HRIS. The assessment provider was responsible for ensuring assessments were appropriately developed and validated, programmed into the assessment engine, and interfaced appropriately to the ATS. The translation vendor was responsible for accurately translating text and ensuring that translations were properly displayed in both online and print formats. Finally, the HR outsourcer provided oversight and support to ensure the ATS was properly interfaced to the HRIS and data warehouse (and subsequent reporting vehicles), which they manage as part of the agreement.

Organizational Challenges

A number of organizational challenges threatened this project from getting off the ground and sustaining flight once launched. First, it was challenging for HR leaders to convince key business executives, hotel operations leaders, and owners to make such a large investment in HR-related technology, particularly since the company resides in a moderate margin industry. To convince this group required positioning the business case as an efficient, customer service, and risk-mitigation initiative. Multiple forums were used to convey the business case. It started with the executive vice president of HR conveying the strategy directly to the CEO and COO and gaining support for further pursuit. This opened the way for gaining input and buy-in from leaders at the market hotel operations level; this group was critical, as they would be charged with selling the strategy to owners and

embedding the program into day-to-day operations at the hotels. Local leadership meetings were held in each hotel operating market. In these meetings the senior HR leader of that region presented the business case and answered questions from the local market vice presidents and hotel general managers. Finally, to gain hotel owner support for the investment, senior HR leaders had to attend and present the business case at an owner and franchisee financial committee meeting that weighs investments supported by hotel owner funds.

Another challenge at the onset of the project was due to the organizational structure. At the time, several of Marriott's larger brands and regions operated as distinct business units with independent P&Ls, strategies, and decision-making bodies. The business case was dependent on the positive economies of scale associated with all businesses sharing the development and maintenance costs. The initiative required reigning in all business units as a single strategic team. Moreover, some of these businesses had distinct employee selection procedures. This also required a clear and compelling business case and expected benefits to convince leaders of the value of moving to a common global solution, regardless of brand or business.

Equally difficult was convincing hotel operations leaders that applicants would be willing and able to complete an online application and assessment process and that line managers would not have to spend an inordinate amount of time off the floor and away from guests to use the system.

Another challenge emerged in the middle of system development. The economic recession hit the business. It took a lot of effort and courage on everyone's part to keep all stakeholders engaged and supportive of the initiative as the company faced new financial challenges.

Finally, development of the system required coordination among multiple external vendors (who were also competitors of each other) to partner as one team.

In the end, the business case was compelling enough to gain support from all stakeholders. The project team developed a sound research-based illustration of the initiative's benefits: cost savings associated with a shared service model, validity of assessments (in terms of predicting job performance and customer

service), the accessibility of the assessments to the applicant base, a profitable partnership to all external vendors, and staffing process efficiencies facilitated by technology. Sustained focus and buy-in on the initiative during development and after implementation can be largely attributed to a highly effective change management strategy.

Technological Challenges

As can be expected, a number of technological challenges had to be overcome to meet the strategic objectives of the program.

- *Assessment Formatting:* Assessment display (text and stimuli/images) had to be designed with simplistic usability across the many languages, which included multiple alphanumeric and character-based languages.
- *Flexible and Secure Assessment Engine:* Given the expected volume of online assessments around the globe (two million applicants annually), the system needed to be designed to secure assessment content. The goal was to design the assessment engine to randomly present different assessment items and versions of the same assessment; present audio and visual displays across various monitor, browser, and operating system versions; and create time limits for completing and accessing the online assessments to significantly reduce the possibility of applicants coming onto the system and “practicing” assessments before actual submission.
- *Global System Performance:* The high volume of applicants and multiple staffing hurdles to be completed by applicants (application and assessments) required the system performance to be fast and not encumber the applicant. Ensuring system performance was particularly complex because the system had to work across a seemingly limitless number of potential computers (internal and external to the company), operating system versions, browsers, and bandwidth that would be used by applicants around the world.
- *System Component Integration:* The overarching Hourly eHiring System required building automated and bidirectional interfaces across three systems (ATS, assessment engine, and HRIS)

managed by four partners (Marriott, assessment vendor, ATS vendor, HR outsourcer).

- *Global SME Coordination:* Developing assessments to work online in twenty-two languages required coordinating hundreds of subject-matter expert employees from around the world who were fluent in one or more of the languages in order to review translation accuracy and then participate in quality assurance (user acceptance testing [UAT]) of the system. UAT requires participants to follow a time-consuming scripted review process.

These technological challenges were successfully addressed in large part by the efficient collaboration of the different COEs and the great partnership and mutual respect that was formed between Marriott's HR professionals and IT experts (internal and external). We believe success in building partnerships across technical disciplines (HR and IT) has a lot to do with the personal attributes of the project team members. Clearly, members on a project of this complexity must possess the requisite technical skills to develop their respective areas of the talent management system. However, of equal importance are the softer skills required to build trusting relationships across members of the extended team. Marriott and its partnering organizations were very selective in determining project team leadership and membership. These individuals had to have a track record of success in managing complex multiple-disciplinary projects and a reputation for building strong rapport with co-workers and clients. Success would require our project members to work effectively outside of their respective technical areas. Technical leaders who are "well-rounded" (inquisitive and want to learn outside of their professional disciplines, open-minded to the perspectives of others, flexible and hardworking) are best suited to solve problems efficiently and build the trust necessary to create a nimble and high-performing cross-disciplinary team.

Success Measurement

The primary objectives for developing the Hourly eHiring System were to (1) ensure better prediction of talent able to deliver on performance and service objectives, (2) ensure the assessments

were accessible by all job applicants, (3) establish consistent talent selection standards around the globe, (4) mitigate risk through standardized and validated staffing processes and data reporting, and (5) create staffing process efficiencies for hiring managers.

Indicators of Success for Enhanced Assessments

The specific objectives for the system were to improve the capability of the assessments to predict performance (that is, improve validity) and ensure the assessments were accessible to applicants, regardless of their level of computer and language literacy. Validity was established before implementation of the system and followed legal and professional guidelines. A series of concurrent criterion-related validity studies were completed. All of the assessments yielded validity coefficients of .30 or greater (uncorrected).

While these results were impressive enough to go live with the system, the project team was anxious to learn how the enhancements to the heart-of-house assessment would play out in the actual applicant population. Specifically, did the literacy enhancements improve the ability of literacy-challenged applicants to complete the staffing process?

To answer this question, the project team completed two studies after the new system was in place for three months in the United States. The first study examined the heart-of-house assessment score equivalence across applicants that used (and presumably needed) and did not use (and presumably did not need) the streaming audio feature. All available applicant data in this job group ($N = 64,096$) were analyzed. Results of the analysis (Figure 9.5) illustrate assessment score equivalence across applicants that needed the literacy support (audio on) and those that did not (audio off). Table 9.1 also shows that mean differences in assessment scores were only 1 point and not practically significant ($d = .16$). These findings suggest that the literacy support aids provide test comparability regardless of applicant language and computer proficiency.

A second study was conducted to gain feedback about the system from users in the market. Specifically, the lead HR professionals at each full-service hotel in the market were surveyed. These

Figure 9.5. Mean Assessment Scores, by Use of Audio Feature, for U.S. Applicants. Applicants took the assessment within the three-month period after implementation of the new hiring system. Points represent the mean assessment scores for those applicants who did and did not use audio.

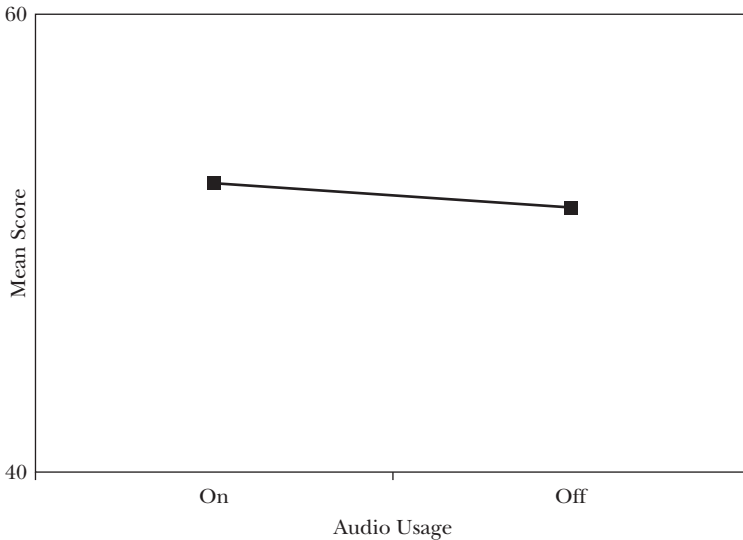


Table 9.1. Mean Differences Between Applicants Who Used Audio While Taking an Assessment and Those Who Did Not Use Audio

<i>Audio On</i>		<i>Audio Off</i>		df	t	d
M	SD	M	SD			
52.59	8.89	51.51	9.41	26,788.65	12.83*	.16

Note. Levine’s test indicated equal variances could not be assumed; thus a more stringent degrees of freedom was used.

* $p < .01$.

individuals are the stewards of the company’s employee selection programs on property and have the opportunity to observe and converse with applicants and hiring managers using the system. A total of two hundred people were surveyed and 116 replied (response rate = 58 percent). Results indicated very favorable

Table 9.2. Mean Survey Responses from Human Resource Professionals Evaluating Global Selection Program Enhancements

<i>Item</i>	<i>M</i>	<i>SD</i>	<i>n</i>
The website layout and navigation features make it easier for hourly/non-management “heart of house” job seekers to apply online.	4.79	.83	113
The new audio technology makes it easier for hourly/non-management job seekers to complete the heart of house assessment online.	4.97	.71	110
The new picture-based questions make it easier for hourly/non-management job seekers to complete the heart of house assessments online.	5.01	.78	110
The new system enables HR and hiring managers to more accurately identify the most qualified applicants to move to the interview phase.	4.85	.91	116

Note. Response scale ranged from 1 (Strongly Disagree) to 6 (Strongly Agree).

perceptions of the system (Table 9.2). For example, a strong majority agreed that the literacy aids make it easier for applicants to complete assessments. Moreover, the majority also agreed that the new system is identifying better talent.

Indicators of Success for Staffing Process Efficiency

In addition to more quantifiable indicators of success discussed above (for example, validity, accessibility to all applicants), the Hourly eHiring System has lived up to the vision in terms of delivering improved process efficiencies for hourly staffing. The system has allowed the company to efficiently transfer staffing responsibilities away for HR staff to hiring managers, allowing the company to keep HR professionals focused on transformational activities. Moreover, the fully integrated nature of the system has facilitated a consistent and standardized hourly staffing process, which enables

the company to maintain talent standards and comply with legal requirements across the globe. Finally, the system has ensured that applicant data at each decision point (screening, assessment, interview, and hiring) are stored in a centralized database for accurate and relatively easy reporting, analysis, and auditing.

Lessons Learned

An entire chapter could be dedicated to the lessons learned at every level of Marriott's Hourly eHiring Initiative. Keeping the focus of the strategic practitioner in mind, we conclude with several "pearls of wisdom" for readers about to embark on a large-scale HR technology initiative. The following areas were found to be the most instrumental to successful implementation of the global Hourly eHiring System.

First and foremost was the importance of building a compelling business case linked to business outcomes valued by non-HR leaders and stakeholders. Leadership buy-in and investment support were garnered by articulating (1) the cost savings to our hotels to be gained over time by using this system and (2) the clear connection to the overarching strategy of the business (customer service). Without executive and operations leader buy-in and effective training of system users, failure is almost guaranteed.

Related to the business case development is change management. The initiative was part of a transformational change in the way the company operated. Change management is truly a professional discipline and a necessary competence for the modern HR organization. A dedicated change management team with a robust communication and training strategy will ensure that all stakeholders understand the business case and are prepared to execute and sustain success, even amid initiative challenges that will most definitely occur.

Similar to change management, professional project management is a discipline that should not be underappreciated or underutilized in complex system development and implementation initiatives. The initiative involved hundreds of project team members from varying COEs, external vendors, and market HR professionals. Our dedicated project management COE, with certified project managers, kept these parties on

task and functioning as a well-oiled machine, ensuring delivery dates were met and on budget.

Most organizations will need external support of some kind when developing and executing a technology-based HR solution. It is paramount to choose your vendor partners wisely when embarking on such a strategy. Not all HR consulting groups are alike. For example, not all assessment firms (despite proven assessment development experience and competence) have the scalability and IT systems know-how. Modern HR practice and organizations have evolved to be inextricably linked with IT; therefore, make sure your assessment vendor has experienced internal IT staff. Similarly, not all HR systems vendors are created equally. Make sure your systems vendor has experience and knowledge in the nuances of HR practice. Many providers have developed systems that do not enable companies to easily follow the strict employment testing regulations and best practices. Involve your in-house IT experts in the vendor (systems and assessment technology) review process—this will also create buy-in from IT partners for the overarching strategy.

Finally, technological solutions, particularly those involving multiple interfaces and text presented in multiple languages, are complex to build and therefore fraught with configuration errors during the build phase. Do not underestimate the investment that must be made in systems development quality assurance processes—SIT (systems integrity testing) and UAT (user acceptance testing). The Hourly eHiring team followed a very sophisticated testing process involving all stakeholders at different testing phases (for example, internal and external IT professionals led SIT testing, HR and hiring managers participated in UAT, etc.). The process was very structured, involving hundreds of people, utilizing predetermined testing scripts and issue-reporting processes, and simultaneous systems fixes and regression testing (making on-the-fly system fixes based on UAT feedback while ensuring fixes did not break other system components) by IT professionals. Therefore, try very hard to break your systems and do not leave any stone unturned before launching to your business. Involving your IT experts in this effort (or having them lead if possible) is wise for quality system development, commitment from your IT leadership, and ensuring your IT function will be

engaged and committed to fixing technological issues that are discovered after the system implementation.

We conclude with the position that realization of the business value of a global talent management program depends on the operational validity of the program—the commitment to a standard and consistent process by users. Sporadic use of assessment tools will do very little for a business; it is the strict adherence to their consistent use that will ensure that desired talent performance standards are reinforced around the globe. We believe that embedding assessment tools and processes within a user-friendly and integrated technological platform can be a highly effective solution to the challenge of global operational validity.

Chapter Ten

PROMOTIONAL ASSESSMENT AT THE FBI

How the Search for a High-Tech Solution Led to a High-Fidelity Low-Tech Simulation

Amy D. Grubb

In this chapter, we describe the journey that brought the Federal Bureau of Investigation (FBI) to design and implement the high-fidelity, remotely delivered, externally administered simulation used today as a promotional assessment for mid-level positions. We will describe how the FBI came to overhaul its promotional process and adopt a virtual assessment approach, following the imposition of a court-issued consent decree and under the cloud of 9/11. The impact of that event on the mission of the Bureau led to changes in the role of mid-level managers that were reflected in the design of the promotional assessment developed and implemented in the years following September 2001.

The FBI

The FBI is charged with protecting the national security of the United States through preventing terrorist acts, protecting against espionage and cyber-based attacks and high technology

crime; combating major economic crime, transnational crime, and organizational and public corruption; protecting civil rights; combating major violent crime; and supporting local, state, national, and international law enforcement and intelligence partners. The FBI has over 33,000 employees working in one of the divisions at FBI Headquarters in Washington, D. C., in one of the field offices across the United States, or in an international location. Approximately one-third of all FBI employees are “Special Agents.” These special agents are sworn law enforcement personnel and trained investigators in a multitude of criminal and national security specialties.

For special agents pursuing a management career track, there are two levels of middle management positions available. The lower positions are classified in the Federal Government’s system as GS-14, and the higher as GS-15. For example, in the field offices, the two levels of middle management positions are the first-line squad supervisor (GS-14) and the assistant special agent in charge (GS-15) who manages the squad supervisors.

The Impetus for Change

In the 1990s, a lawsuit was brought against the FBI by a group of African American special agents claiming discrimination in the process used to promote special agents to mid-level management positions. After several years of litigation, the suit eventually led to a consent decree requiring a complete redesign of the promotion process. The consent decree dictated that the new promotion process be designed to incorporate professionally validated assessment procedures that were blind to race and gender and contained an “automated” component, although no specific parameters for that automation were prescribed. Further, the consent decree mandated that a committee of industrial/organizational psychologists external to the FBI review and attest to the new procedures’ technical adequacy. The decree stipulated that all future promotions for special agent mid-management positions at these levels would be competed through the new process. Beyond these court-mandated changes, administratively, the new system would have to be designed to handle a potential pool of

up to twelve thousand special agents and more than fifteen hundred mid-level management openings each year.

In the aftermath of the 9/11 terror attacks, the FBI increased its emphasis on better integrating the dual responsibilities for law enforcement and for intelligence/national security protection at every level in the organization. This in turn changed the way FBI executives thought about leadership roles. These changes in the middle-manager role dictated that the new promotional assessment would need to:

- Emphasize the importance of management and leadership competencies, not just specialized technical skills or past performance as a special agent, as had been historically the case, and reflect a better balance between these two components
- Provide rich assessment information to stimulate and direct managerial skill development on the individual level and foster a learning culture on the organizational level
- Decrease the subjectivity and increase the perceived fairness of the process to all stakeholders, but especially to the special agents
- Assure that the new promotional process was not only valid, fair, and reliable but that the process itself had face validity, that is, it “looked like” the supervisor job at the FBI for all stakeholders
- Be easy to administer, with minimal disruption to the execution of the FBI’s core and urgent post-9/11 mission and minimization of the time that special agents are taken off their duties

The Promotion Process

The redesigned FBI promotion process uses the new assessment as a threshold screen to create a pool of promotion-qualified applicants. Upon passing the assessment, special agents are eligible to apply for any specific open middle management position, based on their personal interests and based on other job-relevant qualifications (for example, computer intrusion investigative experience, human intelligence collection, public corruption, or overseas experience with an associated language proficiency). The qualifications of each position are determined by the hiring official who

is responsible for that opening. Hiring officials may select up to seven competencies and qualifications on which to evaluate the candidate. The four competencies seen as the most important requirements must be designated from a list of eight core leadership competencies identified through extensive job analysis as being instrumental to leadership success at all levels at the FBI:

- Leadership
- Interpersonal Ability
- Liaison
- Prioritizing, Planning, and Organizing
- Problem Solving
- Flexibility/Adaptability
- Initiative/Motivation
- Communication

Up to three additional competencies can be selected by the hiring official. These target competencies may either be selected from the leadership competency list or be technical competencies.

To apply for a position, a candidate submits a written description of his or her accomplishments and experiences relevant to each of the competencies determined by the hiring official to be essential for that opening. The written description can be no more than eleven lines of text describing a verifiable example (including the situation, actions, and results) when the candidate demonstrated that competency. Every competency example is subject to verification by an individual who supervised the candidate at the time. The applicant's submission is rated by a local hiring panel using behaviorally anchored rating scales developed by executives at the FBI. The local panel may also conduct an interview and/or use other, pre-approved and standardized assessment instruments. All members of the local panel are trained and calibrated in the application of those rating scales. A Division Head Recommendation Form is also considered in the promotion, but only when the division head does *not* recommend the individual for the position. A senior board then makes the final selection decision from the slate of qualified applicants, considering both the local panel's and the hiring official's recommendations.

In Search of an Automated Assessment

To fulfill the requirement of the consent decree for an “automated” component to the promotion process, the FBI initially designed and intended to implement an interactive computerized assessment. This high-tech solution met many of the FBI’s design objectives:

- Completely blind to race and gender
- Standardized and objective scoring
- Minimal burden on administrators
- Ability to test in multiple locations simultaneously at high volume
- Timely (in fact, immediate) feedback of assessment outcome

This fully automated approach ran into strong opposition both at Headquarters and in the field. Applicants in a pilot group found the assessment process very artificial; the situational judgment items, although content validated as job relevant, did not “feel” like real life. Leaders and applicants did not understand the complex algorithms used to score the assessment. In addition, the technology platform for which the assessment was designed was too far ahead of the technology infrastructure in place at the FBI. Indeed, the assessment’s face validity suffered from this technology gap; applicants were taking a sophisticated, fully automated, technology-delivered assessment while, at that point in time, special agents generally did not even have computers on their desks.

Given the opposition, the FBI abandoned the fully automated assessment and sought an alternative solution that would have stronger credibility with the key stakeholder groups and yet meet its design requirements for administrative ease, fairness, and validity as well as fulfill the requirements of the consent decree.

The Winning Solution

The solution the FBI finally settled on was a customized live, role play telephone assessment delivered by a third-party partner that specializes in the design and delivery of simulation-based

assessment. Separate programs, labeled leadership skills assessments (LSAs), were developed for the GS-14 and GS-15 levels of middle management. Each LSA program consists of multiple simulation exercises, with the exercises created using critical incidents collected during the development phase of program design to capture a realistic, highly credible “day in the life” assessment experience. Each candidate is assessed by a team of at least four or five professional assessors using evaluation guidelines developed and validated by FBI subject-matter experts. These assessors interact with the candidate telephonically. They have no access to any background information on the candidate and are fully focused on providing as fair and accurate an evaluation of the candidate’s skill as possible.

The assessment measures each candidate on the eight core FBI leadership competencies. Although the scenarios are embedded within an FBI backdrop, no technical knowledge is measured. The assessment focuses exclusively on the target leadership competencies. Each of these competencies is measured multiple times across the six exercises or scenarios comprising each of the two-and-a-half-hour programs. Initially, two parallel forms were created for each program. Ultimately, two additional parallel forms were created so that in all each LSA has four forms.

The FBI can assess as many as five hundred candidates a week on the LSA with candidates situated at any one of fifty-six field offices and international locations. Assessments are conducted daily, with over 7,500 administered to date. Special agents who are promotion-eligible are scheduled by an FBI coordinator two to four weeks in advance to report for assessment to a nearby site, minimizing the agent’s time out of the field. The site is staffed by FBI personnel to check identification and provide a special identification code; the third-party vendor is privy only to this identification code and is given no information about the candidate, not even the candidate’s name.

At the start of the assessment, candidates receive a sealed, hard-copy packet of background materials to review prior to, and between, the role-play exercises. Hard-copy packets rather than online materials are used to neutralize differences in Internet access across the FBI footprint; this design feature may change in future years as web access becomes universal at the agency at the

necessary bandwidth. Each packet contains a description of the candidate's duties in the target (GS-14 or GS-15) leadership role, an organization chart, a description of the personnel in the hypothetical squad or unit the candidate is managing, a schedule of calls (at least one of the role plays is an additional unscheduled "surprise call"), reports, memos, emails, and other supporting materials. Upon completion of an initial forty-five-minute review period, the first call comes in and the "day" begins.

In the GS-14 version of LSA, for example, the candidate plays a squad supervisor in a fictitious town. The supervisor was promoted to that position recently and is one of a number of supervisory special agents (SSAs) reporting to an assistant special agent in charge (ASAC), who in turn is responsible for the branch. The candidate's squad has a team of special agents, working on a variety of counter-terrorism and criminal operations. The first call of the day is a scheduled coaching session with a special agent on the squad. The candidate has background material describing—through memos, emails, reports, transcribed phone messages—the special agent's strengths and weaknesses. After that initial call, the candidate will speak with a counterpart in another agency regarding an ongoing investigation. He or she will also deal with a counterpart within the FBI regarding personnel and resource matters, interact with other personnel (role played by the assessors) in various typical situations, wrapping up the day with a call-in from the head of the office looking for an update of the day.

During these exercises, assessors operating out of a central call center role play with the candidate and then enter into the data-capturing system their notes and ratings on behavioral dimensions under the eight broad FBI mid-level management competencies listed earlier. For example, under the broad leadership competency, assessors rate three distinct dimensions—mentors, directs, and inspires, each with its own set of behavioral guidelines that provide three to eight illustrations each for less-than-adequate, adequate, and more-than-adequate behaviors that might occur in the course of an exercise. Similarly, for the interpersonal ability competency, assessors rate three distinct dimensions—establishing rapport, respectful, and sensitive to differences. Over the course of the LSA, fifty-five to sixty discrete ratings are captured. Assessors have no access to ratings

provided by other assessors. No assessor evaluates the candidate on more than two exercises. The system also produces statistical process control reports on assessor ratings that can identify when an assessor's ratings on an exercise or on a competency demonstrate a mean or variance that is out of control limits, potentially leading to a review by another assessor. All simulation calls are digitally recorded, and these recordings are used to evaluate reliability in ratings and consistency and realism in role playing. The recordings also provide the FBI with documentation for auditing and review if there is a candidate appeal. The final assessment is an automatically calculated composite on each competency dimension and an overall assessment across competencies. The system produces a detailed assessment report with developmental recommendations (mostly on-the-job experiences, but also readings and training) linked to the candidate's most salient weaknesses demonstrated throughout the assessment.

The FBI invests a great deal of time and effort in assuring that the third party's assessors are familiar and comfortable with the FBI's culture, evolving mission, jargon, and ever-changing procedures. FBI liaison is regularly consulted for guidance on how to interpret and apply the behavioral anchors used to evaluate candidate skill. Further, regular calibration sessions are conducted by the FBI with the assessment team to ensure that the assessors continue to apply FBI standards consistently and can respond realistically to the challenges that candidates present in the course of role playing.

Development and Validation

To create the realistic "day in the life" assessment, a series of SME meetings were held when the LSAs were developed to collect critical incidents, review draft descriptions of scenarios, review scripts, and provide behavioral anchors for the evaluation guidelines. The SMEs were a diverse group of executives who had varying investigative backgrounds (white collar crime, counterintelligence, cyber, etc.) as well as diversity in location, race/national origin, and gender. The SMEs were positioned at a level or two above the target level but had all served and managed people in the target position. No incumbents were used as SMEs

in the initial development to protect test security. Subsequent parallel forms were developed utilizing incumbent supervisors who already had passed the assessment at that level.

Given the history of litigation and the stipulations of the consent decree—in particular the requirement for a panel of industrial/organizational psychologists external to the FBI to review all validation work—the FBI undertook a thorough validation research program using multiple validation strategies. As with most job simulations, a content validation strategy was employed for the LSA. SME panels rated the job relevance and typicality of the role-play exercises; the suitability of the background materials; the linkages of tasks and task clusters to the competencies; and the linkage of all of these to the exercises. The data strongly supported the job-relevance of the LSAs respectively for the two target levels of middle management. The content validation also produced a clear set of evaluation guidelines for which there was strong SME consensus.

The FBI then conducted a criterion-oriented validation study to further build the evidentiary basis supporting the job relatedness of the LSA as well as to comply with the Uniform Guidelines' directive to evaluate the validity of alternate assessment procedures. In addition to the LSA, two cognitive assessments (one of which was customized for face validity) and four personality assessments (one of which was developed solely for this study) were administered to a sample of fifteen hundred incumbents. It should be noted that participants did not receive LSA feedback reports until after criterion data collection was completed.

A research, competency-based performance appraisal measure was developed and content validated for this study. The FBI actually collected criterion data twice. The first, unsuccessful, attempt to collect performance ratings from the managers of incumbents in the study used an online survey tool. The web tool was launched with no specific training, administered with no on-site supervision, and during a major national anti-terror operation. The result: A great deal of missing data and ratings that lacked sufficient variance to have any value. A second attempt was made to collect ratings from managers during scheduled and proctored in-person sessions. Each session began with a twenty-minute introduction on the purposes of the study, potential rating biases, and

how to rate accurately. This approach yielded much higher quality ratings from a sample of 480 supervisors across thirteen representative field offices.

Results of this concurrent study indicated that the LSA produced the highest validities in predicting performance ratings, as compared to cognitive ability and personality alternatives, while demonstrating minimal adverse impact to protected classes. Indeed, the personality scales had very weak correlations with the criterion measures. Any incremental validity obtained by the use of the cognitive ability measures introduced unacceptable levels of adverse impact to the overall composite. A third validation analysis was conducted two years later. Special supervisor performance ratings were collected for the subset of special agents who were originally tested at program launch and were later promoted into a middle management role. The promotion panels for this first group did not have access to the LSA assessment reports; the candidates did, however, receive their own reports as developmental feedback and guidance. The results of this two-year predictive criterion-oriented study confirmed the findings of the original concurrent study.

Benefits

There have been a number of positive outcomes from the implementation of the LSA at the FBI. Candidate acceptance of the process initially and on an ongoing basis is high. The entire promotion process is viewed as a significant improvement in relevance, credibility, and objectivity as compared to the previous process. The simulation is seen by participants as providing a realistic preview of their future roles as mid-level managers. Indeed, some candidates have decided to withdraw from the promotional process after getting an experiential feel for the managerial role through the LSA. In addition, the ease of LSA remote administration—just about any time, from anywhere—has allowed the FBI to meet its applicant flow and time-to-fill needs as positions open.

The perceived quality of leadership as measured through the FBI Annual Employee Survey increased a very material 5 percent for the special agent population two years after the

new promotional process was implemented. Obviously, it is impossible to be sure that these are causally linked. However, there was something of a “natural experiment”: the new assessment was introduced in the special agent population but not for the FBI’s professional staff (non-sworn) population. The difference in quality of leadership between those two populations in the 2009 survey, by which point the mid-management workforce within the special agent population had been almost completely replaced by LSA-qualified staff, was 9.5 percent.

A final point; the FBI leverages the competency profiles emerging from the LSA as a training needs diagnostic to help design targeted developmental interventions on both an individual and agency-wide basis.

The LSA was implemented as part of a resolution to a consent decree stemming from a race-based lawsuit. Thus, promoting qualified individuals while mitigating adverse impact was and remains a paramount issue for the organization. Indeed, the use of the assessment has not created adverse impact. Similar pass rates are found for male, female, white, African American, and Hispanic candidates. Through the implementation of the LSA, the consent decree was resolved. The LSA continues to be used today, not based on any lingering effect of the litigation history, but because it has strong utility to the FBI.

Lessons Learned

This case study has described the application of a technology-enhanced assessment under an adverse set of circumstances (a consent decree) in a one-of-a-kind organization replete with unique challenges. Nonetheless, I believe there are a number of lessons from the FBI that can be shared and generalized to other organizations and situations.

- Most fundamentally, the development, validation, and deployment of a new assessment tool, especially a high-stakes assessment for an internal population, require the simultaneous implementation of a systematic change management process.
- Fidelity of the assessment to the job may be more important than the technical wizardry. The bells and whistles of the

completely automated assessment tool we originally developed had, in our eyes as assessment specialists, a significant “cool factor,” but the assessment was not accepted by the stakeholders and that solution was abandoned. The use of telephone interaction as the medium of assessment was seen by stakeholders as considerably more similar to what occurs on the actual target job. Consequently, a telephone simulation had greater face validity and credibility to those stakeholders. Disciplined change management helps ensure stakeholder commitment through consistent and open dialogue.

- Different stakeholders within the organization add unique value in helping to design specific elements of the solution but not other elements. For example, in our case, incumbent first-level supervisors generated rich simulation scenarios with high fidelity, but had difficulty scaling behavioral anchors for measuring competencies. More senior managers were better able to articulate what good performance in each simulation situation looks like and as a result generated clearer and more useful behavioral anchors than incumbents. Consider carefully and selectively the unique perspective and capability of each source of input into a solution design.
- Validity evidence can be a great communication tool when implementing a new assessment. That evidence in our case was, of course, required by the consent decree. Collecting such evidence is also sound professional and business practice. Beyond those requirements, we learned that hard numbers locally collected can be crafted to effectively address stakeholder resistance. Although, if the LSA had been legally challenged, the content validity evidence would have made a strong case for job relatedness during litigation, having additional criterion evidence from two studies contributed to making a stronger internal business case at FBI. It is instructive to note, though, that different constituencies found different forms of validity evidence convincing. For example, content validity, reflecting fidelity and face validity, was the key to special agent acceptance. Executives, especially those who had helped design the simulation, also found this type of validity evidence most compelling. Middle managers

and their supervisors were most impressed by the criterion-oriented evidence.

- Perceived objectivity and fairness can be enhanced by using third-party professional assessors. Although there was and continues to be some resistance at the Bureau to having “outsiders” conduct assessments and evaluate special agents, the perceived fairness of the process, the narrow focus on scoring leadership competencies, and the arms-length relationship between assessor and candidate have neutralized much of the resistance. The assessors have advanced degrees in psychology or human resources and are hired, trained, and managed to execute their assessment roles effectively. Assuring that these assessors sound like FBI staff requires the Bureau to make an ongoing investment in assessor training and calibration. but the investment has paid off in the high level of acceptance for the LSA within the organization.
- When high-stakes assessment tools are administered to internal candidates, results will be challenged. A special agent with the reputation of being a superstar will inevitably fail the assessment. Phones will ring. The validity of the new tool will be questioned based on this N of 1. If not addressed properly, these incidents early on in a new assessment’s introduction can be fatal. There are two keys, we learned, to responding effectively to such challenges. One is having key executives stand behind the assessment and convey publicly and assertively that the assessment is now an established part of how the organization does business. Strong validity evidence helps give executives the confidence needed to take such a strong stance, as does participation of these key, respected executives on the SME panels. Second, there should be clear, consistent, and transparent messages about how the assessment is scored, how the results are used, and what the retest policy is for those who fail the assessment.
- No one would ever recommend that an organization revise its promotion system or develop new assessments under a consent decree. The process is inherently adversarial, time pressures can lead to poor decision making, and ultimately, the freedom to design what the organization wants and needs

is constrained. The FBI learned, however, that there are some benefits to working under a consent decree:

- Developing and implementing a new assessment tool takes a great deal of effort and money. Often organizations stick with the status quo, even though more effective alternatives are available. The “stick” of a consent decree can stimulate the creativity and focus that potentially produces significantly better results for the organization.
- Access to an external advisory panel of assessment experts can be very beneficial. We learned, though, that the panel’s role needs to be tightly defined. Is the panel there to advise or to approve? In our case, the active participation of the panel in reviewing every step of the development and validation, their freedom to offer advice, and the FBI’s freedom to proceed based on legitimate business needs made for a very productive process.

The FBI continues to tweak and enhance the entirety of the mid-level management promotion system to increase its utility to the organization. Recent changes include an automated application process, advanced collection of local ratings for analytic and feedback purposes, and enhanced feedback capabilities and training and development offerings to the candidate population. Over the course of the past several years, this relatively low-tech process has enabled the FBI to build a significantly stronger leadership culture.

Chapter Eleven

INNOVATION IN SENIOR-LEVEL ASSESSMENT AND DEVELOPMENT

Grab 'Em When and Where You Can

Sandra B. Hartog

This chapter describes an innovative, experiential leadership development program developed in partnership between The Interpublic Group of Companies, Inc., one of the world's leading marketing communication services companies, and its external consultant, a talent management consulting firm. The goal of the program, branded internally as MyLead by IPG, is to provide in-depth, impactful assessment and development to individuals in mid-level to senior leadership roles whose responsibilities include staff leadership, significant client relationships, and business development accountabilities. The target population is dispersed across the globe. Although each office is locally led with a great deal of autonomy, the broader IPG organization has a strong demand for a leadership pipeline.

Three years prior to the launch of MyLead, IPG had instituted a global succession management process. This process helped to define and shape their global leadership development needs.

At that time, the only consistent leadership development program across the enterprise was a leadership program for senior executives—a board-level event. All other leadership development programs were developed and run by local business units with highly variable degrees of quality and regularity. The success of the senior executive development program demonstrated the possibility of success for a consistent leadership development program across business units. Additionally, the succession management process indicated a need for a more consistently available and high-quality approach to development for mid-level and senior leaders.

The assessment and development program addressed the critical issue of developing a common set of leadership standards across a highly diverse, decentralized, global organization, in a manner that blends high-impact delivery with psychological principles of learning and professional development. In designing the assessment and development program, we were committed to capturing the interest and motivation of an elite group of senior leaders who tend to be fast-paced, client-driven individuals, and, as part of a leading marketing communications enterprise, are familiar with the most up-to-date audience impact and visual production technologies. As a group, they tend to place a high value on creativity and innovation and, frankly, a low value on structured programs. Our objective was to develop something that was leading-edge, would create excitement and engagement, and would deliver a high return on investment.

The Challenges

There were several key complicating factors to consider in developing the correct solution. First, there was the need to promote the leadership development of top talent across multiple business divisions, work disciplines, and cultures. At the senior level, the roles within the organization range from functional or technical experts to client relationship or account executives, strategic planners, and creative directors. It was essential to create a program that spoke to the needs of IPG leaders regardless of their role in the organization, agency, or particular office; which did not demonstrate a bias toward one area of expertise, was not unique to any agency or local office structure, and was not culturally or geographically bound.

As with any company, cost was a concern. Per-participant costs needed to be kept very low so as not to be a barrier to entry for any local office. However, costs also needed to provide for the delivery of a high-touch, high-value development solution across thirty-three countries and hundreds of locations.

There was the need for what we began calling a “low drag” experience. This referred to two different notions. First, we needed to provide an in-depth development experience without sacrificing the participants’ available billable hours. Participants could not leave work to attend a program and risk losing potential client revenue, as this would make the cost of any program untenable. Additionally, we needed to accommodate individual work schedules while providing a common experience. With a global, client-driven population of participants, we needed to be very flexible and not dictate set times for participation. Therefore, we wanted to create as close to an “always-on” system as possible that would allow for true flexibility in participation.

Finally, we wanted to provide an individualized experience based on each participant’s unique learning styles and development needs, while also ensuring that there was a high degree of consistency in each participant’s experience. We needed to create learning elements that were flexible enough to accommodate different skill levels and diverse development needs, but that would be perceived as appropriate for the experience and seniority levels of the target audience. Thus it was critical that we integrate multiple business challenges, present them through a range of media, and provide them with one-on-one support and coaching throughout the program. In this way we hoped to cast a wide net with which to draw people to the experiences and insights necessary for true learning to occur.

The challenge was to address all these needs and design an approach to learning and development that would resonate with this audience. The program needed to be immediately compelling, but also offer great value for the time it would demand away from billable work. It needed to speak to the needs of a global senior audience, have low drag, and be flexible regarding time commitments and learning elements that a participant might choose to use. The solution needed to drive company-wide leadership competencies, have individualized learning elements for

each participant, and involve a large degree of personal contact, feedback, and coaching. It needed to offer a consistent experience and, of course, it needed to be low cost.

Identifying the Solution

The assessment and development program was developed with a corporate advisory board from around the world representing all IPG business divisions and agencies convened specifically for the purpose of guiding program development and a cross-agency team of IPG development professionals. By engaging both internal groups from the beginning, we were able to obtain support for the program and gather their subject-matter expertise in helping to prioritize the competencies we identified, as well as their insights into the unique learning needs and optimum delivery modalities of the target audience of participants. The assessment and development program was positioned as something they had a significant hand in creating, not something that was being done either for them or to their constituents. It was also described as a response to senior leader requests for a greater talent pipeline and as a leadership initiative requiring their talents in driving through the agencies.

The first step in the design process was to identify the competencies critical for success in cross-agency, cross-geography, cross-functional senior leadership roles. A competency modeling process was undertaken that involved a representative sample of all functional areas, major agencies, and important global locations of the organization. A set of leadership competencies emerged centered on people leadership, business leadership and profitability, and client leadership. The corporate advisory board reviewed and refined the competency model that was developed.

While developing the leadership competency model, we researched new approaches to senior-level learning, investigated current practices within IPG, and spent time in different agencies to better understand the challenges, motivations, environmental demands, tolerance levels for developmental interventions, and a myriad of other factors. We also spent time considering several different intervention approaches to address IPG's needs.

The first, and most expedient, approach was some type of e-learning solution in which participants could engage, or not, of their own accord. However, the concern with this approach was that it would not necessarily provide the correct content and necessary level of customization for this audience and that e-learning would not be a very engaging or appropriate intervention for leaders at this level. We also knew that, at the other extreme, a university-based program was not going to be acceptable based on the high cost of these programs, as well as the need to send people to a specific location for a significant period of time. Another possibility was a traditional executive coaching program. Coaching programs, however, can be very expensive and also have an inherent variability in quality that makes them difficult to manage. Coaching would also not support the desire to “teach” in any systematic way or single voice the newly identified leadership competencies core to success at IPG. Another consideration was a traditional development-focused assessment center. However, these also generally require individuals to leave their jobs for periods of time and assemble in a particular location. They are generally one-time events and often perceived to be divorced from real-world activities. While effective, they rely on the assessor/coach being able to observe the individual in-situ, which on a large global scale can become very expensive and administratively difficult to manage. There was also heightened sensitivity to the notion of senior leaders “being assessed.” Instead, the corporate advisory board wanted to promote the program as an opportunity for leaders to further their learning, to be challenged, and to be coached in skills they could immediately apply back to the workplace. They wanted neither an evaluative atmosphere around the program, nor any repercussion as an outcome of poor or even superior performance.

This all led us to conclude that we needed to look for a new approach to leadership development that leveraged best practice from multiple approaches and broke set with previous models.

The Solution

The assessment and development program developed for IPG sought to solve these challenges by delivering the joint benefits of a developmental assessment center and executive coaching

delivered over time and distance using Internet-enabled technology. The program brings together four critical elements to create a highly impactful leadership assessment and development experience: the rigor of assessment center methodology, the power of executive coaching, the impact of learner-directed content, and the global reach and flexibility of web-based technology.

This approach allowed us to provide participants with opportunities to experiment with new leadership behaviors in a virtual assessment center setting deliberately turned into a learning environment with both one-on-one support, in the form of role players and coaches, and e-learning elements. The assessment and development program became the focal point for introducing the leadership competencies into the organization, determining participant learning needs, and building the requisite leadership skills. The program contributes directly to organizational strategy by focusing participants on challenges associated with growing the business, improving profitability, leading people, and managing clients. The assessment and development program built in a large amount of participant choice regarding what to do and when to do it. We believed this would be appropriate for leaders at this level and increase engagement as well as leverage the advantages of distributive learning, which maximizes transfer back to real-world settings.

Table 11.1 describes the components and flow of the development center.

Over a period of seven weeks, the assessment and development program participants alternate between assessment/simulation and coaching weeks for three rounds of learning. Each participant is assigned an executive coach. The coach acts as a guide throughout the program, offering feedback and coaching about the participant's areas for development and how to tie learning back to the job. The coach also helps the participant identify development opportunities with the simulation that serve to focus the participant's experience on challenges that will most readily develop the participant in the areas identified.

The program kicks off with a set of introductory materials that describe the program, time commitments required, goals, and coaching and confidentiality agreements (which are to only reveal whether someone completed the program). Participants

Table 11.1. The Assessment and Development Program Structure

Week 1	
Orientation and Preparation	<p>Receive program information and complete introductory materials</p> <p>Complete planning discussion with executive coach</p>
Week 2	
Simulation Module 1	<p>Participate in the simulation for four hours during the week</p> <p>Address issues</p> <p>Receive real-time feedback on phone interactions</p> <p>Receive ongoing access to development resources</p>
Week 3	
Feedback and Coaching	<p>Meet with coach</p> <p>Access development resources and apply learning back to the job</p>
Week 4	
Simulation Module 2	<p>Participate in the simulation</p> <p>Receive ongoing access to development resources and message board</p>
Week 5	
Feedback and Coaching	<p>Meet with coach</p> <p>Access development resources and apply learning back to the job</p>
Week 6	
Simulation Module 3	<p>Participate in the simulation</p> <p>Receive ongoing access to development resources, and apply learning back to the job</p>
Week 7	
Feedback, coaching, and development planning	<p>Meet with coach</p> <p>Create a post-program development plan</p> <p>Meet with manager to review development plan</p>

are required to complete the Honey and Mumford (2006) Learning Styles Questionnaire (LSQ) and a career accomplishment profile via web-administration and engage in an initial coaching interview with their assigned coaches via telephone. This initial set of exploratory materials and coach conversation offer an opportunity for the participant and coach together to begin a process of customizing the experience for the participant. They identify and discuss the learning elements of the program through the lens of the learning styles and self-report assessment of leadership competencies. The coach suggests those elements of the development program that have the most utility for each participant's preferred learning style, recommends leadership competencies to focus on, and engages the participant in a discussion of specific challenges in his or her current role or possible desired stretch roles.

During assessment/simulation weeks, a participant plays the role of a leader in a fictional, global organization. As opposed to our typical day-in-the-life experience, the simulation represents a quarter-in-the-year experience. Each of the three simulation weeks represent a successive quarter for the fictional organization with evolving challenges. Participants address challenges through telephone interactions with role players in a phone bank (portraying any and all of the subordinate team members, clients, or colleagues), and in-basket exercises with email, voice-mail exchanges, business and budget reports, and other information all delivered through the technology platform. Each of them has one extensive role play with his or her coach during each assessment/simulation week, which addresses client, team, or business leadership. Participants choose what to engage in based on the competencies they have targeted for development. By presenting a parallel world in a fictional organization and in a different industry, agency or role biases are removed and the playing/learning field is leveled across all participants. However, we created a world that closely reflected the participants' to ensure fidelity and gain credibility as a real-world experience. Participants engage in challenges such as building a client presentation to generate business, recruiting a high-potential candidate in the industry, negotiating contracts with clients, resolving

turf issues regarding scarce resources, and coaching a derailing manager.

Each activity engaged in by the participant during the assessment/simulation week (for example, email exchanges and role plays) generates feedback. The coach, in turn, integrates all of the feedback data generated from the participant’s role-play interactions and in-basket exercise, and identifies thematic strengths and development needs. During coaching weeks, the participants have calls with their coaches, during which the coach provides feedback to the participant and discusses learnings, opportunities for development, content to focus on in the next simulation week if possible, and ways to transfer insights back to the workplace.

The entire program is delivered via an online platform that allows for learner-directed access to content such as additional e-learning modules, development planning tools, and self-assessment and journaling tools to help learners process their experiences throughout the development program. Figure 11.1 illustrates these elements. Figure 11.2 shows the simulation “desktop” through

Figure 11.1. Screen Capture of Program Interface

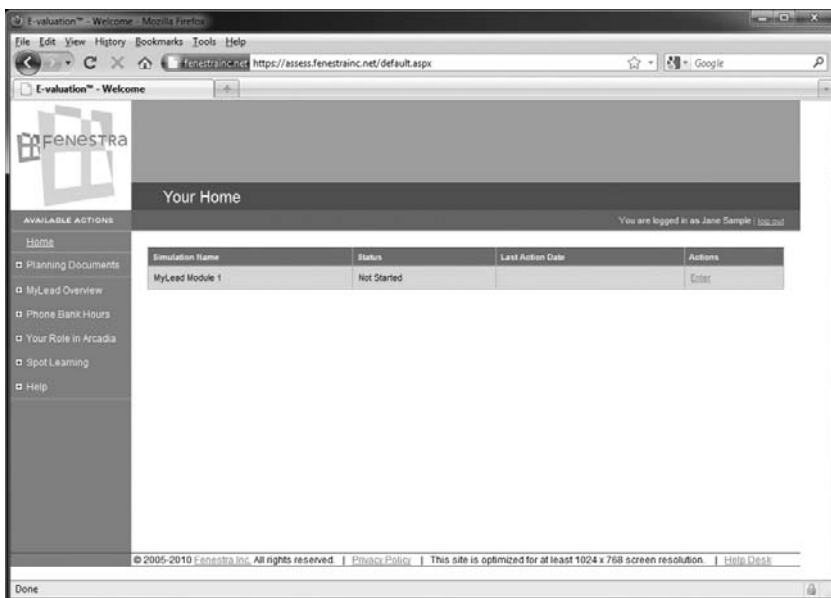


Figure 11.2. Sample Email



which participants engage in the assessment/simulation itself to address challenges of the fictional organization. The assessment platform mirrors the interactivity of a real computer desktop.

Program Innovations

The assessment and development program presents three innovative advances in assessment and development. First, relying on Honey and Mumford's concepts of learning styles (Honey & Mumford, 2006), we developed a set of tools and activities that we believed were adaptable to each of the four learning styles identified by the researchers. This construct defines four learning styles: Activists, Pragmatists, Reflectors, and Theorists.

We believed that the learning style of the participants would impact both their ability to learn from an online leadership development simulation as well as their choices on how to learn and designed the learning experience accordingly. For example, the program was designed to have reflection worksheets at the end of each simulation week to allow the Reflectors ample opportunity to collect the data acquired during the simulation experience and

think about their perspectives. For Theorists, we began the program with an introduction to the leadership model that was the basis for the program. Then, throughout each element of the program, the feedback referred to the part of the leadership model the simulation was addressing. Additionally, each participant's learning style influenced how the executive coach interacted with the participant. For example, a coach working with a Theorist might suggest certain books for the Theorist to investigate new concepts, while, when working with an Activist, the coach might suggest an on-the-job developmental activity. While all elements were available to everyone, developmental suggestions were tailored to each participant's learning style to optimize learning. In this way, the assessment and development program provides a learner with a variety of modalities to acquire skills. These modalities can be readily mixed by a coach to guide the participant to the next round of learning.

A second innovation in our design was to provide a development center experience via web-based technology and combine it with "high touch" components, that is, role players and coaches who would be accessible on the phone and via Internet. This would enable geographically dispersed participants to be engaged in an extensive individualized development experience (thirty hours over seven weeks) with a high degree of scheduling flexibility, but with minimal program cost. Access to the simulation materials remained available on a continuous basis throughout the entire simulation week. Participants could log in and do "work" whenever convenient based on their schedules and from wherever they wanted. Time zones became relatively irrelevant.

The technology-enabled development center platform offered many benefits for IPG's global organization. It provided in-basket exercises, emails, and business cases with embedded information personally addressed to each participant with accurate times and dates, regardless of time zones. This is a great aid to realism and engagement; it is another element contributing to a high-fidelity simulation experience. It also deployed program invitations, feedback reports, post center surveys, and such. It linked, via embedded URLs, to other testing sites or e-learning platforms. It provided for the back-and-forth exchange of emails and other information between a participant and his or her

coach. Additionally, it allowed participants to work with whatever tools they normally used during the course of a day, for example, Internet searches, software programs such as PowerPoint or Excel, and so on.

The technology behind the web delivery of the simulation and assessment materials requires little beyond a high-speed Internet connection, a computer equipped with an Internet browser (Safari, Firefox, etc.), and a traditional land line, mobile, or VoIP phone connection. At this point in our global world, these requirements can be considered mundane and easily obtainable in almost all work, travel, and home settings. The technology behind the web-based simulation works as SaaS (software as a service) and does not require participants to install software onto their computers or require a vast amount of bandwidth for the technology to run smoothly.

The third innovation was a unique intelligence system developed to allow a “phone bank” of English-speaking role players to assume different character roles within the simulation. Role players were junior-level assessors who participated in a four-hour behavioral observation training program, carefully reviewed each simulation and supporting materials, and who were given extensive character notes for each module. They were most often graduate students in industrial/organizational psychology and at times included professional staff from IPG, which helped to keep costs down and added additional organizational perspectives. Since participants had the ability to telephone any character they wanted to during preset hours, we needed a way for role players to keep records of previous participant calls in order to maintain continuity of participant interactions over time and across a number of different assessors/role players. This was important to deepen the realism of the simulation and enable the coach to evaluate behavior over time and integrate the perspectives of multiple assessors. The intelligence system had several components. After each role-play interaction, the assessor/role player completed a competency driven, behaviorally based feedback form that was reviewed by a program manager within one hour for quality-control purposes and then made immediately available to the participant. Role players completed a tracking form organized by participant, in which they recorded the outcome

of the call and any agreed-on actions. This would help the role players to maintain consistency in the characters and story lines across interactions for each participant, continue to “grow the story,” and challenge a participant with higher levels or different types of skills, depending on the participant’s individual needs. For example, a participant, playing the role of a manager in the simulation will reach out to his or her direct report/role player based on some third-party speculation that the direct report has received a job offer from a competitor. This may result in a series of conversations in which the direct report/role player is non-committal about his or her intentions and may try to negotiate certain perks and/or increased responsibilities. The conversation will increase or decrease in certain areas of difficulty, depending on the participant’s needs (for example, political savvy, negotiation skills, ability to provide performance feedback, etc.).

Results

The assessment and development program has been deployed globally across hundreds of local offices from eleven of IPG’s agencies. In twenty-two months, we have had 370 participants in twenty-three countries. Almost two-thirds of the participants have been based outside the United States. Sixty percent of the international participants have been from Europe, 10 percent from Asia, and 20 percent from Latin America and Canada. The assessment and development program has become a core part of the development curriculum in each of IPG’s major business divisions and was prominently featured in the development plans of mid- and senior-level executives in the 2009 succession management process.

Program administration is facilitated by the online platform. Participants can engage in the program twenty-four hours a day, seven days a week, from any computer with web access. If participants choose, they can spend a few hours on their office PCs and resume work from their homes. All pre-arranged contacts with their coaches can be scheduled at the participants’ convenience.

Using the administrative features of the technology platform, participant engagement is monitored by measuring the use of different learning modalities. As of this writing, participants have, on average, three to five role-player interactions, one

coach interaction, and fifteen to twenty email interactions during each of the three simulation modules. Each participant who has completed the program has engaged in at least three of the four planned one-hour coaching sessions and many have requested ongoing coaching engagements. All participants have accessed at least one of the self-directed development tools, while the majority has accessed three or more. Program completion rates are approaching 75 percent.

To determine the measurable results and quality of the assessment and development program, we collected pre- and post-program data from participants. Before the program, we asked participants to rate their skill level on the twelve targeted competencies and to identify the three strengths they would like to build upon or opportunities they want to pursue. Immediately after the program, participants were also asked to respond to a series of questions about their experiences, satisfaction, and competency improvement.

Data from the post-program reaction survey shows that the program was well received by the participants:

- 100 percent of participants agree that the program is flexible, easy to use, and find value in their feedback/coaching calls with their coaches.
- 93 percent of participants agree that the assessment and development program will help them in their current roles.
- 93 percent of the participants would recommend the program to others.

Of the areas each participant targeted for development during the program:

- 100 percent of the participants report improving in at least one of the leadership behavioral areas (people, client, and business leadership).
- 89 percent report improving in at least two of those areas.
- 73 percent report improving in all three targeted areas.

Based on the successful results, positive feedback, and increasing enrollment, IPG has expanded the program as of this

writing. We are exploring additional program evaluation measures, including retention rates, development plans completed, and robustness of the succession pipeline. A similar program for lower-level managers was developed and has been rolled out to a significantly larger population.

Lessons Learned: The Good, the Bad, and the Ugly

Based on the post program survey data and anecdotal feedback, we knew that through a strong partnership with the client we had created a program with significant success. At the end of the program, participants complete long-term development plans and are encouraged to share the plans with their managers. Managers are given support and tools with which to meet their staff at least halfway and have constructive developmental conversations. There was a buzz about development. The learning and development community across the company was telling us about developmental conversations they were having, managers were engaging in more career conversations, and there were several requests for additional coaching from participants. Participants talked about the realism of the simulation, the ability to have spontaneous conversations with characters within the simulation, and the technology-enabled feedback and other support tools that existed that made it both interesting and educational. They talked about the ease of access to the materials and the impact the ongoing coaching had on them.

There were aspects, however, that did not go quite as well as we had expected. One of the guiding principles for the program was to allow as much flexibility as possible to increase the usability of the experience and create a “low drag” intervention. The dropout rate for the program was higher than expected and approached 25 percent. By extending the program over seven weeks, we may actually have compounded a difficult situation. The people we were trying to serve were very busy, client-driven executives. A seven-week time frame, even though it presented a lot of flexibility in the course of each week, likely cut across many different cycles of a work project’s lifespan, including both busy and slow periods. Because of this, even those with the most motivation and best

intentions to begin and finish the development center may have met with circumstances and time commitments beyond their control. Those who dropped out generally left due to work demands, and some rescheduled for a more opportune time.

For the U.S. staff, the most challenging helpdesk support hours were for our Asia Pacific sessions. We decided to run what we began to call our “all Asia, all the time” groups. Basically, it was easier to run a program focused almost exclusively on the time zone needs of an Asian cohort than to blend this cohort across the needs of the rest of the globe. Similarly, the company had specified that English was the language of business for their senior staff and many clients. However, our primary English-speaking helpdesk had to provide assistance to individuals who were not necessarily comfortable with English instruction. Additionally, all simulation and e-learning materials were delivered in English. However, there were requests for feedback and coaching in a range of languages. These requests were accommodated when possible.

We also learned a number of technology-related lessons. One was to always consider the client technology environment itself, specifically, to consider cross-platform and cross-browser issues. IPG had a significant percentage of users on various Mac platforms. To accommodate this, we needed to expand the list of not only the operating systems on which the application would seamlessly run, but also address browser issues specific to each operating system.

A final technology lesson learned was the need for more user-experience testing of the technology related to work flow to ensure that what we thought was simple and intuitive was in fact regarded by other, unfamiliar users as such. While we designed the platform to be as intuitive as possible and tested its functionality with user groups before launch, we found some trepidation among the assessment and development program users about the technology itself, as most had never participated in any kind of online simulation. At the beginning of the program, users received an email that directed them to the platform, where, we and our testers believed, the process was mostly self-evident. However, there was more confusion than we expected. These concerns led to the need to modify the program itself: rather than simply sending an email with instructions, each participant’s

coach, in an initial meeting, would familiarize the participant with the platform by doing a thorough walk-through of the technology. This, in turn, led to much faster acceptance of the technology as the vehicle for the assessment and development program and, hopefully, to use of all the components.

Overall, the experience allowed us to greatly expand the reach of the technology from an assessment-only environment to one that could deliver learning and development programs as well.

New Directions

As technology advances, so do the opportunities. The growing sophistication of technology will bring more opportunities for increased “touch” or remote assessor and coach contact in all parts of the world. Webcams and other telephony services will all but eliminate the need for travel in assessment and development centers, unless there is an explicit reason or desire to travel. Employee portals allowing for self-registration and scheduling technology can match participants, assessors, and coaches. Advance interfaces, including abilities for interactive learning opportunities, avatars, and more responsive simulations, will allow for sophisticated branched learning and increasingly difficult scenarios, more engaging simulations, and richer feedback experiences. An increased use of on-demand learning that integrates a new approach to blended learning will become a more frequent solution to difficult development problems.

Finally, it is extremely important to remember that just because something can be done does not necessarily mean that it should be. In making the switch from a traditional developmental assessment center paradigm to a technology-enhanced one, the project team should also consider the overall goals, client populations, receptivity to technology, access to high-speed bandwidth, a strong technology support infrastructure, and a host of other factors before implementing a program of this type.

Reference

- Honey, P., & Mumford, A. (2006). *The learning styles helper's guide*. Maidenhead, UK: Peter Honey Publications Limited.

Chapter Twelve

CASE STUDY OF TECHNOLOGY- ENHANCED ASSESSMENT CENTERS

Rick Hense and Jay Janovics

This case study focuses on a revision of the selection process for the role of a bank branch manager within a large, global financial services company. Technology was used to enhance the process, particularly through replacement of a more traditional role-play exercise and through advances in personality assessment.

Bank branch managers have responsibility for the operations, sales, and customer service of a single bank branch. They have management responsibility (hiring, performance management) for in-store tellers and bankers. The position is considered a mission-critical role because of its impact on a large population of employees, its potential for revenue and relationship building, and its risk management responsibilities. With many thousands of applicants a year, it is considered a high volume role by the recruiting organization.

The previous selection process included multiple tools designed to assess critical competencies for the role as identified through job analysis. In brief, the process started with an automated minimum qualification screen within the applicant tracking system and a phone screening interview followed by a

web-delivered biodata/personality assessment, a role-play exercise, and two behavioral interviews. The previous role-play exercise was well liked but presented multiple challenges:

1. It was time-consuming. Forty-five minutes were required for the role-play actor and also the observer/note taker, in addition to the time spent scheduling.
2. The geographic dispersion of the role-play administrators limited opportunity for quality training and calibration. When first implemented, training consisted of a three-fourths-day, in-person training. Over the years, the training was shortened and moved to phone administration.
3. The role play was originally designed to be conducted with the candidate in person. However, remote locations necessitated phone administration in some areas.
4. The role play had been in place for many years and the content was compromised. Management likely coached high potentials with the process (not trying to be dishonest, trying to be helpful) and information on the role-play scenarios could be found on the Internet.

Description of the New Assessment Process

Overview

In developing the new assessment, a major design objective was to deploy a realistic assessment with enhanced multimedia (video-based) and psychometric (computer adaptive testing) technology. The intention was to the extent possible create an online version of an assessment center that would be amenable to unproctored use. In particular, in order to be accepted by the organization as an acceptable alternative to the in-person role-play assessment, the assessment battery had to contain a component that would simulate one-on-one interactions with employees.

The project team faced four major constraints from the outset. First, in order to more economically assess geographically dispersed candidates, the organization wanted to have the option to deliver the assessment online in a completely unproctored fashion. Second, given that internal stakeholders had become accustomed

to using assessment methods with high physical fidelity (that is, an in-person role play), the assessment had to include components that were very face valid and realistically portrayed situations encountered by job incumbents. Third, the organization was concerned about the length of the assessment and wanted to keep it to no more than one hour in length. Finally, the organization wanted to keep costs in line with their current online assessment and have the option of replacing their in-person role play without incurring additional costs.

Traditional Biodata and SJT Assessment Content

While the use of new assessment technology was important for meeting the organization's goals of deploying a highly realistic, face valid assessment, the project team did not want to completely break with historically predictive assessment content. To this end, it included three empirically keyed biodata scales from the Supervisory Potential Index source instrument (SPI; PreVisor, 2001). These scales have considerable accumulated validity evidence and yet are short, taking no more than fifteen minutes to administer. The latter point was important given that the new assessment content was expected to require over forty minutes to complete.

Computer-Adaptive Personality Scales

To measure relevant personality characteristics, the Global Personality Inventory-Adaptive (GPI-A; PreVisor, 2008) was selected. The GPI-A is a thirteen-scale general assessment of normal adult personality developed for use in selection and development contexts in organizational settings. The assessment employs a within-trait forced-choice format in which test-takers are asked to select which of two behavioral statements is most true of them. The statements presented reflect different elevations of the same underlying trait, as established by trait ratings collected from a team of industrial/organizational psychologists. A computer-adaptive engine selects a second pair of behavioral statements based on the information about trait elevation obtained in the first pairing, and the test proceeds in this fashion until suitably

Table 12.1. Sample Traits from Global Personality Inventory-Adaptive (GPI-A) Measure

<i>Scale Name</i>	<i>Definition</i>
Achievement	Setting and accomplishing challenging goals; taking satisfaction and pride in producing high-quality work and excelling in one's own efforts; working hard, exerting effort, and persisting despite significant obstacles; competing with self and others.
Confidence and Optimism	Believing in own abilities and skills; feeling competent and successful in multiple areas; remaining self-assured and optimistic even in the face of rejection.
Influence	Persuading and negotiating effectively with others; influentially asserting ideas and thoughts; adeptly moving others to a decision or favorable outcome; effectively networking with others; coordinating individuals' efforts to accomplish work.

accurate theta estimates are generated. Three sample traits from the GPI-A measure are listed in Table 12.1.

The GPI-A was selected for use in this test battery based on its anticipated advantages over traditional personality instruments. For one, administration time would be reduced given the computer-adaptive nature of the assessment. As mentioned previously, keeping assessment times low was of major importance to the organization. The computer-adaptive format was also expected to provide better test security through reduced item exposure. This was particularly important given the desire to employ the assessment in an unproctored setting. Finally, in line with research on forced-choice personality scales, the GPI-A was expected to be less susceptible to motivated distortion compared to traditional scale formats (Christiansen, Burns, & Montgomery, 2005; Jackson, Wroblewski, & Ashton, 2000).

Video-Based Situational Judgment Test

As noted above, a major objective for the organization was to develop an assessment to evaluate how well candidates would perform in one-on-one coaching and developing activities with their direct reports. Because this test would be replacing the role-play assessment, it was critical for the new test to be as realistic as possible. To this end, the project team developed a multimedia situational judgment test (MMSJT; Olson-Buchanan & Drasgow, 2006) based on situations commonly encountered by bank branch managers. This test was designed to evaluate how effectively potential bank branch managers would interact with their subordinates—to identify the extent to which they would employ effective strategies such as listening, encouraging, staying constructive, etc., while interacting with employees in a tactful and constructive manner.

The scenarios and response options included in this test were developed based on in-depth interviews and focus groups with a set of experienced, high-performing job incumbents. Bank branch managers described situations they commonly encountered in a coaching context, such as disputes over vacation time or providing feedback to employees following interactions with customers. Two industrial/organizational psychologists wrote six five-part scenarios based on these situations. Each part consisted of an item stem (employee dialogue, to be converted into a video clip) and six response options (potential manager responses). The item stems were kept short and the interactions limited to five parts each to reduce the size of the video files that would ultimately need to be downloaded by position candidates when completing the test. The intended format involved having test-takers view a video clip, select a response from the options provided, and then watch and respond to the next video clip, thereby simulating an actual dialogue with the employee.

Response options were written to represent a variety of different ways managers might choose to respond to the employee and included desirable coaching behaviors as well as undesirable options. The interactions were written so that the five parts would all logically go together, regardless of which response options

were chosen by the test-taker. Thus, each five-part scenario simulated a short, self-contained conversation.

Thoroughly edited scripts were used in video production. Complete video clips were then presented to a sample of fifteen subject-matter experts from the organization's learning and development team. These individuals had considerable leadership experience ($M = 16.4$ years) and were involved in providing coaching training to the organization's management ranks. The SMEs were asked to view the video clips and rate the effectiveness of the response options provided on a 4-point scale (1 = ineffective; 4 = highly effective). The initial set of six response options was winnowed to four options based on the magnitude of and agreement between SMEs' ratings.

Once the final four response options (and associated item scoring) were identified for each item stem, the test was assembled and deployed online. The test was configured such that for each of the six scenarios, the test-taker would view some background information regarding the interaction that was about to take place and then view each video stimulus and select the most effective and least effective responses from the four options provided. The test would then advance to the next part of the scenario, or in other words the next stage of the "conversation." A screen capture from the final version of the assessment is provided in Figure 12.1.

Validation of the New Assessment

Prior to use with job candidates, the newly developed assessment was evaluated in a concurrent validation study. Current bank branch managers were asked to complete the entire assessment, unsupervised, on their work computers at the time of their choosing. While they did complete the assessments on company property, relaxation of other common test-taking requirements (that is, at a specified time on an unfamiliar computer in a supervised testing room) resulted in a reasonable simulation of the unproctored environment that would be encountered by actual job candidates.

Figure 12.1. Screen Capture from Video-Based Coaching SJT

Step 1: Scenario

One of your personal bankers, Dave, recently transferred to your banking center after working for ten years at a lower-volume center in a nearby city. Your banking center is very busy, with a high volume of customers during peak hours. So far, Dave has had trouble keeping up with the work pace, and you've noticed that it takes him much longer to service customers than the other bankers. There are always customers waiting in the lobby to see a banker when Dave is on shift. He has a tendency of being overly chatty with customers, talking with them about

Watch the following video and choose the most and least effective course of action from the options below.

Step 2: Choose

	Most Effective	Least Effective
Yes, that's exactly what I wanted to talk to you about. I'm concerned that the work pace here might be a bit more than you're used to.	<input type="radio"/>	<input type="radio"/>
Yes, it's busier than your last banking center. And I've noticed that we have much slower customer service and longer lines during your work shifts. I'd like to talk about some strategies for helping you service our customers more efficiently.	<input type="radio"/>	<input type="radio"/>
Right. Your approach may have been suitable at your last banking center, but I need you to be much faster and more efficient than you have been so far.	<input type="radio"/>	<input type="radio"/>
Yes, I can tell that it's much busier than you're used to. What can I do to help you adapt to the pace? I've noticed that you have struggled a bit when we get busy.	<input type="radio"/>	<input type="radio"/>

Next

Participants' managers provided performance ratings using an on-line performance appraisal. The performance evaluation was developed based on a job analysis and consisted of forty-two different ratings of various relevant performance dimensions as well as more general work effectiveness ratings (overall effectiveness, overall achievement relative to peers, likelihood of re-hire if given the opportunity to do so, etc.).

Two hundred twenty-seven bank branch managers completed at least a portion of the assessment battery. Of these, usable performance ratings were available for 158 participants. Final sample sizes for the analysis ranged from 138 to 158 for the various assessment scales. Some participants completed only a subset of the coaching SJT; they were included in the analysis if they had responses for an entire scenario, but composite SJT scores were only computed for those bank branch managers who completed the entire SJT.

Table 12.2. Observed Correlations Between Coaching SJT Scores and Performance Ratings

<i>Coaching SJT Scale</i>	<i>Performance Composite</i>	
	<i>Managing Talent</i> ²	<i>Overall Rating Composite</i> ³
Scenario 1	.22*	.17*
Scenario 2	.21*	.22*
Scenario 3	.15	.11
Scenario 4	.26*	.21*
Scenario 5	.16	.15
Scenario 6	.04	-.05
Composite of All 6	.28*	.24*
Composite of Best 5 ¹	.31*	.28*

$N = 138$. * $p < .05$. Values in table are observed, uncorrected validity coefficients.

¹The Composite of Best 5 includes scenarios one through five.

²Managing Talent is a composite of three performance ratings: Building Relationships, Developing Employees, and Handling Conflict.

³The Overall Rating Composite is a simple average of all forty-two performance ratings.

Observed validity coefficients for the individual coaching SJT scenarios are presented in Table 12.2. Note that these validity coefficients are based on composites of most effective and least effective responses within each scenario. As the table shows, correlations between individual scenarios and the overall performance composite ranged from $-.05$ to $.22$, with all but one of the scenarios exceeding $.10$. Similar relationships were found when these scenario scores were correlated with more specific, conceptually-aligned performance criteria such as the Managing Talent performance dimension composite. The sixth scenario was the weakest predictor of performance, correlating just $-.05$ with the overall performance composite. The total SJT score, represented by the "Composite of all 6" row in the table, correlated $.24$ ($p < .05$) with the overall performance composite. A composite of the best

five scenarios (excluding scenario 6) was slightly more predictive at $r = .28$.

Many of the other assessment scales were also predictive of bank branch manager performance. The strongest predictors included two of the biodata scales (observed r s = .39 and .38) and the Influence (.25) scale of the GPI-A. The final version of the assessment recommended for use with job candidates included the five most predictive scales in the coaching SJT, the three biodata scales and the eight most predictive GPI-A scales. A unit-weighted composite of these scales had an observed validity estimate of .41.

Development of the Mini-Role-Play Interview

As part of an end-to-end review of the entire selection process, the current interview guides were revisited. Hiring managers had grown to appreciate the interactive and rich evaluation of coaching that the previous role-play provided. Thus, “mini-role-play interviews” were included in the guides in addition to behavioral questions. These brief situational questions gave hiring managers the opportunity to evaluate the candidates’ demonstration of coaching skills without the need for candidate materials and extensive assessor training.

Configuration of the Final Assessment

After completing the concurrent validation study, there was much discussion about the best way to configure it for use with job candidates. Given the difficulty and expense associated with developing the coaching SJT, the preference was to only administer that test in a proctored environment. But unfortunately, the organization did not have any good options for administering the assessment in bank branches, because candidates were too widely dispersed geographically. The option of using local testing centers was considered and rejected.

The project team identified a compromise option of having a two-part unproctored assessment, with all of the assessment content except for the coaching SJT included in the first stage. These tests are all either empirically keyed biodata or

computer-adaptive personality scales, mitigating concerns about their widespread unproctored use. To protect it from exposure, the coaching SJT was included in the second stage with single-use assessment links that were only sent to those candidates who achieved passing scores on the first set of tests. Consequently, the coaching SJT (1) was only completed by a subset of the most qualified candidates and (2) could not be re-taken or forwarded from one candidate to another.

As a further means of addressing test security, the instructions of the coaching simulation included a statement warning candidates not to share answers or discuss the content of the assessment with other parties. This was particularly important given that the test would be used for both external hires and internal promotions.

The final assessment was therefore configured as a two-part assessment with two separate hurdles. In order to be recommended for hire, candidates must achieve passing scores on both a composite of the tests included in the first stage as well as the coaching SJT administered in the second.

Organizational/Political Challenges

As with any change to the selection process, organizational/political challenges must be addressed. The changes required substantial buy-in from multiple groups, including the business, human resources, staffing leadership, and front-line recruiters. Challenges and buy-in were primarily addressed through process development and communication.

The tool development and validation approach was designed to improve buy-in. The coaching simulation scenarios were created through SME observation and focus groups, and scoring was developed using business experts. The concurrent validation approach made the internal sell much easier. Favorable validation study results helped communicate clearly that the new assessment tools could differentiate performance of the organization's current population. In addition, new interview guides were developed together with recruiting and business experts and received unusually positive feedback, especially considering previous problems with interview utilization.

A thorough communication and training plan was created with different messages created for each group and targeted at appropriate level for non-technical audiences. For example, the business presentation included an executive summary followed by two pages of example results using expectancy charts. Communication tools specifically addressed the issue with most expected resistance (replacing the role play).

Technological Challenges

In general, there have been few technological challenges associated with moving from the existing role-play assessment to the online assessment. The only exception to this has to do with wireless Internet connections, which do sometimes interrupt downloads for the coaching SJT. In response to this, the organization added a statement in the introduction to the assessment advising candidates that they should complete the assessment only on a computer with a hard Internet connection.

Other difficulties have been limited to occasional reports of problems from individual candidates. In every case, scores were captured by the computer system and were available when looking up candidate results. This suggests that any reported problems may have been related to candidates not realizing that they had completed the assessment.

Overall, the assessment has been in use for six months with fewer problems than expected based on past experiences with assessments requiring broadband connections. This may indicate greater technical savvy among job applicants, or possibly more widely available access to broadband Internet connections. It is also possible that problems have been avoided because of some conscious decisions made when the assessment was developed. For instance, by downloading video content directly into the local PC's browser cache, the video takes a moment to load, but then plays seamlessly (and is then deleted when the browser's cache is cleared). Similarly, video files were deliberately kept as small as possible to limit download times, and the download occurs while candidates are reading the background information for the scenario. To avoid any problems with candidates missing an important piece of information, test-takers are

given the opportunity to replay videos if they choose to do so. These features result in a more streamlined process that limits or outright avoids any potential for difficulty when completing the assessment.

Measuring Success

In addition to the criterion-related validity evidence already presented, anecdotal feedback about the assessment and the process has been positive from recruiters and the business. Pass rates and adverse impact analyses are conducted frequently and consistently meet established targets. To further support the initiative, an efficiency analysis showed very large per-candidate savings. The project team evaluated the organizational time spent on the previous role play (forty-five minutes each for a recruiter and hiring manager) compared with the time requirements for administering the coaching simulation (five minutes for the recruiter), along with the opportunity cost savings using average salary for administrators. The opportunity cost saving just for the recruiting organization for the first six months easily demonstrated positive return on the development cost investment. Finally, a predictive validation study is planned for next year.

Lessons Learned and Recommendations

Based on this experience, the organization has several “lessons learned” that would influence the approach if this process were repeated. These also serve as recommendations for anyone else seeking to move from a “high-touch” in-person assessment to an unproctored web-delivered assessment.

First, it is important to get early buy-in and participation from project stakeholders. Regular update calls and frequent communication with constituents ensured that they had input and knew what the assessment would include. In addition, such steps as using job incumbents as SMEs (to enhance SJT realism) and internal content experts to develop the SJT score key (to increase confidence in how responses are evaluated) were very helpful in encouraging ultimate user acceptance and buy-in. Organizational

stakeholders were particularly pleased with the look and feel of the SJT, and it was helpful to be able to explain how their internal coaching experts helped develop the score key.

In terms of communicating the “why” behind deploying a new assessment method, the value of having favorable concurrent validation results cannot be overstated. When presenting the assessment solution to stakeholders, the project team was able to show not only the proposed process and tools but why their use was justified. The team shared empirical expectancy charts based on specific performance dimensions that would be particularly relevant to stakeholders, such as “managing talent” and “customer service.”

The project team also learned that, when implementing a new technologically advanced assessment, it is important to be prepared to react to unexpected challenges, such as the technological hurdle of wireless Internet connections. This relatively minor issue was still one that required swift action and was addressed by adding a notification for candidates. Anticipating any potential challenges was important, so internal testing of the multimedia assessment was essential, particularly given that the platform and technology were new. In this case, it was beneficial to have a single point of contact from the organization’s information technology department to review and test the assessment content and provide feedback.

Finally, it is important to recognize that tradeoffs are inevitable. Ideally, the coaching assessment would be delivered in a proctored setting. But administering the assessment on-location was not a realistic option. Once it was decided that the assessment had to be administered in unproctored settings, the project team came up with ways of accommodating that approach. For instance, exposure to the coaching SJT was limited by including it in a second stand-alone assessment stage completed only by those candidates who passed the other tests and the recruiter interview. This second stand-alone assessment was deployed with single-use links to the assessment sent out on a candidate-by-candidate basis. This made it impossible for candidates to forward the link to other potential candidates or to take the test repeatedly, thereby preserving test security. In addition, in an attempt to reduce the chance that the test would be compromised, it includes a

warning about not sharing responses or information about the test with other parties.

Conclusion

This chapter described a case study involving the development, validation, and implementation of a technology-enhanced pre-employment assessment. The assessment battery consisted of both multimedia technology (a video-based situational judgment test) as well as advanced psychometric technology (computer-adaptive personality scales). The assessment was validated and successfully implemented as part of the selection process with only minor and easily resolved technological issues. This new technology-enhanced online assessment successfully replaced the existing in-person role-play assessment, offering substantial time and cost savings while maintaining favorable reactions from organizational stakeholders.

References

- Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering the use of forced-choice formats for applicant personality assessment. *Human Performance, 18*, 267–307.
- Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance, 13*(4), 371–388.
- Olson-Buchanan, J. B., & Drasgow, F. (2006). Multimedia situational judgment tests: The medium creates the message. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application*. Mahwah, NJ: Lawrence Erlbaum Associates.
- PreVisor. (2001). *The Supervisory Potential Index (SPI) technical manual*. Developed by Richardson, Bellows, Henry & Company. Minneapolis, MN: Author.
- PreVisor. (2008). *PreVisor Computer Adaptive Personality Scales technical manual*. Atlanta, GA: Author.

Chapter Thirteen

VIDEO-BASED TESTING AT U.S. CUSTOMS AND BORDER PROTECTION

Jeffrey M. Cucina, Henry H. Busciglio,
Patricia Harris Thomas, Norma F. Callen,
DeLisa D. Walker, and Rebecca J.
Goldenberg Schoepfer*

In this chapter we discuss the practical application of video technology for assessing applicants for law enforcement officer positions at U.S. Customs and Border Protection (CBP). As in many law enforcement positions, incumbents at CBP must have excellent judgment and be exceptionally skilled at interacting with both the public and co-workers. CBP has developed a video-based test (VBT) for assessing the judgment and interactional skills of applicants. The VBT uses video technology to enhance the realism of the situations presented to applicants as well as applicants' responses to the situations. This chapter describes the VBT approach used by CBP, outlines the process for developing VBTs, describes the scoring process, presents psychometric and related evidence in support of the VBT, and provides practical suggestions to industrial/organizational (I/O) psychologists who

*Opinions expressed are those of the authors and do not represent the position of U.S. Customs and Border Protection.

are interested in developing and implementing VBTs. The chapter begins with a description of the VBT approach and highlights its effectiveness in enhancing the realism of applicant assessment while also reducing the burden on the organization.

Description of the VBT

Many of CBP's uniformed law enforcement officers work at various entry points to the United States, where they handle a number of duties associated with passport control, customs, immigration, agricultural, and anti-terrorism issues. A job analysis conducted by CBP identified a number of critical competencies, or knowledge, skills, abilities, and other characteristics, that are required for successful performance in these positions. Some of the critical competencies, such as logical reasoning and math skills, are measured by written tests and educational requirements. Other critical competencies are more difficult to assess using written tests and include the four competencies that are measured by the VBT: interpersonal skills, emotional maturity, cooperativeness/sensitivity to the needs of others, and judgment/decision making. In the past, CBP measured these four competencies using a situational structured interview, during which interviewers presented applicants with a verbal description of a situation. When responding, applicants described what they would do in the situation. Although the structured interview had benefits, such as affording an opportunity for face-to-face interaction with applicants, it lacked the realism that occurs when the four competencies are used on the job.

The VBT allows for higher fidelity in the delivery of questions to applicants through the use of videotaped scenarios that depict situations more realistically than a spoken narrative. Using more realistic depictions can reduce biases caused by differences in applicants' interpretations of situations. Additionally, applicants respond in a more realistic fashion; that is, they must respond as if they were actually in the situation rather than merely describing what they would say or do in the situation. The greater realism of the VBT also provides an excellent realistic job preview to applicants.

The VBT approach has also dramatically reduced the organizational burden associated with assessing applicants. The situational

structured interview that was previously used required approximately three supervisory staff hours per applicant and had to be conducted during normal business hours at various locations across the country. This created scheduling challenges for CBP, which conducts operations twenty-four hours a day. In addition, travel costs were incurred for locations that lacked a sufficient number of interviewers, which required interviewers to travel to various locations. Also, the structured interview posed a considerable burden on interviewers by requiring extensive written documentation of applicants' responses and subsequent justifications for ratings. Since raters are not required to be onsite during the administration of the VBT, they can score VBT responses on their own schedules. This eliminates the scheduling issues that were associated with the interview and allows the VBT to be administered in one location and rated in another, nearly eliminating travel costs to the agency. Furthermore, time savings accrue because raters are only required to view the actual applicant responses and do not have to sit through the instructions and scenario presentations for each applicant. The situational structured interview also required extensive note-taking on the part of the interviewers to document what the applicant said or did during the interview—this is not needed under the VBT approach. Typically, CBP had added a third interviewer to the process to assist with note-taking and question reading. Under the VBT approach, only two raters are needed because applicant responses to the VBT are recorded and kept indefinitely and scenarios are presented to applicants using videotapes. The two VBT raters can focus on observing and rating the applicants rather than administering questions and recording responses. In all, the VBT only requires two supervisors to spend fifteen minutes each when rating an applicant, compared to the three supervisors who were needed for one hour each when interviewing and rating an applicant under the structured interview process.

Development of VBTs

The development of a VBT involves eight steps. The first step entails conducting a comprehensive job analysis of the position to identify the critical tasks and competencies that are needed

for successful performance. In the second step, critical incidents (Flanagan, 1954) are collected from subject-matter experts (SMEs) to document situations that arise on the job involving the targeted competencies. In step three, I/O psychologists and video production staff review the critical incidents and identify those that are suitable for testing and amenable to filming. In the fourth step, the critical incidents are used to create scenario briefs, which are short, one-paragraph descriptions that portray the situation (but not the behavior and consequences) from a critical incident. Step five consists of pre-production activities, which include the (a) use of scenario briefs to draft scripts; (b) identification and auditioning of actors and actresses; (c) procurement of props; and (d) creation of a schedule for filming. Pre-production at CBP has been a joint effort between professional video production staff and I/O psychologists. The sixth step is video production, the filming of the scenarios. At CBP, all scenarios were filmed at an operational worksite, resulting in an accurate portrayal of the job and the work environment. The filming locations included a sample of the relevant job sites where critical CBP activities occur. In the seventh step, SMEs review the filmed scenarios and rate each scenario on importance, difficulty, frequency, and competency coverage. Probable applicant responses are generated for each individual scenario and are used in the creation of benchmarks for each competency-based rating scale. In the eighth and final step, individual scenarios are pieced together into multiple versions of complete VBT tests, matched by difficulty and competency coverage.

Administration of the VBT

The VBT is administered by trained test administrators in a specially equipped test room. Figure 13.1 presents a photograph depicting the test room and equipment setup. Applicants enter the test room and are provided oral and written instructions by the administrator. Next the administrator begins playing the VBT test tape and departs, leaving the applicant alone in the room while taking the VBT. The VBT test tape includes a narrated introduction to the VBT, two sample scenarios, and eight evaluated

Figure 13.1. Typical Setup for Administering the VBT

scenarios. A typical scenario might begin with a voice-over and a widescreen shot of a job site to give some context and background information. Next, the applicant might encounter a disgruntled or agitated individual and a situational dilemma occurs. The applicant has forty-five seconds to respond to the individual and effectively resolve the dilemma or problem. The applicant responds as if he or she were actually in the scenario by talking directly to the individual shown on the TV monitor. A camcorder positioned next to the TV monitor records the applicant's response for later viewing by a panel of trained raters.

Applicants are alone in the room while viewing and responding to scenarios shown on a twenty-inch standard TV monitor. The scenarios are stored on VHS videotapes and applicant responses are recorded using a mini-DVD camcorder placed on a tripod. Applicants are allowed to sit or stand during the test session and most choose to sit when viewing the scenarios and stand

when responding. Raters view the applicant responses using a DVD player that is attached to the same TV as shown above. Typically, raters will rate a number of VBTs in one sitting and the raters do not have to be at the same location where the VBT was administered.

In the final scenario, applicants are shown three travelers and must indicate which individual(s) warrant further inspection. Applicants are given ninety seconds to respond to the final scenario. This type of scenario has proven to be a very effective means of determining how well applicants make judgments about other people, apply attention to detail, and discern behaviors that might be indicators of illegal activity. Once the VBT is completed, the administrator thanks and debriefs the applicant. The recordings are later scored by a panel of raters who have completed an intensive one-and-one-half-day rater training session covering the scoring and rating process. All raters are CBP employees who have experience working in the occupations covered by the VBT.

Present Use of the VBT

Each year, as many as fifteen thousand applicants take the VBT at fifty-six locations across the United States. All locations have a VBT coordinator responsible for overseeing local VBT activities and at least one VBT administrator, who is typically in an administrative support position. The larger locations have a pool of raters who have attended the one-and-one-half-day rater training session.

Psychometric and Related Evidence Supporting the Use of the VBT

Scoring and Reliability

Applicant responses to the VBT scenarios are scored in a two-phase process. In the first scoring phase, two raters independently use a three-point scale to provide scores for the competencies measured by each scenario. The decision to use two raters was guided by rater reliability analyses, cost, defensibility against challenges, and best/standard practices in federal hiring (whereby two or three raters are typically used with subjective scoring

systems). Next, the two raters independently total up their scores for each competency collapsing across the eight scenarios. The total scores are compared to predetermined cutoff scores for each competency, which were established by a panel of SMEs. In the second phase, the raters share their individual scores and reach consensus on their total scores for each competency. An applicant must reach the cutoff scores for all competencies in order to pass the VBT.

CBP recently conducted an analysis to ascertain the reliability of the VBT scoring process and reported evidence of high rater agreement on the scores. Prior to consensus, individual raters agreed on the pass/fail status of the competency scores 96 to 98 percent of the time and on the overall pass/fail status of applicants approximately 95 percent of the time. In terms of the pass rates for the VBT, CBP designed the VBT to have a similar level of difficulty and passing rate as the situational structured interview that it replaced.

Validity and Adverse Impact

The VBT has been validated using content and construct validity strategies. In terms of content validity, the scenarios were designed to measure key competencies from a traditional job analysis, and a critical-incidents job analysis approach was used to develop the scenarios and provide further job analytic support. Throughout the development process, the scenarios were reviewed at each stage to ensure that they were part of the content domain of situations encountered on the job and that they were linked to the critical competencies measured by the VBT. SMEs rated the scenarios on importance, frequency, and competency coverage at various points in the development of the VBT. Furthermore, the VBT approach has very high job fidelity, because applicants view situations that occur on the job and then respond as if they were incumbents on the job.

Construct validity evidence for the VBT has been demonstrated through the relationship between performance on the VBT and scores on the other assessments used to select applicants.¹ For example, applicants who passed the VBT scored higher on a composite of cognitive tests ($t = 3.024$, $p = .003$,

$d = .29$), as well as on two of the component tests: logical reasoning ($t = 2.953$, $p = .003$, $d = .35$) and writing ($t = 2.189$, $p = .029$, $d = .26$). A relationship with logical reasoning can be expected given the prominence of cognitive ability in predicting performance in many different settings (for example, Kuncel, Hezlett, & Ones, 2004) and the relationship of cognitive ability with situational judgment and practical intelligence. In addition, the correlation between writing test scores and VBT scores seems reasonable, given Carroll's (1993, p. 620) work showing a higher order factor for language incorporating aspects of written communication and oral communication. We examined the relationship between performance on the VBT and scores on another situational measure: a low-fidelity paper-and-pencil simulation/situational judgment test designed to assess applicants' suitability for a law enforcement career. Applicants who passed the VBT scored higher on the low-fidelity simulation ($t = 2.454$, $p = .016$, $d = .42$), providing additional construct validity evidence.

In terms of the VBT's criterion-related validity, meta-analytic estimates for related measures suggest a validity coefficient within the range of paper-and-pencil situational judgment tests ($\rho = .20$; McDaniel, Whetzel, Schmidt, & Maurer, 2007) and situational structured interviews ($\rho = .50$; McDaniel, Whetzel, Schmidt, & Maurer, 1994). To further substantiate validity, CBP is planning to conduct a local criterion-related validity study that will include marker measures, work samples, and supervisory ratings of job performance. The VBT pass rates for minority groups are very similar to those for majority groups and adverse impact ratios are well in excess of the 4/5 rule of thumb prescribed by the *Uniform Guidelines*.

Cost-Effectiveness

Exhibit 13.1 shows a list of the costs associated with developing and implementing VBTs. Despite the higher development and implementation costs, CBP has experienced substantial post-implementation cost efficiencies using the VBT approach. The VBT costs \$59 per applicant to administer and score, which is about half the cost of the situational structured interview (\$137) it replaced. Much of this cost savings is due to a reduction in rater time and travel expenses.

Exhibit 13.1. VBT Costs

Development

- **SME panel travel costs.** Typically three week-long panels with six to eight SMEs each are used to (1) collect critical incidents and review scenario briefs, (2) review scripts, and (3) review final filmed scenarios and create scoring guidelines. You should consider the costs of a typical SME panel in your organization.
- **Test developer time/salary.** VBT development lasts about one year and would require approximately three test developers each working part-time on the project, for a combined total of one full-time equivalent.
- **Extra reviews required for each scenario.** More time is required for reviewing VBT scenarios than is typically needed for interview or paper-and-pencil item review. This is due to the cost that is involved in creating VBT items and the difficulty in changing a scenario after it is filmed.
- **Video production costs.** Typically, video production costs can approach the six-figure mark, but depend heavily on the availability of in-house resources for producing videos.
 - *Script writing.* It is best to have a professional script-writer turn the scenario briefs into a script. Alternatively, a test developer with script-writing experience could be used.
 - *Salary and expenses for video production staff.* This involves paying for a producer, an assistant producer, a production assistant, a cameraperson, an audio engineer, and a grip.
 - *Travel costs.*
 - *Pay for actors and actresses.* Typically the “day rate” for the Screen Actors Guild is used.
 - *Post-production editing costs.* This is the cost of editing the filmed footage and requires renting a studio and paying for editing staff.

- *Props and wardrobe.* Props and wardrobe can typically be obtained cheaply from second-hand stores, and the total cost is usually under \$500. If incumbents in your organization wear uniforms, you may obtain uniforms in all available sizes for actors and actresses.

Implementation

- **Test production and distribution.** There will likely be costs involved in producing copies of the VBT scenarios and distributing them to field locations.
- **Equipment for administering the VBT.** This could involve computers and webcams, or in CBP's case, TVs, VCR/DVD players, AV carts, camcorders, tripods, etc.
- **Supplies for administering and rating the VBT.** Typical supplies include blank mini-DVDs/tapes when camcorders are used to record applicant responses, servers/hard drives for storing applicant responses when an online testing system is used, batteries, rating forms, shipping supplies, safes for securing test material, etc. Purchasing items in bulk can save money.
- **Rater and administrator time.** Consider salary costs, management buy-in, etc.
- **Program management time.** One to two psychologists will likely be needed to work part-time on the project to manage the VBT implementation; conduct rater training or train-the-trainer sessions; provide guidance to raters, administrators, and coordinators in the field; brief management; review applicant challenges; conduct psychometric and statistical analyses of the data; etc.

Applicant and Test User Reactions

In general, both applicant and test user reactions to the VBT approach have been quite favorable. All applicants are invited to voluntarily complete a six-question anonymous survey immediately after completing the VBT. The survey responses are analyzed annually, and a summary of the findings are presented in Exhibit 13.2. In terms of test user reactions, administrators,

raters, and coordinators in the field have been fully supportive of the VBT approach. The VBT initially began as a seven-location pilot program, after which the raters, administrators, and coordinators voted unanimously to expand its use nationwide. When asked, the test users in the pilot program cited more precise measurement of skills, more efficient use of time, a faster rating process, and increased reliability of ratings as the main benefits of the VBT.

Exhibit 13.2. Results of Applicant Reactions Survey

1. How would you rate your performance on this test?
 - I performed exceptionally well on this test. 11%
 - I performed above average on this test. 39%
 - My performance on this test was about average. 48%
 - I performed below average on this test. 2%
 - I performed poorly on this test. 0%
2. How comfortable were you in responding to the scenes on the TV monitor?
 - I was very comfortable responding to scenes on the TV monitor throughout the entire test. 21%
 - I became comfortable responding to the scenes on the TV monitor after the practice scenes were given. 32%
 - I became comfortable responding to the scenes on the TV monitor after responding to a couple of real test scenes. 44%
 - I was not at all comfortable responding to scenes on the TV monitor throughout the test. 3%
3. How sufficient were the instructions for this exam?
 - The instructions were sufficient for responding to the scenes on this exam. 95%
 - The instructions were somewhat sufficient, but I could have used more instruction prior to responding to the scenes on this exam. 4%
 - The instructions were not sufficient, resulting in confusion. 0%

4. How would you rate the forty-five-second response time at the end of each scenario?
 - Too long 24%
 - Just right 75%
 - Too short 1%
5. What did you like most about the video-based test?
 - Depicted real-life situations 19%
 - Gave a realistic job preview 15%
 - Miscellaneous (categories that were only mentioned by two respondents or fewer) 11%
 - Was alone in the room and didn't have to appear before a panel 8%
 - Instructions were good/detailed 7%
 - Efficient/fast/concise 3%
 - Test was good/fair 2%
6. What did you like least about the video-based test?
 - Provided a positive comment/didn't dislike anything about the VBT 22%
 - Miscellaneous (categories that were only mentioned by two respondents or fewer) 18%
 - No feedback from characters/actors on TV monitor 15%
 - VBT format was hard to get used to 9%
 - Response time too long 8%
 - Nervous on camera 4%
 - Impersonal (no interaction with a real live person) 3%

Suggestions for Future VBT Developers and Lessons Learned

This section describes lessons learned in developing VBTs and outlines some considerations for I/O psychologists who wish to develop VBTs. Much of the initial apprehension over the implementation of CBP's VBT turned out to be unfounded. One initial concern was possible applicant objections to being videotaped. This concern proved to be unfounded, as only a very small number of applicants have objected to being videotaped, and most

appear to forget that the camcorder is on after the first few scenarios. Videotaping applicants was deemed to be preferable over having applicants respond in front of a live audience of raters, which might be awkward for the applicants and time-consuming for the raters. Audiotaping applicants was also not considered desirable because it does not allow for the observation of non-verbal communication and complicates positive identification of applicants during legal challenges.

Another initial concern was the length of the forty-five-second time period allotted for the applicants' responses to each VBT scenario. CBP has found that the vast majority of applicants actually respond in thirty seconds or less, and the forty-five-second response time is more than adequate for responding orally to interpersonal situations. Yet another concern was the potential for applicants to fail to adapt to the role-play format of the VBT. CBP has found that very few applicants experience problems with the role-play format. When responding to VBT scenarios, applicants play the role of the job for which they are applying and thus, display the typical reactions and behavior required to handle a certain job situation. This results in less "acting" and more "reacting" on the part of the applicants. Finally, CBP learned that applicants quickly become familiar with the VBT process and understand how the process works. Providing applicants with two unscored sample scenarios and extensive instructions has been especially beneficial. A list of other suggestions for future VBT developers is presented in Exhibit 13.3.

Exhibit 13.3. Suggestions for Future VBT Developers

Development and Pre-Production

- **Plan on developing multiple versions of your test at once.** In terms of economies of scale, it is more cost-effective to develop enough versions to last at least several years than to develop a new version every year.
- **When reviewing critical incidents from SMEs, consider the "filmability" of the situation.** Scenes that

require special effects, large numbers of actors and actresses, stunt performers, etc., may be beyond your budget. In addition, situations without a lot of action, interpersonal interaction, etc., may not present well on video and could be better measured using a paper-and-pencil test.

- **Envision the possible range of applicant responses to a scenario before you decide to film it.** Think about the different plausible responses an applicant could have to a scenario and consider the competencies underlying each response.
- **Remove unnecessary job knowledge from scenarios.** Sometimes this will create a scenario that is not 100 percent representative of the job, but still measures the critical competencies. This is critical if your applicants are applying for an entry-level job that does not require job knowledge at entry.
- **Make sure that all types of major work locations and functions are represented.** Often, B-roll, which is secondary footage that sets up a scenario by showing the work location, can be of assistance.
- **Read the scripts out loud when reviewing.** It is very helpful to actually play out the scenarios and read them aloud. In addition, writing for a script is often very different from writing to be read off paper. Script writing tends to be more colloquial with less formal language and shorter sentences.
- **Thoroughly review your items, scripts, etc., at every step.** In contrast to paper-and-pencil or online tests, the cost of a filmed video item is very high, and it is very difficult to change a filmed video scenario.
- **Balance the demographics in your scenarios.** Create a matrix of different demographic groups and ensure that no group is cast in a negative or positive light more often than the other groups.
- **Conduct a scouting visit to the filming location and ask video production staff to come with you.** The scouting visit helps to ensure that the space is suitable for

(Continued)

Exhibit 13.3. Suggestions for Future VBT Developers (*Continued*)

filming and reflects not only the typical work setting but also the image the organization is trying to convey. The video production staff will be examining the lighting, adequacy of the power supply, etc.

- **Ensure you have “staging areas” for filming.** You will need space nearby for actors and actresses to be briefed and to rehearse, changing rooms, break areas, space for catering setup, secure overnight storage for video equipment, etc.
- **Obtain filming permits, if necessary.** Usually you will not need to obtain a permit to film if you are on “company property.”
- **Purchase good props.** Try to find props that show up well on video (that is, are large, easily recognizable, etc.) and avoid purchasing props with brand names or logos.
- **Purchase props well in advance of video filming.** Do not wait until the day of filming to obtain props or assume that they will easily be available at the film location.

Video Production

- **Consider upcoming changes to the job before filming.** Are aspects such as the job duties, uniforms, etc., likely to change in the near future?
- **Film at an operational work location.** Filming at an actual busy work location will give applicants a better realistic job preview. A mock work environment for training purposes can often look very deserted and unlively unless a large number of actors and actresses are brought in.
- **Thoroughly review the script beforehand.** It does not hurt to memorize the script yourself before filming, even if you will not be one of the actors or actresses. Knowing the script yourself will make it easier to spot mistakes made by actors and actresses that will impact the assessment.
- **When possible, use real actors and actresses instead of employees.** A particular employee you hoped to film may not be available the day of the film shoot due to

operational issues. Employees can also forget their lines and “clam up” in front of the camera and film crew.

- **Audition your actors and actresses.** During the auditions you will evaluate the performance of the actors and actresses and gauge how well they reflect the position and workforce you are portraying. If your film crew is unionized, you may be required to use actors and actresses who are in the Screen Actors Guild.
- **Have an SME on hand at all times during filming.** The SME should be knowledgeable about the attire/uniform/dress code of employees, the wording used, mannerisms, realism, job-relevance, etc.
- **Think deeply about on-the-fly changes to the scenarios.** Any changes you make during filming might not be rectifiable later on; consider filming the scenario both as is and with the suggested changes.
- **Be a test security vigilante.** Avoid having an actor or actress in more than a few scenes to prevent overexposure to the test material; have everyone on set sign security agreements; ensure the test video is kept secure; keep track of all printed copies of the script; do not provide actors and actresses with the script before the day of filming or allow them to take the script out of the filming area.
- **Be the assessment expert/advocate.** You will likely be the only expert on test development present during filming and will need to ensure that the filming results in a scenario that meets professional and legal testing guidelines.

Implementation

- **Consider how the VBT will be implemented and what the equipment requirements will be.** It is often best to ensure that all testing locations have the same equipment. This is not only important for standardizing test conditions for applicants but also for standardizing training materials for test administrators.
- **Consider how technologically savvy your test administrators and raters are.** Training often has to be developed for the lowest level of knowledge, as some test users may have no experience with audiovisual equipment.

Conclusion

In conclusion, CBP's VBT is a very successful assessment tool that uses technology to produce high-fidelity simulation, while resulting in considerable resource and cost efficiencies. Future VBT developers can contact the authors for more information and to receive a copy of a tutorial and exercise booklet that has been presented at recent conferences.

Note

1. The sample described here was comprised of 735 applicants who were tested at nine locations in the summer of 2003.

References

- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *41*, 237–258.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, *86*(1), 148–161.
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60*, 63–91.
- McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*(4), 599–616.

Chapter Fourteen

GOING ONLINE WITH ASSESSMENT

Putting the Science of Assessment to the Test of Client Need and 21st Century Technologies

Eugene Burke, John Mahoney-Phillips, Wendy Bowler, and Kate Downey

Setting the Scene and the Scope of This Contribution

With the growth in the Internet and the impact of associated technologies and business models on public and private organizations such as software-as-a-service (SAS) and cloud computing, it now seems inevitable that assessment and, specifically, testing would move online. If one looks back at the history of testing, testing and assessment have always been influenced by technology, whether that be the facility to mass produce test forms on paper (a factor that enabled testing programs as far back as the sixth century in China; Weiner, 2008), standardized testing of military personnel during the First World War, and automated scoring machines developed in the decades that followed, through to the delivery of testing via micro computers in the 1980s alongside the growth of computer adaptive testing (CAT)

using such technologies and the facilities for test distribution offered by local and wide area networks (LAN/WAN). As Tippins stated at the 2008 Society for Industrial Psychology Conference in San Francisco, the Internet testing train has left the station; the challenge to us in terms of science and practice is very clearly how to safeguard the value of testing and assessment when they are administered under conditions in which a supervisor or a proctor are not physically present, although Drasgow, Nye, Guo, and Tay, L. (2009) have recently questioned whether proctored testing is the gold mark that it is often believed to be, and when such administration and distribution of materials may be open to such threats as piracy of content and cheating.

Another factor to consider is how the facilities offered by Internet and PC technologies have also influenced our attitudes to how and when we undertake personal transactions. In the UK and particularly the U.S., the last two or three decades have seen an easing of the time constraints that previously defined working hours and personal hours such as when and where one could browse products and execute a shopping transaction. In other words, we have come to expect greater flexibility and immediacy to the transactions we undertake in our daily lives, and this is no different in the world of work when we are looking for employment opportunities. Today, the world of employment is populated with job boards and recruitment sites as well as major applicant tracking systems (ATS), talent management systems (TMS), and learning management systems (LMS), all of them providing portals through which external job applicants and internal candidates can access materials and processes in order to pursue employment opportunities, whether that be as a new hire through on-boarding to succession and exit from an organization. As such, and just as in our personal lives, these technologies provide applicants or candidates with the convenience of determining when and where they engage in an employment transaction.

We as a profession have systematically claimed that testing and assessment adds value to such processes through predicting training and job performance, and through guiding both organizations and applicants/candidates¹ in terms of person-job, team, or organizational fit (Robertson & Smith, 2001; Salgado, Anderson, Moscoso, Bertua, de Fruyt, & Pierre, 2003; Schmidt &

Hunter, 1998; Schmidt, Shaffer, & In-Sue, 2008). The acceptance of that evidence base has to be seen today in the context of increasing pressure to demonstrate the return on investment in talent management processes such as recruitment and selection. This, in turn, has put pressure on human resource (HR) professionals to demonstrate the value added by the processes they design, purchase, and implement. In many ways, the base case for going online with assessment is similar to the one that drove the growth in automated scoring and CBT systems in the late 20th century, that is, greater process efficiencies and cost savings as well as improvements in quality.

However, just as the talent management agenda has developed over the past decade, today we see other drivers and metrics coming to play beyond the basic economics of acquiring and managing talent at a lower unit cost. One such driver is employer brand and the impact, positive or otherwise, that HR processes, and assessments as part of those processes, have on company brand. Concerns around the role of assessments in supporting an employer's value proposition (EVP) become more accentuated when one considers that a job applicant might also be an existing or a potential customer for the organization's goods and services, and candidate experience is increasingly to the fore of conversations around assessment generally and online assessment specifically. Indeed, the term "candidate-centric" processes, with the advent of social networking sites, is now being used with increasing frequency in discussions and surveys around online HR processes, talent attraction and acquisition, and, therefore, online assessment.

The focus of this contribution is to share experiences over the past decade in addressing business needs for two major and international companies. We will start with the development of a solution to unsupervised Internet testing (UIT) targeted at graduate or campus recruitment that grew out of a need to show efficiencies in talent acquisition processes in the early 1990s, but has since developed to look at broadening the assessment of cognitive abilities to meet wider talent needs as well as ensuring that the solution is equivalent across different languages and geographies. How the solution sits within a broader talent management framework will also be described and discussed.

We will then move onto a different organizational need and a solution for call center operatives in customer service. Here the key driver was how much assessment was required to provide for selection and placement to three different roles—what the first author calls the “assessment window” and which points to a tension between the length of an assessment that is usually (and often incorrectly) associated with validity, versus the potential negative impacts and attrition among applicants who see the assessment process as onerous and who then disengage as applicants. This solution also provides us with the opportunity to look at how an assessment solution, in addition to meeting issues of security and validity of UIT, must also address the social and political context in which an organization has to operate (the example is in banking, which is a very hot topic at the time of writing) as well as be shown to address the talent agenda and EVP of that organization.

Throughout the case materials we will share, we are conscious of the facets of validity articulated in Messick’s (1989, 1995, 1998) seminal work. In addition to the technical quality that should be expected of any assessment and psychometric instrument, whatever the mode in which it is administered, the key aspect that the case materials will focus on is that of consequential validity. That is, in considering whether a solution is fit for purpose, are the consequences of using that solution seen by various stakeholders to meet the needs of the organization, its representatives such as recruiters and managers, and the candidate?

Starting at the Beginning: The Driver for UIT Administration of Cognitive Ability Tests

Our story begins in the late 1990s and early 2000s or the “noughties” as they have come to be called in the UK. The setting is an annual graduate recruitment program for the organization in which the second author is employed, a strong international brand with a track record of attracting applicants from among the highest rated universities and business schools in Asia, Europe, the UK, and the U.S. The organization already had a structured competency-based assessment process supported by criterion and construct validity in place. At that time,

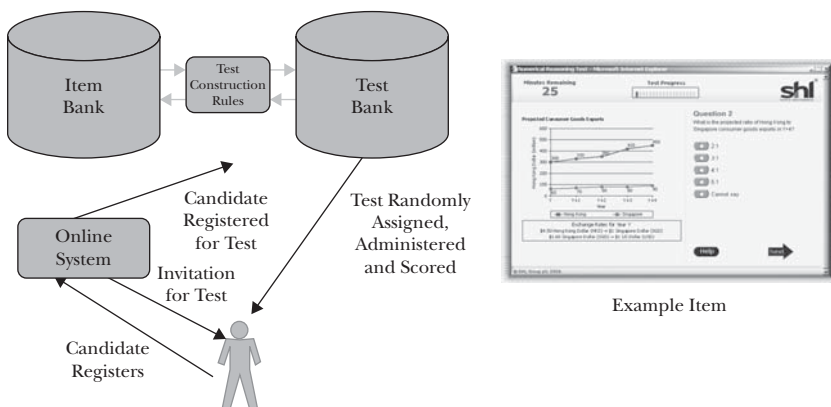
the program was required to supply 250 graduate hires into various business functions such as investment banking from a large pool of applicants, and the program was run from the autumn into the early spring of each year. The organization had already introduced an online pre-screening questionnaire as part of an Internet-based process for capturing applications, and so had already begun to explore ways in which Internet technology could be used to help manage the cost per hire and to track the progress of applicants through the talent acquisition process.

One component of the assessment centre was the administration of numerical reasoning tests providing information against competencies such as problem solving and analysis. The remit for exploring UIT was whether a numerical reasoning test could be developed that would be shown to tap into the same construct measured by the paper-and-pencil test administered within the assessment centre and offer security against potential cheating through the piracy and exchange of test content in the form of items and/or test forms (a concern fuelled by experiences in educational testing as described by Cizek, 1999). In other words, a single fixed form for UIT was not seen as offering sufficient security, and there were concerns around whether several equivalent forms would meet the security concern.

In considering the client's needs, and after a review of the literature, the solution, while drawing on the published work around that option, was not seen as CAT. Indeed, case studies and reviews such as those of Davey and Nering (2002)² have since shown that CAT alone may not offer sufficient security for UIT settings. The solution that was developed was based on the linear-on-the-fly-testing or LOFT model using an item response theory (IRT) calibrated item bank developed using item blueprints matching the content of the paper-and-pencil numerical reasoning test already in place. On-the-fly test construction rules sample the item bank by item difficulty while also checking for item enemies (for example, items that have similar content and answer structures that could increase the likelihood of applicants identifying the correct answer to an item irrespective of their true abilities), and check for equivalence of properties of tests so constructed prior to them being registered in a test bank (that is, if a test does not meet several checks in terms item and test parameters,

then that test is rejected and is not registered in the test bank). Following the pre-screening questionnaire, applicants are taken via a link to register for the UIT numerical reasoning test and issued a username and password. We have moved to the present tense as this application used by the client today retain the basic functions and processes described. A test is then assigned to the applicant at random, is downloaded using an encrypted package onto the applicant's workstation, which could be a home PC or one available at, say, a university location. At the point the applicant wishes to sit the test, the application slaves the workstation and uses the workstation's clock to time the test. On completion of the test or when time for administration of the test is reached, the application then closes and, on next connection with the Internet, the entire application is then sent back to the provider's server, where the responses are scored and processes such as norming applied. As the reader will note, and at a time when Internet connectivity was constrained by dial-up connections and relatively small bandwidth, the solution also removed the need for the applicant to maintain Internet connection to sit the test (that is, the application operates client rather than server side), and security was maintained through a unique configuration of test items for which the scoring template for that configuration was not downloaded with the test items (that is, all scoring information is retained server side on secure servers). See Figure 14.1.

Figure 14.1. Schematic Summary of LOFT UIT Process



The evaluation of this solution addressed two audiences. The first was the client's I/O psychologist and HR staffs involved in implementing the solution and managing the graduate recruitment assessment processes with a focus on whether the results from the first UIT test correlated with results from the second paper-and-pencil proctored test. The second was the client's general management in terms of whether the solution retained quality (as indicated by addressing the issues of UIT and proctored test relationship) and reduced the cost per hire.

The second author's organization has adopted a strong evidence-based approach to talent management processes. Among regular reviews of processes and their value to the organization, the I/O group led by the second author also conducts regular criterion validations across its various business functions and operational territories. For example, evaluations of numerical reasoning tests from the Management and Graduate Item Bank (MGIB) administered in supervised conditions as part of assessment and development centers obtained observed validities³ of 0.31 against assessment centre ratings for problem analysis, problem solving and decision making (aggregated sample size of 660 across four studies of graduate recruitment assessment centers conducted between 2001 and 2003 in Australia, Switzerland, and the organization's European operations), and 0.59 against development centre ratings of intellect (two studies conducted for investment banking and operations in Europe between 2001 and 2005). Accordingly, in extending the assessment of numerical reasoning out further to earlier stages of the talent acquisition process, it was important that a strong relationship be shown between the existing supervised numerical reasoning tests and the UIT numerical reasoning tests that would now precede them. Evaluations conducted in between 2001 and 2006 of primarily graduate recruitment campaigns conducted in Asia and Europe yielded a sample weighted correlation between both test scores of 0.45 for an aggregate sample of 1,349 across six studies. Correcting for range restriction only, and reflecting the high selection ratios for this organization's graduate and experienced hire recruitment programs, yields an estimated correlation between test scores of 0.66.⁴

In terms of the savings of interest to the wider management audience, the reductions in numbers of applicants proceeding to the assessment centre stage of graduate recruitment netted a 32 percent reduction in the cost per successful hire while maintaining the volume of hires sought by the organization. From a cost base of around \$1.2 million (factoring in the human resources management time as assessors, as well as facilities and materials), this amounted to an annual saving of around \$380,000. This only captures part of the full value of the solution when other less direct returns and costs are factored in, such as maintaining the quality of hires while also reducing the opportunity cost to the organization in management time spent as assessors and selectors.

More recently, the organization has extended the range of UIT test types to include inductive (also referred to as abstract) reasoning alongside the numerical reasoning test, which can be reasonably cast as a facet of deductive reasoning (as described by Fleishman & Reilly, 1992). A recent evaluation of both UIT test types against competency based interviews (CBIs) showed that both test scores correlated with interviewer ratings of applicants on the client competency of judgment (problem solving and analysis), while UIT inductive reasoning scores predicted interviewer ratings on innovation. For a sample of 126 applicants, hierarchical regression models show CBI ratings of judgment were predicted with almost equal weight for numerical and inductive reasoning ($R = 0.20$, $p = 0.012$). For the same sample, regression models showed a significant improvement in predicting CBI ratings of innovation when inductive reasoning scores were included in the model ($R = 0.22$ vs. 0.10 , $p = 0.007$). Overall ratings across both competencies were predicted by a unit weighted composite of numerical and innovation test scores with a Multiple R of 0.34, in line with observed validities for proctored reasoning (cognitive ability) tests as reported by various meta-analyses.⁵

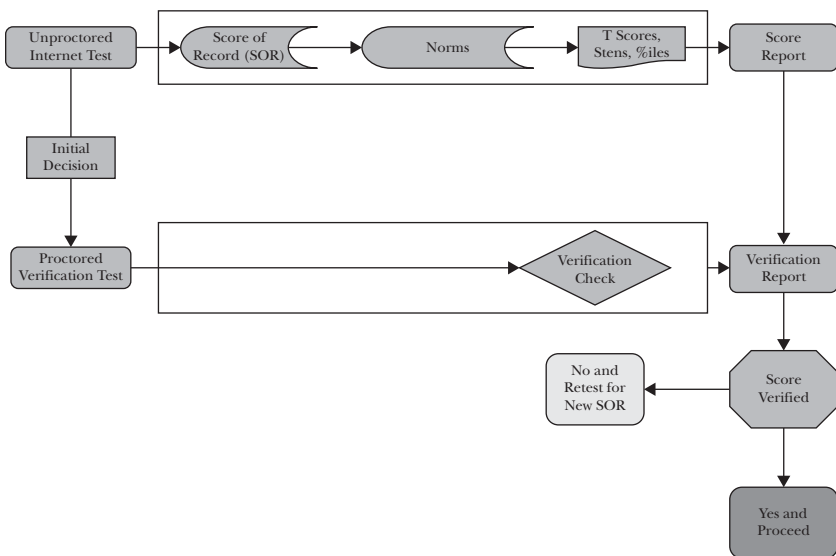
Extending the UIT Solution

Briefly, this initial solution has since been further developed to a stronger UIT model in which the UIT test, referred to in this model as the Verify Ability Test (VAT), supplies the score of

record (SOR), and in which a short proctored test, referred to as the Verify Verification Test (VVT), is used to validate UIT scores (Burke, 2008a, 2008b, & 2009; Burke, van Someren, & Tatham, 2006). This represents a shift away from what might be called a two-test solution in which a UIT is used to screen candidates, and that score is then discarded, followed by a subsequent proctored test that then supplies the SOR.⁶ This approach was developed based on two pieces of research: a survey conducted in 2005 of a range of clients to identify where they saw online assessment moving to over the next five years (further information on that survey can be found in Burke, 2006); and criterion validity studies coupled with a number of data forensic audits of UIT scores. See Figure 14.2 for an overview of the verification process.

The client survey showed that organizations, public and private sector, saw an increasing need to effectively outsource components of their talent acquisition and succession processes by moving to online tools, and saw that assessments including cognitive ability tests would have to follow that trend. Any solution to UIT was also seen by the majority as having to meet two requirements. The first

Figure 14.2. Schematic Overview of Verify Testing Process

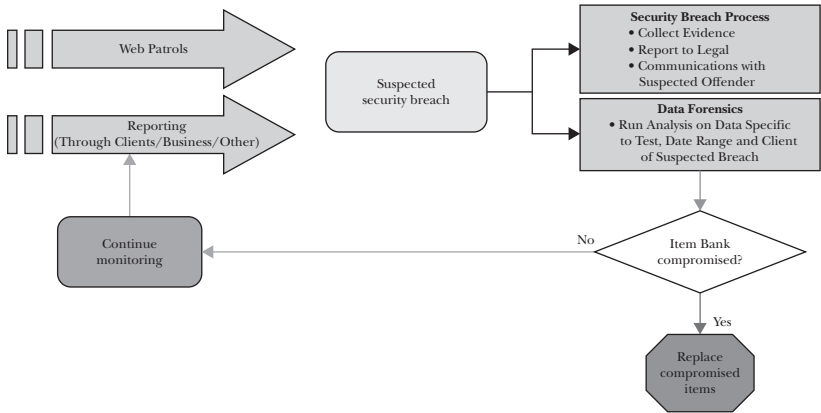


was a simple and consistent use of testing in which a single score would provide the basis for all decisions in a given process from the point at which a UIT was administered. For example, if a score was used earlier in the process for screening applicants, many clients surveyed wanted to have that score available for subsequent selection decisions so that decisions affecting any applicant at any stage in the process would have been based on test score information that could be shown to be consistent throughout that process. So a clear need expressed in the client survey was for a UIT solution that could be shown as demonstrating procedural justice⁷ (Gilliland & Hale, 2005).

The second requirement was that of managing the security over the SOR and, in essence, meeting the concerns that have been expressed in the article by Tippins, Beatty, Drasgow, Gibson, Pearlman, Segall, and Shepherd (2006) over piracy and cheating, and which will be briefly addressed next.

Following Impara and Foster's (2006) principles for the security of testing programs, the new process has features aimed at defenses against cheating prior to administering the UIT as well as features aimed at defenses during UIT administration. The former defenses rely on the design of the tests using the LOFT approach and multiple equivalent test forms assigned randomly to candidates registering for the first stage UIT in line with the guidelines set out by the International Test Commission for computer-based and Internet tests (ITC, 2006), as well as short and efficient verification tests used at a second proctored stage (these tests might be used at a very late stage of the overall process, for example, the last ten short-listed applicants of several thousand assessed at earlier stages, and take literally a matter of minutes to administer). These features of the testing process sit within an overall security framework that actively monitors security threats such as pirate sites as identified by web patrols (systematic web searches for test content and offers for coaching on the tests) as well as more reactive reporting of potential security incidents by candidates, clients, and third parties. See Figure 14.3.

One method of evaluating the efficacy of these actions and related to the first stage UITs is the application of data forensic (DF) algorithms. These algorithms search for a number of statistically unlikely patterns of test responses, such as similar patterns of

Figure 14.3. Overview of Security Framework

right and wrong answers suggesting collusion between applicants/candidates, aberrant scores where patterns of correct responding do not correspond to the applicant's/candidate's ability levels (for example, more difficult questions are answered correctly while easier questions are answered incorrectly), as well as fast responding (small latencies on questions) allied with correct answers indicating prior access to the answer key for questions (see Burke, 2006, for a reproduction of an article by Maynes as well as Maynes, 2006, for further details of these algorithms). As reported by Burke (2008c), DF analyses of the first stage UITs show low frequencies of abnormal and aberrant question responses and overall test scores. For a data slice of circa 30,000 in vivo UITs administered in 2007, the frequency of DF indices ranged from 0.003 percent (fast latencies and high question accuracy) to 1.12 percent (identical test responses though further analyses showed a large proportion of these tests were administered through different testing programs conducted in different geographical locations such as Australia and the UK which, allied with the LOFT model and different test content, indicate low likelihoods of shared UIT content between applicants). Overall, this DF analysis showed 2 percent of applicants to have one or more DF indices flagged as abnormal or aberrant, and this is typical of the regular audits conducted of Verify data on a quarterly basis. However, 2 percent of 100,000 applicants could suggest that 2,000 applicants might achieve scores

exceeding the cut-score levels set for various client testing programs. These statistics suggests that, while LOFT as the basis for the UITs is effective, they also warn against complacency and emphasize the need for additional security measures such as verification testing.

Ultimately, the value of security efforts will rely on whether they influence applicant and candidate perceptions of the testing process, and, in turn, their behavior when engaging in the testing process. The following is a verbatim record of a thread between job applicants obtained from a UK graduate site as identified from a web patrol conducted in early 2007 and cited in Burke (2009; all spelling is as per the thread identified):

“Posted at 12:09 I got 20 real SHL numerical test questions. If someone needed, send me email <email address given>. Only 20 pounds. If you want to pay me a little money, then push you to the assessment stage, just email me. Worth or not, you decide.

“Posted at 16:01 There are a number of reasons why everyone should ignore the original poster. Firstly I have taken a number of SHL numerical tests and by and large the questions are of a similar level but different. Yes, there may be some that are repeated but out of 20 questions there is not much. More saliently, lets assume that by some miracle all 20 questions come up. Yes, you will pass but will be retested at the next round . . . so a waste of money and time. I really hate it when cretins like <web name cited> take advantage of people. If she really wanted to help, she would have offered them for free. You have been warned . . .

“Posted at 17:23 I agree, what a stupid thing to advertise. 1. It is not probable that you will get these 20 questions when you take the test <this person then goes on to list another 9 reasons why the first posting is unhelpful becoming increasingly emotional about the posting and the person who posted it, hence the termination of the account of this thread at this point>.”

To add further support to this singular and anecdotal case, several criterion validation studies as described in the Verify technical manual show validities in line with those from meta-analyses of proctored cognitive ability tests (meta-analysis citations are as given earlier and operational validities for the VATs reported in

the technical manual are in the region of 0.39 to 0.50 from studies conducted in various industry sectors and job levels in the UK and the U.S.; see Burke, van Someren, & Tatham, 2006, for further details).

Accordingly, and while also acknowledging that test security can never be 100 percent and is a continuous area for improvement, the evidence gathered to date does suggest that the Verify model has gone some way to developing a UIT solution that uses the Internet and computer technologies as well as sound psychometric principles to offer measurable levels of test security. In 2008, the first author had the benefit of participating in the first test security summit sponsored by the Association of Test Publishers (ATP) at the ATP's annual conference in Dallas, Texas. While the following principles were drafted after the development of the solution, they serve as a summary of the key objectives of the program as well as the key features of the new process. Those principles are as follows (as taken from Burke, 2008b):

- **Enforcing** test security means
 - Actively managing intellectual property breach
 - Monitoring candidate behaviors
- By **identifying** test fraud through
 - Policing content and monitoring through critical incident procedures
 - Regular data audits to check for piracy, cheating, and item exposure
- And **preventing** test fraud through
 - Designing cheat resistance into the score of record

Meeting Other Challenges to Online Assessment: The Assessment Window, the Employer Value Proposition, and the Candidate's Experience

We will now move to a different set of challenges to online assessment which center around the candidate's experience of sitting an online assessment, and the experience gained with the third and fourth authors' organization in developing talent acquisition

programs for call centre staff. These programs had to address a number of stakeholder concerns, among which were:

- Internal stakeholder (such as recruitment and line manager) concerns about having sufficient assessment to evaluate an applicant's fit to roles while also avoiding over-length assessments that might drive talent to other employers (that is, that the assessments would require so much effort from an applicant that he or she would abandon the process and look for employment opportunities elsewhere). This balance is what the first author refers to as the "assessment window."
- Industry regulator guidelines and expectations for the organization in addressing customer needs and experiences of the services provided to them.
- The simple fact that applicants might also be current or potential customers of the organization and their perceptions of the online assessment process would influence their general perceptions of the organization.

We will address these concerns by first describing a modular and criterion-driven approach to assessment design to mitigate concerns over applicant fatigue in sitting online assessments (that is, how much assessment is required to obtain a valid but efficient solution) and then by describing the context of regulator expectations and how the assessment solution maps to the organization's actions to ensure that its talent management processes address those regulator expectations.

The specific project was to develop an online assessment solution for call centre operatives encompassing three roles: inbound customer service agents dealing with customer inquiries; outbound customer service agents offering products and services to customers; and premier agents servicing higher value customers. The solution sought had to meet the needs of selecting to each of these three different roles while also providing data on where an applicant would best fit should they meet the minimum requirements for any of the three roles. The solution was developed using the criterion-centric approach as described by Bartram (2005) and Burke and Bateson (2009). Essentially, the first step was to define critical behaviors required

in each role using the Universal Competency Framework (UCF) taxonomy of behaviors (Bartram, 2006). This hierarchical taxonomy offers 112 behaviors organized into twenty dimensions which in turn are organized into eight factors. The structure of the UCF reflects its organization around construct and criterion validation studies through which empirical links have been established between criterion behaviors as classified by the UCF and predictors of those behaviors, such as cognitive ability, personality and motivational constructs, and measures.

Competency analyses with operations supervisors and managers coupled with a review of the organization's competency profiles for the three customer service roles identified six UCF dimensions as critical to effective performance across the three roles. These were:

- Adhering to Principles and Values (respect for and adherence to organizational values and respect for others, and mapping to the Big 5 construct of Agreeableness)
- Persuading and Influencing (mapping to the Big 5 construct of Extroversion as related to engaging others)
- Analyzing (problem-solving mapping to the deductive reasoning facet of general mental ability)
- Delivering and Meeting Customer Expectations (mapping to the dependability facet of Conscientiousness)
- Following Instructions and Procedures (again mapping to the dependability facet of Conscientiousness)
- Achieving Goals and Objectives (mapping to the achievement orientation aspect of Conscientiousness and to Need for Achievement)

Cast in the context of the organization's own competency language, as well as the wider regulator context to be described in more detail below, the UCF profile just described can be summarized as *the sort of person who will tend to be results orientated, energetic and competitive but persuasive; flexible, participative, and modest; who will be structured, conscientious, and detail conscious; who can be relied on to be punctual, dependable, and customer focused; and who can sell through service.*

While this may seem a demanding set of requirements, the behaviors and associated applicant characteristics to be captured by the assessment process can be relatively easily translated using the UCF framework into requirements for cognitive ability, dependability, and a targeted set of personality and motivation scales. The next step in the design of the solution was to develop a modular suite of assessments targeted on the criterion behaviors. In effect, the basis of the design was a set of composite predictor scores reflecting the criteria against which successful applicants would be subsequently judged by operational supervisors and managers in terms of job performance. In short, the final online solution combined elements of cognitive ability (short versions of Verify verbal and numerical tests), a dependability questionnaire (the Dependability and Safety Index; Burke & Fix, 2009), a short bespoke situational judgment test (using the knowledge format for SJTs), and a short personality questionnaire (containing targeted scales derived from the UCF framework and using a forced choice format to manage potential faking). In total, the assessment suite takes thirty minutes to administer.

The question that such an approach might raise is that of the reliability, and therefore the accuracy, of such a short series of measures. However, and as argued some time ago by Cureton (1950), this is only a concern if internal consistency is chosen as the index of reliability, and, while it may be a necessary condition, reliability is not a sufficient condition for validity. As argued by Burke and Bateson (2009), the fixation on internal consistency has been a limitation in efficient test design and ignores alternative approaches to reliability and measurement error (such as stability or test-retest, alternate forms as well as generalizability theory). Modern test theory in the form of IRT models applied to both cognitive and non-cognitive measures (for example, Brown & Bartram, 2009) shows that more efficient assessments are possible once the hegemony of internal consistency is rejected as a basis for evaluating the quality of an assessment solution. However, as in the case of the solution developed for call centre operatives described here, the design of an efficient solution must be based on a clear, strong, and validated theoretical framework (in this case the UCF) in line with Messick's model for validity.

The third step in the design of this particular solution for call centre operative assessment was a criterion validation study with three hundred operatives in all roles, one half of the sample comprising a concurrent validation with existing employees and one half comprising a predictive study with employees selected using the assessment solution (this study was conducted in two phases, with the first phase being the concurrent study). Whether the data were collected using a concurrent or predictive design was not found to moderate the predictor-criterion relationships significantly, and data across the two designs (concurrent and predictive) were pooled for final analysis. The criteria used for collecting supervisor and manager ratings of performance used behavioral items from the earlier UCF competency analysis (that is, as well as specifying the predictors, the UCF analysis was also used to define criterion measures in line with the criterion-centric approach), and a factor analysis of these ratings identified three performance dimensions:

- A conscientiousness and dependability factor labeled Results Orientation capturing behaviors such as “checks work thoroughly,” “ensures that work is accurate,” “sticks to company regulations,” “follows supervisors instructions,” and “adheres to company work methods”
- A hybrid extroversion and achievement factor labeled Persuasive capturing behaviors such as “can easily sell an idea or a proposal to others,” “negotiates well,” “demonstrates enthusiasm,” and “works energetically to achieve goals”
- A second hybrid extroversion and achievement factor labeled Action capturing behaviors such as “can make decisions under pressure,” “acts on own initiative,” “communicates clearly,” and “adapts communication style to suit different people”

These factors were themselves validated against performance targets in the three roles, showing the first factor to be the most substantive dimension in performance ratings for the inbound role, the second factor as the most substantive dimension in performance ratings for the outbound role, and all three factors as significant in performance ratings for the premier role. Specific criterion-predictor relationships were identified in line with the

expectations from the UCF analysis conducted in the first stage of the solution's design (the validation design followed a confirmatory approach with statistical analyses using directional, one-tailed hypotheses to evaluate the results): the Results factor was correlated 0.41 with a composite of scores from the dependability measure, SJT, and targeted personality scales; the Persuading factor was correlated 0.41 with a composite of scores from cognitive ability, SJT, and targeted personality scales; and the Action factor was correlated 0.29 with a composite of scores from cognitive ability, SJT, and targeted personality scales. The overall composite (unit weighted) of all predictors correlated 0.43⁸ with supervisor and manager ratings of overall potential of participants in the study (as rated on omnibus items such as "How well do you see this person progressing in their current role?" and "What do you see as this person's potential for progression to a more senior role?").

The results from this validation not only supported the value offered by the solution in terms of selection to each specific role, but also provided the basis for designing a simple placement system based on composite scores from the assessments mapped to the three behavioral factors identified as underpinning supervisor and manager ratings of performance. Accordingly, minimum cut-scores were set for each of the three factors to provide a basis for initial screening of applicants. Fit scores using the three factors were then used to indicate roles for which the applicant had highest person-job fit. These fit scores helped to manage subsequent steps in the organization's selection process by suggesting which roles applicants should be interviewed for at a later stage.

The data generated by a study such as this can, of course, provide information to brief stakeholders on the value from such a process as well as how the constraints in terms of the assessment window were met, while also maintaining relevance and empirical validity against performance dimensions. However, organizations operate in a wider economical and political context, and assessments have to be shown to be relevant to that context as much as to the more usual metrics used to evaluate the utility from assessments. In the UK, a significant influencer in the wider context of banking is the Financial Services Authority (FSA). Of particular relevance to the solution just described for call centre

operatives is the FSA's Treating Customers Fairly (TCF) initiative (FSA, 2009, although the FSA materials used as part of the evaluation of the assessment solution were obtained in 2007).

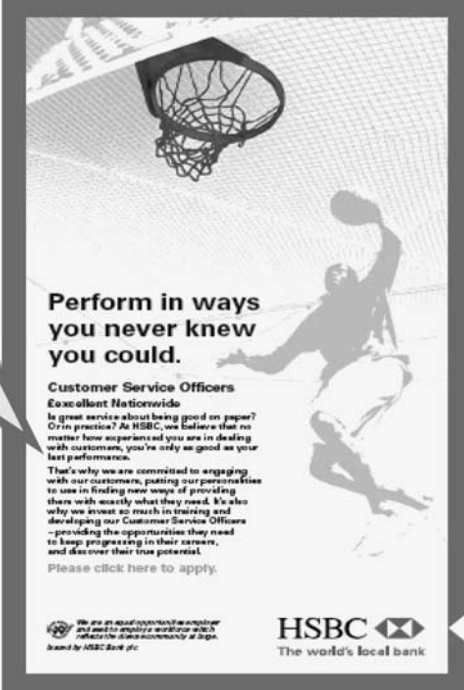
This initiative sees the culture of financial organizations as critical to the fair treatment of customers and sets out clear expectations as well as outcome measures for how organizational leadership and culture will be evaluated by the FSA. For example (and working from the 2007 FSA materials), "A firm can influence the delivery of fair consumer outcomes by *recruiting staff with appropriate values and skill*"; [as an example of good practice] "The firm ensured that all new staff, regardless of their previous sales experience, had *similar TCF values to their own.*"; [describing expectations for a culture framework to be demonstrated by a financial organization] "Management make positive *behaviors and attitudes* to the fair treatment of customers *a key criterion in the selection of staff.*"

In response to these expectations, the third author's organization had reviewed and revised its customer service framework (CSF) and incorporated these internal standards into its performance and competency measures for front-line customer staff. Examples of the behaviors flowing from this TCF orientated review into front-line role competencies include *continually enhances own skills and knowledge; rapidly learn new tasks and demonstrates understanding of new information; accurately records information, follows procedures and delivers quality results; checks work thoroughly and is concerned about the needs of others; exercises effective judgment when dealing with customer facing situations; takes ownership and shows initiative; understands others and makes an effort to help them; builds constructive relationships; perseveres to get things done.* These behaviors form the foundation of the employer value proposition (EVP) reflected in the organization's recruitment advertising, and represent a critical frame of reference against which the relevance of the assessment solution has to be evaluated and within which applicants will themselves evaluate the relevance of the assessments they are asked to sit in seeking employment with the organization. See Figure 14.4.

The following summarizes the components of the assessment solution in the context of the organization's EVP through which each assessment component was mapped to the organization's CSF to demonstrate relevance to expectations set out by the

Figure 14.4. Messaging in Recruitment Advertising Exemplifying EVP and CSF

.... We are committed to engaging with our customers, putting our personalities to use in finding new ways of providing them with exactly what they need. It's also why we invest so much in ... providing the opportunities [to our Customer Service Officers] to keep progressing their careers and discover their true potential



Perform in ways you never knew you could.


Customer Service Officers
Excellent Nationwide

Is great service about being good on paper? Or in practice? At HSBC, we believe that no matter how experienced you are in dealing with customers, you're only as good as your last performance.

That's why we are committed to engaging with our customers, putting our personalities to use in finding new ways of providing them with exactly what they need. It's also why we invest so much in training and developing our Customer Service Officers - providing the opportunities they need to keep progressing in their careers, and discover their true potential.

Please [click here](#) to apply.

We are an equal opportunity employer and need to employ a workforce which reflects the diversity of our customers. [Apply to HSBC Bank plc](#)

HSBC 
The world's local bank

FSA's TCF initiative (the reader will hopefully see the links with the three performance dimensions of results orientation, persuasive, and action described earlier):

- Cognitive ability—continually enhances own skills and knowledge; rapidly learn new tasks and demonstrates understanding of new information
- Dependability—accurately records information, follows procedures and delivers quality results; checks work thoroughly and is concerned about the needs of others
- SJT—exercises effective judgment when dealing with customer-facing situations; builds constructive relationships
- Personality scales—takes ownership and shows initiative; understands others and makes an effort to help them; perseveres to get things done

In Conclusion

In the course of this contribution, we have attempted to share experiences in the development and implementation of online assessment solutions that characterize how the field has developed over the past decade as technologies such as the Internet have themselves developed and driven client processes and needs. Going online with assessment clearly presents challenges to the science of assessment and specifically to the secure administration of cognitive ability tests, but this is now the technology through which our personal transactions are largely conducted, and this is the technology that forms the foundation for most business processes worldwide, including those managed by HR functions. In addition to technical challenges in the form of assessment design and delivery, the increasing pervasiveness of Internet technologies in talent management also highlights the need to address issues of communication and relevance to a broad range of stakeholders that extends beyond the traditional client in the form of an organization's HR function. Internet-based assessment increases the visibility of assessment processes to potential customers as well as public bodies concerned about wider industry best practice.

As well as challenges, the growth of technology in talent management is also providing increased opportunities for assessment as organizations recognize its value in addressing rapid economic and demographic change. The following quote from a recent report published by the Aberdeen Group, a leading research and analytics group focused on the value from technologies, serves to demonstrate the opportunities for innovations drawn from our science and practice in meeting the challenges of Internet based assessment:

“With rising unemployment and increased uncertainty, the talent market is expanding rapidly with active job prospects. Although budgets for staffing are down and hiring freezes seem commonplace, the rapidly expanding talent pool makes screening and selection even more critical, as organizations must ensure that they are interviewing and placing the best candidates in terms of skills, behaviors, and cultural fit. . . . Another macro pressure that requires attention is the changing demographics within the

workforce and labor market, and the implications of changing work expectations. Globalization is a reality that most organizations face. In addition, there are now four generations of workers in the labor market. As a result, organizations must be keenly aware of cultural and demographic nuances of employees and job prospects. As such, organizations are placing greater emphasis on hiring and placement decisions based on “fit”—which includes elements such as attributes and behaviors. . . . Assessments provide the intelligence necessary to make decisions by ascertaining alignment with long-term objectives, thus, best positioning an organization for strategic growth.” (Saba, Martin, & Madden, 2009)

Notes

1. We are indebted to our colleague Professor Dave Bartram for flagging a distinction in terminology in which an “applicant” can be distinguished as someone engaged in an assessment for a job vacancy (usually high-stakes settings such as selection and promotion), while “candidate” is more widely used to refer to anyone taking an assessment in an employment setting (which could be a low-stakes setting such as a development program).
2. This relates to an incident in which a CAT version of the Graduate Records Exam (GRE) was compromised within a matter of a few weeks by a coaching company in the United States, and this experience led to the development of such frameworks as the linear-on-the-fly-testing or LOFT model, which was used in developing the solution for the client.
3. These estimates are uncorrected for any artifacts such as range restriction and measurement error in criterion measures.
4. Correcting for both range restriction and measurement error in both scores yields an estimated operational validity between the UIT and proctored scores of 0.88.
5. Note that Multiple R’s have not been corrected for range restriction or for interrater consistency, and therefore underestimate the true relationships between UIT scores and interview ratings. For example, correcting for an interrater consistency of 0.7 and internal consistency reliabilities of 0.8 in the two test scores, the true correlation between the composite of interview ratings on judgment and innovation and the composite of the two UIT scores would be estimated to be 0.45 prior to any corrections for range restriction due to prior selection on the UIT test scores.
6. Please note that, at the time of writing, the second author’s organization still uses the UIT numerical and inductive reasoning tests to

screen prior to proctored tests administered as part of their graduate entry assessment centers.

7. Procedural justice relates to whether a process is seen as offering a fair opportunity for participants in that process to demonstrate their suitability for a position or role. The accuracy and stability of an instrument allied with strong validity evidence, criterion and construct, are critical elements of the scientific evidence supporting positive perceptions of procedural justice. Also, evidence that shows that an instrument functions equally well for different candidate groups and that it is free from any biases in its content and scoring is also important in supporting positive perceptions of procedural justice.
8. All correlations have not been corrected for the effects of range restriction or unreliability of the criterion.

References

- Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203.
- Bartram, D. (2006). *The SHL Universal Competency Framework*. SHL, UK: Retrieved November 30, 2009, from www.shl.com/OurScience/Documents/SHLUniversalCompetencyFramework.pdf.
- Brown, A., & Bartram, D. (2009, April 2–4). Doing less but getting more: Improving forced-choice measures with IRT. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, Louisiana.
- Burke, E. (2006). *Better practice for online assessment*. SHL, UK. Retrieved July 28, 2008, from www.shl.com/SHL/en-int/Thought_Leadership/White_Papers/White-Papers.aspx.
- Burke, E. (2008a, April). Preserving the integrity of online testing. In N. T. Tippins (Chair), *Internet testing: Current issues, research solutions, guidelines, and concerns*. Symposium conducted at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, California.
- Burke, E. (2008b, March). Dealing with the security of online employment testing. Case study presentation at the Association of Test Publishers Test Security Summit, Dallas, Texas.
- Burke, E. (2008c, July). Applying data forensics to defend the validity of online employment tests. Paper presented at the Conference of the International Test Commission, Liverpool, UK.
- Burke, E. (2009). Preserving the integrity of online testing. *Industrial and Organizational Psychology, 2*, 35–38.

- Burke, E., & Bateson, J. (2009, April). Technology assisted test construction, delivery and validation. In J. Weiner (Chair), *Technology-based assessment in the 21st century: Advances and trends*. Symposium conducted at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, Louisiana.
- Burke, E., & Fix, C. (2009). *Dependability and Safety Index: Technical manual* (2nd ed.). Thames Ditton, UK: SHL.
- Burke, E., van Someren, G., & Tatham, N. (2006). *Verify range of ability tests: Technical manual*. Thames Ditton, UK: SHL.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement*, 10, 94–96.
- Davey, T., & Nering, M. (2002). Controlling item exposure & maintaining item security. In C. G. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Drasgow, F., Nye, C. D., Guo, J., & Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology*, 2, 46–48.
- Financial Services Authority. (2009). *Treating customers fairly*. Retrieved November 30, 2009, from www.fsa.gov.uk/Pages/Doing/Regulated/tcf/index.shtml.
- Fleishman, E. A., & Reilly, M. E. (1992). *Handbook of human abilities: Definitions, measurements, and job task requirements*. Palo Alto, CA: Consulting Psychologists Press.
- Gilliland, S. W., & Hale, J. (2005). How do theories of organizational justice inform fair employee selection practices? In J. Greenberg & J.A. Colquitt (Eds.), *Handbook of organizational justice: Fundamental questions about fairness in the workplace*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hollinger, R., & Lanza-Kaduce, L. (1996). Academic dishonesty and the perceived effectiveness of countermeasures. *NASPA Journal*, 33, 292–306.
- Impara, J. C., & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.
- International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing*, 6, 143–172.

- Maynes, D. (2006). *Recent innovations in data forensics—2006*. Retrieved November 30, 2009, from www.caveon.com/articles/df_innovations06.htm.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*, 741–749.
- Messick, S. (1998). Alternative modes of assessment, uniform standards of validity. In M. D. Hakel (Ed.), *Multiple choice: Evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology*, *74*, 441–472.
- Saba, J., Martin, K., & Madden, K. (2009, March). *Assessments in talent management: Strategies to improve pre- and post-hire performance*. Boston, MA: Aberdeen Group.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Pierre, J. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology*, *88*, 1068–1081.
- 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.
- Schmidt, F. L., Shaffer, J. A., & In-Sue, O. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, *61*, 827–868.
- Tippins, N. T. (2008, April). *Internet testing: Current issues, research solutions, guidelines, and concerns*. Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, California.
- Tippins, N. T., Beatty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, *59*, 189–225.
- Weiner, J. (2008, July). The potential impact of cheating in online testing: Good news, bad news. In E. Burke (Chair), *Understanding the impact of cheating and content piracy on test scores and decisions based upon them*. Symposium presented at the Conferences of the International Test Commission, Liverpool, UK.

Chapter Fifteen

IMPLEMENTING COMPUTER ADAPTIVE TESTS

Successes and Lessons Learned

Mike Fetzer and Tracy Kantrowitz

Organizations receive more applications for job opportunities than ever, and as a result, the identification of cost- and time-efficient processes to help determine top applicants for critical positions is of paramount importance. Pre-employment testing effectively identifies applicants who are well suited to positions, but recent economic conditions have forced organizations to slot testing earlier in the hiring process to help reduce the more time- and resource-intensive phases of hiring. As a result, unproctored (unsupervised) Internet testing (UIT) has emerged as mode of administration that brings multiple advantages to organizations, including decreased time-to-fill and recruitment costs (Beaty, Dawson, Fallaw, & Kantrowitz, 2009). UIT presents its own set of challenges, however, including increased exposure to test content, increased opportunity for cheating, and unstandardized test environments.

Identification of innovative methods for assessing applicants has been critical to helping support organizational trends while maintaining the integrity and security of selection processes.

A major advancement in pre-employment testing is computer adaptive testing (CAT), which combines science and technology to deliver a more targeted and secure testing experience. Computer adaptive testing (CAT) provides the maximal balance of accuracy and efficiency. Not to be confused with computer-based testing (a term that refers to *any* type of test administered using a computer), CAT is a method of testing that “adapts” to each individual test-taker. In other words, CAT provides a tailored testing experience based on the test-taker’s level of knowledge, skill, ability, or other (KSAO) characteristic being evaluated by the test. As a result, a CAT requires fewer items and produces a more accurate score than traditional “static” or randomly generated tests. CAT also presents a number of other advantages to recruiters and hiring managers, including reduced testing time and increased reliability (compared to “static” test equivalents). These key advantages make CAT a more appropriate alternative to UIT programs than traditional or static assessments.

Over the past few decades, CAT has been used extensively in the areas of education, certification, and licensure. There are over thirty large-scale CAT programs around the world that evaluate an estimated four to six million people each year (Fetzer, 2009). Recently, the benefits of CAT have started to be realized in the area of personnel selection (see discussion of sample CAT programs by McCloy and Gibby, Chapter 5 in this volume). Both public- and private-sector organizations have traditional as well as newly emerging needs with regard to assessment that can be addressed with CAT.

This case study focuses on the development and implementation of an approach to pre-employment testing that focuses on CAT. We present CAT as a practical and effective solution to the challenges of UIT. Specific examples and descriptions of several types of computer adaptive tests are provided as part of this case study. In addition, an innovative approach to confirmation testing using CAT technology is described below. For a more in-depth discussion of CAT, we direct the reader to Chapter 5.

This case study discusses the implementation and use of CAT within a public sector agency—the human resources department of Riverside County, California. The department supports a workforce of more than eighteen thousand employees across

fifty departments and agencies. Riverside County is the fourth-largest county in California and is home to more than two million residents.

Organizational Challenges

The most significant challenge for the department was the need to move to a more unproctored testing model. The department had established large-scale, proctored testing centers devoted exclusively to testing applicants for job openings in the county in 2002. The time and resources required to run its various proctored testing centers were extensive, and, starting in 2008, looming budget cuts were threatening their existence. In addition, the department wanted to “cast a wider net” in order to recruit applicants from outside its immediate area. Anticipating the need to overhaul the current testing process, the department forged a relationship with a SHLPreVisor to architect and validate a new method of assessing applicants using CAT.

Historically, the department routinely included static assessments of cognitive ability, personality, and hard skills as a part of its assessment process. Moving to an unproctored model presented some inherent challenges, especially with regard to test security. In general, the jobs it recruits for are highly sought after, which increases the risk of applicants attempting to cheat by obtaining questions and/or answers prior to their scheduled testing sessions.

A second challenge was the need to further distinguish among highly qualified applicants. As the economy worsened, its applicant pools grew—often by a larger portion of applicants with higher levels of KSAOs than was typical for most jobs. Thus, a considerable percentage of applicants were “maxing out” on certain tests (obtaining the highest possible score), which provided more qualified pools of applicants from which to select but created a situation in which the test scores among those applicants were no longer useful in the decision process.

Finally, the success of the current testing program within the department was a double-edged sword. Because the tests were useful tools for evaluating applicant qualifications, the department’s internal customers wanted more tests administered to

incoming applicants. This, of course, resulted in an increase in the amount of testing required for each applicant. The fact that each applicant was required to complete more assessments caused some concern among the major stakeholders with regard to potential applicant reactions. In addition, some internal customers were quite satisfied with the traditional testing batteries put into place as part of the proctored testing center program. They felt the current process yielded good results and needed hard evidence that changing test batteries was warranted.

Even in isolation, these challenges were significant to say the least. In combination, they presented the department with a problem that didn't seem to have an easy (or even moderately difficult) solution, or so they thought. At about the same time the department was faced with these challenges, SHLPreVisor was in the process of developing and validating computer adaptive versions of cognitive ability and personality assessments to supplement its line of computer adaptive knowledge tests. The timing was perfect, and the agency formed a partnership with us in order to implement these leading-edge assessments and address all of its challenges with one solution.

Technology-Enhanced Solution: Computer Adaptive Testing

A key element of the department's testing program was a "whole person" approach, utilizing cognitive ability, personality, and skills tests in combination to generate a more accurate evaluation of applicant qualifications. With the advent of CAT versions of cognitive ability and personality assessments, the department was able to continue to leverage the whole person approach and extend that model to an unproctored setting. Further, new methods of confirmation testing based on CAT technology were implemented in order to supplement the unproctored cognitive testing phase to mitigate the potential for cheating. The selection process at the department leveraged the "whole person" approach by making selection decisions based on overall scores at the unproctored, screening phase and the proctored, selection phase. That is, composite scores were created at each stage in the process that represented a combination of hard skills,

cognitive ability, and personality, and top-down selection decisions were made.

CAT-Based Hard Skills Tests

The first phase of this project focused primarily on identifying an alternative to hard skills (knowledge) testing the department had been conducting in its testing centers. For many years, the department used simulation-based or static assessments of knowledge and skills (for example, computer literacy, software knowledge). As part of its library of hundreds of CAT tests, SHLPreVisor offered CAT-equivalent tests the department could utilize as suitable replacements. Having already established the job relevance of these knowledge and skill areas through internally conducted job analysis processes, scores from these tests enabled hiring managers to quickly and easily determine applicant qualifications to supplement the resume review and/or interview. Because these tests were computer adaptive, they could be administered unproctored and thus allowed for culling down the applicant pool to only those who met or exceeded basic qualifications before bringing them onsite for more resource-intensive steps in the hiring process.

CAT-Based Cognitive Ability Tests, with a New Twist

Given the breadth of jobs and levels the department was responsible for, they chose to implement three different CAT-based cognitive ability tests: Quantitative Skills, Verbal Proficiency, and Deductive Reasoning. These three tests were designed to tap into different facets of cognitive ability and are administered individually or in various combinations, depending on the target job.

Each test consists of large “pools” of several hundred items that span the range of difficulty level, with a higher concentration of items in the average difficulty range. Constructing pools in this manner reduces item exposure and thus enhances the security of the test, since items of average difficulty will tend to be administered to applicants more often than will high- or low-difficulty items. In addition to the active (scored) items, unscored items are also included in the pools in order to collect data on

these items so that they may eventually be activated. This process ensures that the test content is continually refreshed.

The cognitive tests incorporate a stopping rule that terminates each test when a standard error of .38 (or below) is reached. This standard error stopping rule is roughly equivalent to ending the test when a “static” internal consistency reliability of .85 is met for each person, a reliability coefficient typically recommended for selection tests (Gatewood & Feild, 2007). As some individuals take longer than others to reach this standard error threshold, a second stopping rule is used in conjunction with the standard error stopping rule. If an applicant does not reach the .38 standard error rule by the time he or she completes twenty items for the Deductive Reasoning and Verbal Ability assessments, or thirty items for the Quantitative Skills test, the test ends and his or her current theta (ability) score is calculated. These rules were put in place to limit the potential testing time to be a reasonable length. Thus, as it is possible that a few applicants will not meet the standard error stopping rule within the maximum number of items, the overall test reliabilities may vary slightly from the .85 value stated above.

Use of reliability as a stopping rule introduces the notion of “controlled reliability” and greatly improves the accuracy of CATs over traditional/static tests. This control results in more reliable test scores over a broader range of ability levels, and is especially useful in hiring situations when finer distinctions among top performers are critical to making the right hiring decision. The stopping rules can be modified as needed, depending on the needs of the particular department or agency, but they generally don’t deviate from the default settings.

Given the large item pools and adaptive nature of these cognitive tests, they were particularly well-suited to help solve the department’s main concerns around the potential for cheating in unproctored environments. The variable nature of the testing experience greatly reduced the ability for an individual or group to obtain copies of all (or even a significant portion) of the items. However, there was still the risk that an applicant taking the test from home could enlist the aid of his or her smart(er) friend or relative to help answer the questions or complete the entire test in his or her place.

The department needed a method to confirm the unproctored scores but did not want to subject applicants to a second full-length cognitive test once they were onsite. To address this need, we had also developed a unique method for confirmation testing that utilized the CAT technology. This confirmation method utilizes the applicant's final ability score collected at the completion of the unproctored session and uses it to determine the starting point for the short, adaptive assessment completed onsite. While the test-taker is completing the onsite assessment, the algorithm quickly converges on the applicant's ability score based on precision-based stopping rules if consistency is found in his or her responses. If the item response information is inconsistent, the system administers additional items until a consistent pattern emerges and a reliable, valid score can be reported. This innovative confirmation process greatly enhances the security of the unproctored cognitive tests as it leverages completely separate item pools.

CAT-Based Personality Assessment

SHLPreVisor's computer adaptive personality assessment is a general assessment of normal adult personality with a focus on workplace applications. This assessment consists of thirteen separate scales that correspond to thirteen dimensions of personality (see Table 15.1). Through the use of CAT technology and an alternative item type (forced-choice) the assessment not only adapts to the applicant's trait level, but greatly reduces item exposure and potentially reduces faking. As with the cognitive tests, the department utilizes the individual personality scales in various combinations, depending on the target job requirements.

The item response theory (IRT) model operationalized is based on the ideal point paired comparison approach defined by Zinnes and Griggs (1974) and extended by Stark and Drasgow (1998). Applicants select which of two statements representing different levels of a personality trait are more descriptive of them. They are then presented with two additional statements, selected using an updated trait level estimate based on their previous responses. Sequences of statement pairs are selected in a manner that maximizes item information at each step. The adaptive

Table 15.1. Computer Adaptive Personality Scales

Achievement	Innovation
Collaboration	Influence
Composure	Reliability
Flexibility	Self-Development
Independence	Sense of Duty
Confidence and Optimism	Thoroughness
	Sociability

personality assessment is supported by more than 2,500 personality statements (approximately two hundred per trait), thus the probability that any two applicants will receive a similar test is quite low indeed. The assessment uses precision-based stopping rules to determine when sufficient information about a test-taker is gathered; the stopping rule is roughly equivalent to an internal consistency reliability of .85.

Implementation and Maintenance

The implementation of the suite of computer adaptive tests was relatively straightforward, but it was not without its challenges. As noted previously, the department had historically utilized a battery of static primary skills and knowledge tests, given in a proctored environment at various testing centers throughout the county. This process was well engrained across the various agencies and departments, and the change to new tests given in an unproctored environment met with some resistance.

The primary challenge originated from the simple fact that the department was moving to a new set of tests. As with any organizational process change, stakeholders may have been resistant simply because it required moving from the old and familiar to the new and unfamiliar. Hiring managers were used to utilizing the scores from the previous tests to make hiring decisions, and the move to new tests required a shift in their decision-making process in order to appropriately leverage the scores from the new tests. In addition, many hiring managers felt the

previous tests were very effective and thus did not quite understand the need to make the transition to adaptive testing.

This challenge was addressed by explaining the need for (and benefits of) unproctored testing. Internal industrial/organizational psychologists at the department spearheaded efforts to communicate the new testing approach and its benefits. This information was relayed to hiring managers and department heads as part of the concurrent validation process through presentations and other internal communication methods (such as email, intranet). Many stakeholders were impressed by the fact that the department was utilizing leading-edge testing technology to provide a broader pool of more qualified applicants in a shorter period of time. Although some may have been resistant at first, the rationale behind the new testing process was quickly adopted and accepted as the new status quo.

The second challenge involved determining the extent to which the computer adaptive cognitive ability and personality tests were job-relevant. This was addressed through a large-scale validation effort that utilized incumbent test data and job performance criteria across three broad job families. Entry-level jobs were categorized into two families (Clerical I and Clerical II), and the third job family encompassed professional/individual contributor jobs. Across all three job families, data from 668 employees were analyzed to determine the validity of the scores generated from the computer adaptive cognitive and personality measures.

The results indicated that the three cognitive ability measures and various combinations of scales from the adaptive personality test (chosen based on individual scale validities and then combined through a weighted composite approach) were significantly correlated with the criteria. Observed correlations between individual test/scale scores and certain job performance criteria ranged from the low teens to the high .20s (not surprising due to the range of predictors and criteria), and observed correlations between predictor and performance composites ranged from the mid .20s to the high .30s. When corrected for criterion unreliability, correlations among composites were in the .30s and .40s.

Despite the positive outcomes, the validation process itself was fraught with challenges. The departments had been undergoing

major budget cuts or reductions in force (RIF), and there was constant concern that entire groups or departments would be cut with short notice. It made collecting data during this sensitive time particularly difficult, even though communication relayed to stakeholders and employees (done through managers, emails from HR, and emails from department executives) made it clear that the purpose of the study was for research only. Participation in the study was voluntary for each department. The largest department in the county chose not to participate—they just did not see the value and thought it would require too many resources.

The final challenge involved making the shift from a proctored testing process to one that was unproctored. Historically, applicants came to one of the proctored testing centers and were administered the appropriate test(s), depending on the jobs they were applying for. In order to move to an unproctored testing process, online access to the tests had to be provided as part of the application process. This was accomplished through the department's careers website, which enabled 24/7 access for applicants to complete both the application and the appropriate test(s). Since the new CATs required no special software or hardware, any applicant with Internet access could complete the process at his and her convenience, and results were immediately available for review. Applicants without Internet access were still able to apply and test onsite, in order to accommodate everyone.

In order to further enhance the security and integrity of the testing process, the department has implemented several safeguards. First, applicants are discouraged from submitting multiple applications for the same job through the use of unique identifiers and warnings about potential disqualification. For example, unique identifiers are created by a combination of information (for example, birth date and last four digits of Social Security number). Thus, applicants who submit multiple applications/tests for the same job (or those who attempt to do so by slightly modifying their information) are flagged for review if the unique identifiers match.

Second, the department is leveraging the new confirmation testing process described previously on a random basis, enabling them to confirm test scores from the unproctored test session.

In other words, once applicants are brought onsite, they may be asked to complete the proctored confirmation test. Scores from this test are then used to make decisions about progression to the next stage in the hiring process. This process is being used randomly to balance checking applicants' identities and scores with administrative costs.

The new tests, once configured for the appropriate job families, are relatively maintenance free for the department. The platform through which the tests are delivered utilizes the "software as a service" (SaaS) model, thus all updates are handled by SHLPreVisor without the need for resources from the department. These updates include periodic review of test scores, updating item pools with new active and unscored items, and system-generated reports of test activity/usage. All the department has to provide are computers with Internet connections for the random proctored testing.

Success Metrics

The implementation of the new CATs was a success on several levels. First, the move from proctored to unproctored testing was a tremendous savings in terms of departmental resources. First and foremost, the department was able to close a full-time testing center staffed by thirteen people. This resulted in an annual cost savings of over \$500,000. Second, the department was also able to greatly reduce time to hire, since the time required to schedule and administer testing was reduced to zero. As a prime example, the department held a one-day hiring event for which the top-ranked applicants (based on their unproctored test scores) were invited for interviews. Each of these top-ranked applicants was guaranteed at least three interviews, which resulted in nearly all applicants attending the event. At the end of the day, the department had filled nearly 70 percent of the open positions, and the interviewers and hiring managers were extremely impressed with the caliber of applicants in attendance.

Third, the feedback received from many departments/agencies in the county indicated a universal increase in quality of applicants in general since the inception of the unproctored testing process. This has largely been attributed to the department's

ability to access the passive job-seeker market. Specifically, the previous hiring process, involving mandatory testing at a proctored testing center, greatly reduced the ability for passive job seekers to apply and complete the testing process because the testing centers were only open during normal work hours.

The validation results indicated the tests were predictive of job performance, thus providing substantial utility to the department (and Riverside County) in terms of increased employee productivity. In addition, the validation results were utilized to sell the new program internally, as stakeholders were able to see first-hand how effective the tests are in identifying top applicants. The studies also provided the necessary support from a legal perspective, a critical piece of the puzzle for a public-sector agency. The department proposed the new CAT model to stakeholders by focusing on validation data and the many inherent benefits of unproctored computer adaptive testing to demonstrate how test processes and outcomes could be significantly improved.

Finally, the department ran some analyses to determine whether or not the move to unproctored testing would result in a greater risk for cheating. Test scores from both proctored and unproctored testing conditions were compared over a one-year period, and no significant differences were found. Thus, the steps they put in place to further enhance test security, combined with the adaptive nature of the assessments, appear to be successful in deterring potential cheaters. The department conducted proctored confirmation assessments for several candidates who received conditional job offers based on unproctored scores. The proctored verification process revealed that less than 1 percent of unproctored scores could not be verified in a proctored environment.

Lessons Learned

The change from static to computer adaptive tests and the subsequent move to an unproctored testing process resulted in several lessons learned. First, the department indicated that working with a reputable vendor with existing CAT capabilities and I/O expertise was a core component to its success. The development of a testing platform, test content, and the data requirements to

establish even one computer adaptive test are daunting indeed, and would have been beyond the budget and capabilities of the department. Having a vendor that provides these services enabled the department to focus on the implementation and use of the tests without having to be concerned with the technology and resource requirements to develop and deliver computer adaptive tests.

Second, the department was somewhat surprised to learn that the use of computer adaptive tests (as opposed to traditional static tests) was not a big issue for applicants or internal stakeholders. Some applicants did question the methodology behind the tests, but were satisfied with high-level explanations that equated the new testing method to that of tests used for education and certification. Similarly, most hiring managers and other internal stakeholders were eager to adopt the new testing technology, as it provided a sense of being on the “leading edge” of assessment practices. As CAT was the key to moving to a successful unproctored testing process and produced the benefits noted above, the shift to CAT was not very difficult at all.

Finally, keeping the explanation of why the department moved to CAT at a high level and focused on the benefits to stakeholders was another key to the program’s success. Attempting to explain the theory and application of item response theory and CAT algorithms would not have gone over well with its internal audience. Knowing what is important to those who will be most affected by change and keeping the discussion on how the change will result in positive outcomes is just as important for implementing computer adaptive testing programs as it is with any other organizational change. The stakeholders were primarily concerned with changing test processes that were not perceived as “broken”. They were also concerned with issues of employee fairness perceptions and face validity. Some of the older tests had been in place for many years and possessed high face validity, while the proposed CAT cognitive ability and personality assessments possessed low face validity. The department’s demonstration of robust validation data and alignment to other widely accepted testing processes reassured stakeholders and gathered the needed support to move to the new CAT system with minimal resistance.

References

- Beaty, J. C., Dawson, C. R., Fallaw, S. S., & Kantrowitz, T. M. (2009). Recovering the scientist-practitioner model: How IOs should respond to UIT. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2, 58–63.
- Fetzer, M. (2009, April). *Validity and utility of computer adaptive testing in personnel selection*. Symposium presented at the Society for Industrial and Organizational Psychology Annual Conference, New Orleans, Louisiana.
- Gatewood, R., & Feild, H. S. (2007). *Human resources selection*. Cincinnati, OH: South-Western College Publishers.
- Stark, S., & Drasgow, F. (1998, April). Application of an IRT ideal point model to computer adaptive assessment of job performance. Paper presented at the Society for Industrial and Organizational Psychology Annual Conference, Dallas, Texas.
- Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika*, 39, 327–350.

Chapter Sixteen

PRACTICE AGENDA

Innovative Uses of Technology-Enhanced Assessment

Michael J. Zickar and Christopher J. Lake

One of the few constants in today's world is rapid change. Although historians may debate whether this period of advanced gadgets and complex technology has brought about the largest amount of change compared to previous times of tumultuous change (for example, the Industrial Revolution) and the introduction of other tools that radically changed the world (for example, development of the printing press, the growth of the railroads), the influences of technology on personnel selection specifically and human resources practices in general are undeniable. Applied psychologists and practitioners, and the organizations for which they work, must not only adapt their knowledge bases and practices to the latest technologies, but they must anticipate future advances and be open to change. The challenge of continuously being innovative is especially difficult because the pace of technological advances far outpaces most of our abilities to understand it. At times, it seems like a full-time job to be able to understand the latest Internet and personal computing advances.

In this chapter, we highlight several aspects that must be considered when adapting to new technology: ethical, scientific, and practical issues. We work through several examples of cutting-edge

topics related to assessment, discussing the various pros and cons for these difficult topics. We discuss and evaluate the practice of employers scouring for personal information on the Internet to help aid selection decisions. In addition, we discuss the potential for brain-scanning and imaging techniques for personnel selection, as well as discuss potential ethical issues involved. Finally, we evaluate the potential for virtual reality assessment to help organizations better equip their employees with the experiences needed to succeed on the job. These examples were chosen because they are areas that have been much speculated about, even though there has been little systematic research conducted by industrial/organizational (I/O) psychologists. Finally, we review general strategies for psychologists and organizations to keep current with technology in ways that are feasible for most of our schedules.

Digging for Digital Dirt: Scouring the Internet for Traces of Personality

People post silly things on the Internet, information that they would never reveal in the context of a job interview or an application blank. Just scouring through friends' status updates on the social network site Facebook, we observed information relating to political views (a friend liked the group *Just Tell Dick Cheney to Shut the Hell Up*), religious views (one friend posted a Bible verse: "he was pierced for our transgressions; he was crushed for our inequities"), sexual orientation (a friend posted pictures from her first anniversary celebration of her same-sex marriage), and work pace (a friend complained on her status update "I'm not a slow grader. I'm glacial"). These bits of information are all things that individuals would most likely not want future employers to know, although in some cases employers may want to know these things. If you do a Google search for the first author of this paper, you will find posts about his personal politics, students comments on ratemyprofessor.com ("Great teacher! Really nice guy, makes class interesting. All of his tests are based on the notes, so don't buy the book") and, with a bit of ingenuity, his number of speeding citations. Again, depending on the type of job for which a person is interviewing, he may or may not want potential employers to see this information.

This scouring for “digital dirt” is quite prevalent among many employers. A 2006 survey of executive recruiters found that 77 percent use search engines to screen for candidates and that 35 percent of these recruiters have eliminated candidates based on the results of these searches (Maclim, 2006). A 2009 survey found that 45 percent of hiring managers search social networking sites such as Facebook, LinkedIn, MySpace, and Twitter to glean information on applicants (Gransz, 2009), with the largest percentages of searches occurring in the information technology sector (63 percent) and the professional businesses services sector (53 percent). The most cited reasons for denying a candidacy based on this information were inappropriate photos, references to illegal drug use, and badmouthing previous employers. Besides using Internet information to exclude applicants, the 2009 survey found that some employers were more likely to hire candidates after a social networking search because of better perceived fit with the company, that the profile supported their qualifications, or that it showed creativity.

Given the growth in Internet intelligence gathering, it is not surprising that an industry has grown up around cleaning up this “dirt.” Firms such as Reputation Defender, Reputation Management Consultants, and International Reputation Management work to remove negative content as well as to promote positive content by manipulating search engine algorithms (Ali, 2008). Although industry has modified its hiring practices to include Internet information and candidates are beginning to adapt to this new digital world, there are many questions that should be answered by I/O psychologists and HR professionals so that such information can be used in a more systematic and valid manner.

Soiling Your Hands with Digital Dirt: Ethical Issues

Before proceeding into issues of validity, it is important to discuss the ethical issues related to the use of digital dirt. As with the use of many technological advances, ethical issues become prominent as technology allows companies to probe into different aspects of their employees’ and candidates’ lives. First off, there is nothing in traditional employment law that precludes companies from

using digital information to influence hiring decisions as long as the use does not result in discrimination based on a protected class (see Palfrey, 2007). As Lenard (2006) notes, discrimination on protected classes may work in subtle ways in this context; employers who use digital dirt may unconsciously discriminate by having different standards for different ethnic or gender categories. He notes that it may be possible for employers to dismiss pictures of white applicants standing in front of a fraternity holding a beer, while eliminating from consideration an African American candidate who is in hip-hop gear at a nightclub holding a beer. Similarly, it is conceivable that someone may discount a Facebook post by a male applicant bragging about his sexual prowess, whereas a female candidate might be eliminated from consideration for a similar type of post. Organizations need to use the same caution that they apply to all other personnel techniques when using information gleaned from the Internet, paying especial attention that standards are applied evenly across all protected groups.

Lenard (2006) mentions that there are other laws, besides the typical employment laws, that may come into play with the use of Internet-based information. For example, the Fair Credit Reporting Act may require companies to disclose that they have sought information on the Internet using a third-party service; this would not prevent them from using the information but merely require them to disclose that such information was used to influence the hiring decision. Employers who “hack” through privacy protection to gather information about a candidate may be subject to invasion of privacy lawsuits. Therefore, companies that create fake online profiles as a means of accessing an unsuspecting applicant’s personal information could be sued under invasion of privacy laws.

The question of whether searching for information about applicants on the Internet is ethical is different from the legal question. Different individuals and companies will have different opinions on the ethics of this practice. Clearly many social network users believe such data-mining techniques are wrong. A quick Facebook investigation finds user-generated groups such as *Employers Using Facebook as a Background Check Is Wrong!* (64 members on 3/11/2010), *Stop Businesses from Facebook Stalking*

You (467 members), *Stop Nosy (sic) Employers from Spying on Your Facebook* (176 members), *Search for 1,000,000 People who Agree that Facebook Shouldn't Effect (sic) Jobs* (343 members), and so forth. Employers who use Internet-based information will likely invoke negative employee and applicant reactions. In fact, an employee or applicant who was annoyed with an organization's use of the Internet to mine data might use these same social network sites to spread negative information about the offending organization and its products. These individuals who joined those Facebook groups believe that their online personas should be separate from their work personas. For example, one individual in the aforementioned groups complained, "So what if we don't like our jobs, the fact that we turn up is good enough, and what's good about getting up at 7 in the morning to go and spend all day at work 5 times a week, of course we will have some shit days. we are only human!"

Employers might argue that this is additional information that may provide insight into future employee behavior that could not be gleaned from other sources. In addition, companies may argue that the distinction between online persona and actual workplace persona may not be as clearly defined as individuals who use Facebook think (as alluded to in the previous quote). For example, the *Harvard Business Review* (June 5, 2007) had a hypothetical case study in which an applicant for a multinational company that was expanding into China had been discovered to have posted anti-Chinese political positions on political blogs. This information may be relevant for how that individual is treated within China. At a minimum, companies that use such information should be up-front about such practices, alerting applicants that Internet searches will be conducted. In addition, companies should respect individual privacy settings and never use deceptive methods of finding out information about applicants.

More importantly from an I/O psychology perspective is the question of whether such information is useful in making predictions about whether applicants will succeed on the job or not. We could only uncover one study that has examined the use of Facebook information in a scientific manner, although there are many studies in progress based on a recent discussion with researchers at the annual Society for Industrial and

Organizational Psychology conference. Clearly, the use of scouring the Internet has advanced far past scientific research. In the following section, we use a series of questions that should be answered before justifying such practices.

Is There Any Validity to Knowing That He Likes Brittany Spears? Scientific Issues

The question that surrounds the use of digital dirt in informing personnel decisions is whether digital dirt can produce any job-relevant information. Perhaps there is a correlation between the propensity to post personal information online and other personality and character traits that may be of relevance to organizations. One could argue that this propensity could be correlated with traits often deemed negatively by organizations such as impulsivity. In addition, there may be other traits that could be correlated with posting information on social networks that would be deemed positively by organizations (at least for some jobs), such as positive energy, need for privacy, openness to experience, and spontaneity.

Future research should investigate whether important information can be gleaned from postings on blogs, social media sites, and general information that would show up in Google search results. Studies could be conducted that compare coder ratings of personality traits based on Internet searches to actual personality trait scores. It is our guess that the correlation between observer ratings and personality test scores would be low, and that by itself would suggest that such data should not be used in making hiring decisions. The one study that we found that examined Facebook information, however, suggests that there may be some utility in assessing personality via Facebook.

Back, Stopfer, Vazire, Gaddis, Schmukle, Egloff, and Gosling (2010) found modest correlations (ranging from .13 to .41) between observers' assessments of personality based on Facebook profiles with self-ratings of actual personality. Correlations were smaller between observers' assessments and self-ratings of ideal personality, suggesting that Facebook information is more linked to actual personality than some kind of idealized view of the self. The main finding, however, is that there was some modest

correspondence between assessment of personality based on Facebook profiles and self-reports (although it should be noted that the relationship between observer assessments of neuroticism and self-ratings of neuroticism was not significant). The magnitudes of these correlations, however, were small enough to suggest that Facebook assessments cannot be used as surrogates for self-ratings of personality assessment. In addition, the validity of these observer assessments to predict external criteria such as job performance needs to be examined. Finally, it should be noted that the magnitude of correlations was higher when the average of multiple observers was used (nine to ten observers were used for each profile in their study).

In addition to using an observational method similar to Back, Stopfer, Vazire, Gaddis, Schmukle, Egloff, and Gosling (2010), researchers could take an empirical validation approach, similar to that used by many biographical data researchers, to develop empirical keys that could be used to predict job performance. For example, researchers could correlate the presence or absence of certain types of cues (for example, mention of athletics versus mention of literature) with various performance outcomes. The danger of this approach, just like empirical keying of any type of measure, is that of capitalization on chance. Given empirical keying's success with biographical data, however, this approach might have some warrant (see Mitchell & Klimoski, 1982).

Research suggests a lot of harm could come from employers' use of Internet-related information in an unsystematic manner. Research on job interviews shows that interviewers often focus on job-irrelevant information, which can distract interviewers from important information or bias interviewers in positive or negative directions (Arvey & Campion, 1982). This seems possible with digging for Internet dirt, where people may focus on information irrelevant to job performance such as whether a candidate supported McCain or Obama in the 2008 election or whether he or she likes rap music or is an agnostic. The amount of irrelevant information is likely to be much higher in the Internet context, compared to other settings that have stronger norms of people guarding personal information. In addition, research in judgment and decision making suggests that irrelevant information

may dilute the effect of relevant information. Nisbett, Zukier, and Lemley (1981) found that when individuals are presented with diagnostic (information found to be valid) information along with non-diagnostic information, individuals make less accurate predictions than when presented with diagnostic information alone.

Research studies could be done to examine the effects of Internet information supplementing existing information that has already been found to be valid and relevant. It is our suspicion that employers will search the Internet for pictures of applicants and that these pictures might bias decision-makers. Photographs might especially bias applicants who are overweight (Roehling, 1999), who fail to fit sex-role stereotypes, or who are physically unattractive (Cash, Gillen, & Burns, 1977). For example, research has shown that people will judge applicants who are more physically attractive as more qualified for the job compared to other applicants even though the resumes are indistinguishable on other characteristics (Cash, Gillen, & Burns, 1977). It is possible that employers could see pictures of tattoos and piercings that applicants might be able to hide in interview settings and would probably lead to unfavorable perceptions (see Dale, Bevell, Roerch, Glasgow, & Bracy, 2009).

Conclusions About Digital Dirt

Employers are digging through information on the Internet at high levels. The question is whether the information is being used in appropriate ways. The ethics and legality of such practices need to be considered but the area where I/O psychologists can really help is the scientific investigation of the utility of considering such information. I/O psychologists should provide answers about what type of digital information is valid in helping make personnel decisions and which information is just noise that will distract decision-makers from making unbiased decisions. It is our suspicion that most of the information currently being considered by those who troll the Internet for information on their applicants is worthless at best and can potentially lead employers to make worse decisions than without it, but the research needs to be conducted.

Scanning the Brain for Intelligence and Personality Characteristics: Modern-Day Phrenology or Modern-Day Assessment

In futuristic movies and science fiction novels, the idea that machines could read the brains of individuals and determine their future seemed either like Orwellian nightmares in which organizations are allowed to peer into our most sacred spaces. Or, perhaps to some, these futuristic dreams seemed like efficient ways of making better selection decisions. For example, the 1981 movie *Scanners* portrayed characters who had telepathic abilities that allowed them to learn all kinds of sordid thoughts about people. Although these fictions might have seemed far-fetched at the time, there has been a fair amount of research aimed at understanding the potential for humans based on images in the brain. Many technologically advanced techniques now exist for scanning the human brain; these techniques are collectively known as neuroimaging. Some of the more recognizable forms of neuroimaging are magnetic resonance imaging (MRI) and positron emission tomography (PET) scanning. There are two general classes of neuroimaging relevant to this conversation: structural and functional. Whereas structural imaging methods allow for measurement of static brain features (for example, quantity of grey matter; brain size), functional imaging allows for measurement of dynamic brain processes (for example, blood flow; glucose metabolism). The question is whether information gathered from these assessments can provide reliable, valid, and practical assessments related to work performance. Can data from brain images be used to predict performance or are they just modern-day phrenology?

Research on Brain Imaging

Although brain scans have been used to aid medical and psychiatric diagnoses for some time, scanning techniques have more recently been used in the study of intelligence, personality characteristics, and emotional regulation. Joseph Matarazzo, former American Psychological Association president, notably stated that physiological measures obtained from neuroimaging could one

day supplement or even replace traditional psychometric measures of intelligence (Matarazzo, 1992). More recently, professor and researcher Willem Verbeke has stated that neuroimaging techniques will soon become part of pre-employment testing like an interview (Oosting, 2009; Over 5 jaar scannen, 2009). In one of the first studies of its kind, Verbeke and colleagues (Dietvorst, Verbeke, Bagozzi, Yoon, Smits, & Van der Lugt, 2009) used neuroimaging measures to validate a psychometric scale measuring salespeople's ability to empathize with a client's thought process and dynamically adapt to changing sales situations. Participants whose brains showed specific activation patterns received higher scores on the psychometric measure, indicating greater sales ability than those without this activation pattern.

Are brain scans an alternative method of assessing intelligence and personality? As compared to the other methods of assessment reviewed in this chapter, brain scanning has a relatively large body of peer-reviewed scientific evidence from which to draw. Across multiple studies, there does appear to be a relatively consistent ability to find brain-measure correlates of intelligence and personality. First, consider how neuroimaging-derived measures relate to psychometric measures of intelligence. Anderson (2003) reviewed fourteen structural imaging studies of brain size (volume) and found that all but one reported a significant correlation between brain size and psychometrically-measured (for example, WAIS, Vandenberg Mental Rotation Test, Raven's Progressive Matrices) intelligence. Anderson estimated the correlation between brain size and intelligence to be $r = .35$. A meta-analytically derived average correlation $r = .33$ was later calculated by McDaniel (2005). More recently, seven studies reported in Jung and Haier (2007) report statistically significant correlations between intelligence and measures of grey matter in certain brain regions. In addition to the structural measures just described, functional imaging measures have also been shown to correlate with psychometrically measured intelligence. Numerous studies by Haier (for example, Haier, Cheuh, Touchette, Lott, Buchsbaum, Macmillan, Sandman, Lacasse, & Sosa, 1995; Haier, Siegel, Neuchterlein, Hazlett, Wu, Peak, Broning, & Buchsbaum, 1988) have shown a significant inverse correlation between localized brain glucose metabolic rate and measures of intelligence.

That is, high IQ seems related to the presence of highly-efficient neurons in the brain (Jung & Haier, 2007). Jung and Haier (2007) reviewed thirty-seven structural and functional neuroimaging studies and found converging evidence across studies to suggest that specific brain lobe regions are especially important to intelligence and reasoning.

Now consider the ability of neuroimaging to detect personality characteristics. Although Haier (2004) reports that the use of neuroimaging measures in personality studies is still in its infancy relative to the study of intelligence, several functional imaging studies show a significant correlation between localized brain activation and psychometric personality measures (for example, personality dimensions scores from the NEO-PI-R and other personality inventories). Canli (2004) reports correlations of .79 and .71 between localized brain activation and extraversion in two studies. Haas, Omura, Constable, and Canli (2007) report significant correlations of .47 and .55 between psychometric neuroticism scores and localized brain functioning. Although not studied as thoroughly, at least one study (Haas, Omura, Constable, & Canli, 2007) reports a significant ($r = .42$) relationship between localized brain functioning and agreeableness.

In the quest for reliable and objective assessment, the prospect of using a physiological brain scan as a means of employee assessment may sound appealing at some level. Farah, for example (2002, p. 1127), notes that “measures of brain function are one causal step closer to [underlying] traits and states than . . . more familiar measures [such as] responses on personality questionnaires.” Whereas psychometric measurement scales may be biased by faking or contextual factors, neuroimaging may provide a relatively uncontaminated measure.

Practical and Ethical Problems with Brain Imaging

Using neuroimaging technology in practice would prove very difficult for several reasons. First, the cost of imaging equipment can be quite prohibitive. We found that imaging machines can be purchased used for as low as about \$500,000USD and new machines may require a multi-million-dollar investment. Without having to purchase a machine, Haier (2003; 2004) reports that

MRI brain scans can be obtained for about \$400USD per person and PET scans can be obtained for about \$1,200USD per person.

Cost aside, an additional problem arises due to the inherently medical nature of neuroimaging. Brain scans are capable of detecting brain abnormalities (for example, lesions) and indicators of psychiatric abnormality (for example, depression or anxiety disorders). The localized brain areas that Dietvorst, Verbeke, Bagozzi, Yoon, Smits, and Van der Lugt (2009) used to validate their psychometric sales-ability scale, for example, are also used by clinicians to assess the presence of autism. Furthermore, Haas and Canli (2008) note that indicators of personality characteristics and psychiatric disorders are quite often found in the same areas of the brain. Simply knowing that a job applicant has low activation in these brain regions means having knowledge about what could potentially be a medical diagnosis.

From a legal perspective, using neuroimaging in employment practice may invite lawsuits by those claiming violation of the Americans with Disabilities Act of 1990. This act prohibits employers from asking about the existence of any mental or physical disability or using such information to discriminate against an applicant in the hiring process, unless it is specifically related to the job requirements. Whereas intelligence or personality characteristics are often measured with psychometric methods, using brain scans to accomplish the same task may appear to judges and juries to be more of a medical than cognitive procedure. In addition, candidates may resent the intrusion and form negative opinions of the process.

There also seems to be a fear that neuroimaging services will be offered before the state of the science has concluded what, exactly, neuroimaging tells us about a person's future behaviors. Parallel to the issue discussed here is the use of fMRIs as lie detectors. At least two companies now offer lie-detection services via fMRI even though there is still much scientific debate as to the utility of such measures (Farah, 2009). In spite of people's fears of pre-employment brain scanning, Willem Verbeke believes that neurological imaging promotes the societal good by allowing employers to screen out psychopaths (Over 5 jaar scannen, 2009).

Conclusions About Brain Imaging

We conclude the neuroimaging section with a word of caution about the alleged objectivity of brain-level measures. It is important to keep in mind that brain measures rely on statistical inferences just as any psychometric measure would. A study by Weisberg, Keil, Goodstein, Rawson, and Gray (2008) demonstrates the “seductive allure” of neuroscience-based explanations of psychological phenomena. These researchers found that inserting entirely irrelevant brain scan–derived information into an explanation of phenomena markedly increased participants’ satisfaction with the explanation.

Virtual Reality Assessment

Another cutting-edge use of technology is the use of virtual reality to improve the training and assessment of organizational members, whether new employees or long-term ones. *Virtual reality* is a phrase that has many different connotations, though common elements from many different definitions include heightened sensory information (typically a three-dimensional environment) along with the mechanisms to track input from users. As Fox, Arena, and Bailenson (2009) state, the “goal of a virtual environment is to replace the cues of the real world environment with digital ones” (p. 95). This can be accomplished by providing increased sensory information in a variety of ways. In one combat game, virtual combatants wear a vest that allows them to feel some of the effects (albeit in less serious amounts of pain!) of being hit by bullets or being close to grenades as they explode (Cormack, 2008).

Virtual reality caught on initially with video gamers. Games such as *Dactyl Nightmare* and *Legend Quest* helped gamers feel as though they were closer to the action and that the worlds that they were exploring were more realistic. It took many years, however, until the promise of virtual reality gaming provided an experience that truly flooded the senses and took advantage of virtual reality technology (Cormack, 2008). Recent technological advances such as the combat vest and the Emotiv EPOC have advanced the gaming industry in significant directions. The

Emotiv EPOC system is a headset that uses electroencephalography (EEG) technology that claims to be able to respond to users' thoughts and emotions as well as head movement to provide a more realistic gaming environment. The current version is available for \$299USD and hooks up to traditional PC systems (www.emotiv.com, 2010).

Research on Virtual Reality Assessment

Virtual reality technology has been used for training and assessment purposes in fields in which the skills being trained or evaluated are quite technical or the cost of errors is quite high. One example is that of surgical training and assessment simulations. Seymour, Gallagher, Roman, O'Brien, Bansal, Andersen, and Satava (2002) studied whether virtual reality training resulted in better operating room performance in a gall bladder surgery compared to randomly assigned residents who had received the standard training. The virtual reality training involved a 3D box that represented an accurately scaled operating room. The group that had virtual reality training in addition to the standard training had significantly fewer errors in gall bladder surgery. Seymour and colleagues conclude that virtual reality training helped because it allowed as much training as needed to master the task. Virtual training and assessment often allow a trainee to monitor his or her own performance and observe performance improvements over time. Performance data from such training simulations could also be relayed to a supervisor for review. Gallagher and Satava (2002) found that a virtual reality laparoscopic surgery training device was valid in that it was able to discriminate the performance of experienced surgeons from the performance of inexperienced surgeons. The researchers also report high test-retest and internal consistency reliabilities, indicating that such a device may be quite useful in assessing the psychomotor skills required to safely perform specific surgical operations. Although these examples were conducted in the context of training, they could be used to develop work sample tests, which have been shown to possess a high amount of validity across a variety of jobs (Roth, Bobko, & McFarland, 2005).

Military operations is another field in which technical skills are required and the cost of errors is high. Gately, Watts, Jaxtheimer, and Pleban (2005) describe a virtual reality assessment of decision making used at a military training facility. Teams of soldiers use the VR platform to carry out missions in war-like environments. As the scenario plays out, key statistics such as deviation from route, number of shots fired, and fratricides (friendly fire incidents) are logged by the system, along with an audio recording of all interpersonal communications. Trainers then analyze the results and provide feedback to the soldiers, discussing key decision-making points in the mission and the soldiers' performance.

Virtual reality may also be used in the training and assessment of interpersonal skills. Heiphetz and Woodill (2010) report that the Canadian Border Services Agency provides border agents the opportunity to practice interpersonal communication in a virtual environment. Agents interact with a virtual reality program using avatars (characters that the participants choose to represent themselves in the program). Agents then interact with virtual people who are attempting to cross the border. Heiphetz and Woodill also report on the use of sales training via virtual reality software. Again, avatars are used to represent the employee and other people (potential clients, in this case). The salespeople interact with the potential clients to learn skills such as getting past a gatekeeper (receptionist) and sales pitch openings. These and other interpersonal communication VR programs can serve as both training tools and assessment tools. When used as an assessment tool, a minimum level of competency may be required before an employee is deemed ready to work in the field. The sales program, for instance, provides a score that is purportedly indicative of performance in the scenario. It also provides feedback about strengths and weaknesses of the conversation that took place between the employee and potential client.

Although the applications to organizational practice have been minimal, virtual reality could have implications for improving the efficiency and effectiveness of assessment, as demonstrated by some of the aforementioned examples. As noted in Cosman, Cregan, Martin, and Cartmill's (2002) review of surgical

virtual reality simulations, the benefits of VR implementation are two-fold: skill acquisition and assessment. As it pertains to use in organizations, VR is often discussed in terms of training benefits; meanwhile, the assessment benefits are seemingly overlooked. In our opinion, there are many situations in which it might make sense to use virtual reality to assess someone's level of competency. As mentioned previously, although most of the research has been conducted in the context of training, virtual reality technology could be used to create work sample tests that have high amounts of validity.

Conclusions About Virtual Reality

An HR or I/O practitioner interested in using virtual reality for assessment should consider that goal when the virtual reality protocol is being designed. Whereas a VR program used solely for training may not require that the user receive a specific score or feedback, these elements should probably be included in any VR assessment. Cannon-Bowers and Bowers (2010) propose that a method of tracking dynamic performance throughout the course of the assessment, and providing this information in real time to the employee, may be helpful.

Additionally, we feel that fidelity is a real concern with the use of virtual reality assessment. There is clearly a continuum of realism that can be used for assessment purposes. 2D computer simulations, such as the initial flight simulators, might be considered relatively low-level simulations of realism, whereas full 3D simulators, as mentioned previously, might be considered full virtual reality experiences. Low fidelity simulations might not adequately capture task performance and may not appear credible to a person being assessed. However, the incremental value of making a low fidelity experience into a full-blown virtual reality experience is a prime question that experts need to consider. In cases in which the behavior required is complex and the cost of errors is prohibitive, it might make sense to expend large amounts of money to make sure that assessment is as realistic and effective as possible. The areas in which advances in virtual reality training and assessment have been most extensive have been in the military and medical domains,

both areas in which the cost of errors is measured in human lives in addition to money. It should be kept in mind, however, that the cost of virtual reality technology will be decreasing significantly and so training tools that might have seemed out of the reach of most personnel budgets might be feasible in the near future.

Significant research questions remain about the use of virtual reality assessment. As alluded to above, the incremental value of virtual reality assessment over other forms of assessment should be one of the first questions to be answered. For many purposes, the additional utility of the virtual reality assessment will prove to be minimal and not worth the cost to develop.

Overall Conclusions

Increases in technology will happen at a pace much faster than those of us who focus on personnel psychology and human resource management can handle. It is unreasonable to expect that we can maintain content expertise as well as keep track of the latest technological advances. Given the assumption that most of us cannot devote a significant portion of our time to keeping up with the latest technology, we provide some suggestions on how to keep up with the latest technology without making it a full-time profession.

Go Social!

Social networking sites are not just for teenagers anymore! The fastest growing population among Facebook users was the age group from thirty-five to fifty-four, which grew by 276.4 percent over a six-month period ending at the beginning of 2009; the second fastest growing group was fifty-five and over, which grew at a 194.3 percent rate over the same period (Corbett, 2009).

We recommend that you log on to a few social networking websites and explore their features. Do not be afraid to ask others for advice. Interact with others and learn the technology simply through hands-on experience. And learn to appreciate the mentality that others use the website.

Read General Technology Websites and Magazines

Magazines such as *PC Magazine* and *Wired* provide easy-to-understand articles about the latest technological advances. The price of an annual subscription is minimal, and perusing articles on a regular basis will keep you somewhat informed on the latest developments. One website to peruse is www.cnet.com, which provides reviews on latest technological gadgets and software. The website www.lifehacker.com provides tips on how to integrate technology into your life, making tasks easier and more efficient. In addition to technology websites, there are a good number of human resource websites that would provide insight into technology and HR practices. The blog (a personal website written by one or more individuals with individual perspectives) www.sixdegreesfromdave.com provides insight into HR practices with a focus on social networking sites. The website www.hrvendornews.com focuses on HR vendor news updates, providing insight into new HR products, especially those with a technological bent. Just as with the magazines, following one or two of these blogs requires little investment and might result in some good ideas.

Go to a Focused Conference

Many conferences have broad themes that cover a wide variety of topics. Large annual conferences of groups such as the Academy of Management, Society for Industrial and Organizational Psychology, and Society for Human Resource Management will all have sessions that are related at least in some ways to the use of technology in HR and I/O psychology practice. At times, however, the discussion in the breakfast rooms and bars at these conferences becomes too diluted by the myriad of topics covered in the conference. There are conferences focused exclusively on HR practices and technology, and attending one of these conferences occasionally may provide you with additional contacts and ideas. For example, there is an annual HR Technology Conference sponsored by *Human Resource Executive* magazine. The 2010 conference has sessions such as Great New Technologies Just for You—including Twitter!, Cool New Technologies for HR, and MetLife Tackles Workforce Analytics—Twice (www.hrtechconference.com).

Find a Technology Confidant

It is embarrassing to appear ignorant in front of others. This can happen frequently with technology. You should consider finding a person with whom you feel completely comfortable asking stupid questions. Take that person to lunch on a regular basis. Discuss your latest HR problems and needs and see whether he or she has insights into your problems. Your confidant can be an employee within the organization where you work or an acquaintance with no connections to your business life. Either way, be nice to him or her and pick his or her brain on a regular basis.

Do Not Forget Your Core Training

The core principles used to evaluate personnel practices, whether training programs, hiring procedures, or assessment devices, apply regardless of whether one is evaluating a paper-and-pencil test or a test that is based on some sophisticated piece of technology that purports to determine someone's personality from brain scans. The basic concepts of validity, reliability, and adverse impact are still relevant, regardless of how fancy the technology. Generally, the individuals who are working on the technology side of an assessment system have little insight into these issues, so it is especially important that testing professionals be involved throughout the lifespan of the project.

Often, the issues involved in assessing reliability, validity, and adverse impact present the same challenges as other forms of assessment. For example, the questions of what point a test-taker becomes an applicant is a central concern for all Internet testing because of the effect the answer has on the level of adverse impact that may be found. Do you need to count everyone who stumbles upon your testing page or should you count only people who complete all portions of your assessment? For digital dirt and other forms of assessments that require rating of a work product, several forms of reliability may be important. First, interrater reliability would be important to establish. Different raters who observe the same phenomena might draw vastly different conclusions. Second, temporal reliability may be especially important when information on the Internet changes on a regular basis.

Given that many users of social network sites change their information on a regular basis, the temporal stability of ratings based on this information should change. To properly assess reliability, validity, and adverse impact, knowledge of the particular technological information is important and the evaluations should be tailored to such knowledge.

Remember Legal and Ethical Issues

All of the techniques that we consider have advantages in that they allow employers to assess aspects of candidates' capabilities that might not be able to be assessed via other means. Although research may ultimately show that these techniques provide incremental validity over traditional methods, these techniques are likely to cause resentment among some applicants who view these assessments as unfair intrusions into their personal lives. The question of whether the additional validity justifies the resentment among some applicants is a debate that companies will need to consider. In addition to ethical reasons, companies will need to determine how various practices relate to legislation. Although most of us are familiar with the Civil Rights Act, other laws may become relevant with these new technologies. As assessments forge into new areas, we will need to adapt these practices to new laws. In addition, if these assessment techniques are used improperly and irresponsibly, it is likely that additional legislation would be created to address these misuses.

Final Conclusions

Modern technology provides many opportunities for advancing personnel practices, especially training and assessment. Unfortunately, to fully integrate such technologies requires knowledge that most personnel experts do not have. In this article, we reviewed three technologies that have the potential, if done right, to advance the practice of what we do. As seen throughout these examples, however, technology often advances faster than the understanding of its ramifications and utility. Companies that use cutting-edge technology for training and assessment need to worry, not only about validity, reliability, and

adverse impact, but also about reactions that such techniques might cause among employees and applicants. Potential negative applicant reactions might be much more severe with these technologies than with traditional assessments, given that some of these technologies probe into aspects of lives previously left alone. We hope that we have provided some ideas and examples that will stimulate your thinking!

References

- Ali, L. (2008, February 18). Google yourself—and enjoy it. *Newsweek*.
- Anderson, B. (2003). Brain imaging and g. In H. Nyborg (Ed.), *The scientific study of general intelligence* (pp. 41–51). Boston, MA: Pergamon.
- Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, *35*, 281–322.
- Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, *20*, 1–3.
- Bevill, S., Roerch, T., Glasgow, S., & Bracy, C. (2009). Body adornment: A comparison of the attitudes of businesspeople and students in three states. *Academy of Educational Leadership Journal*, *13*, 69–78.
- Canli, T. (2004). Functional brain mapping of extraversion and neuroticism: Learning from individual differences in emotional processing. *Journal of Personality*, *72*, 1105–1132.
- Cannon-Bowers, J., & Bowers, C. (2010). Synthetic learning environments: On developing a science of simulation, games, and virtual world for training. In S.W.J. Kozlowski & E. Salas (Eds.), *Learning, training, and development in organizations* (pp. 229–261). New York: Routledge.
- Cash, T. F., Gillen, B., & Burns, D. S. (1977). Sexism and beautyism in personnel consultant decision making. *Journal of Applied Psychology*, *62*, 701–710.
- Corbett, J. (2009). 2009 Facebook demographics and statistics report: 276 percent growth in 35–54 year old users. www.istrategylabs.com/2009/01/2009-facebook-demographics-and-statistics-report-276-growth-in-35-54-year-old-users/. Accessed March 30, 2010.
- Cormack, B. (2008). Whatever happened to virtual reality gaming? <http://news.gotgame.com/whatever-happened-to-virtual-reality-gaming/19106/>. Accessed March 30, 2010.

- Cosman, P. H., Cregan, P. C., Martin, C. J., & Cartmill, J. A. (2002). Virtual reality simulators: Current status in acquisition and assessment of surgical skills. *ANZ Journal of Surgery*, *72*, 30–34.
- Dale, L. R., Bevill, S., Roerch, T., Glasgow, S., & Bracy, C. (2009). Body adornment: A comparison of the attitudes of businesspeople and students in three states. *Academy of Educational Leadership Journal*, *13*, 69–77.
- Dietvorst, R. C., Verbeke, W.J.M.I., Bagozzi, R. P., Yoon, C., Smits, M., & Van der Lugt, A. (2009). A sales-force-specific theory-of-mind: Tests of its validity by classic methods and functional magnetic resonance imaging. *Journal of Marketing Research*, *46*, 653–668.
- Farah, M. J. (2002). Emerging ethical issues in neuroscience. *Nature Neuroscience*, *5*, 1123–1129.
- Farah, M. J. (2009). Neuroethics. In A. L. Caplan, A. Fiester, & V. Ravitsky (Eds.), *The Penn Center guide to neuroethics* (pp. 71–83). New York: Springer.
- Fox, J., Arena, D., & Bailenson, J. N. (2009). Virtual reality: A survival guide for the social scientist. *Journal of Media Psychology*, *21*, 95–113.
- Gallagher, A. G., & Satava, R. M. (2002). Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. *Surgical Endoscopy*, *16*, 1746–1752.
- Gately, M. T. Watts, S. M., Jaxtheimer, J. W., & Pleban, R. J. (2005). *Dismounted infantry decision skills assessment in the virtual training environment* (Technical Report No. 1155). Arlington, VA: United States Army Research Institute for the Behavioral and Social Sciences.
- Gransz, J. (2009). Forty-five percent of employers use social networking sites to research job candidates. www.careerbuilder.com. Accessed March 11, 2010.
- Haas, B. W., & Canli, T. (2008). Emotional memory function, personality structure and psychopathology: A neural system approach to the identification of vulnerability markers. *Brain Research Reviews*, *58*, 71–84.
- Haas, B. W., Omura, K., Constable, R. T., & Canli, T. (2007). Is automatic emotion regulation associated with agreeableness? *Psychological Science*, *18*, 130–132.
- Haier, R. J. (2003). Positron emission tomography studies in intelligence: From psychometrics to neurobiology. In H. Nyborg (Ed.), *The scientific study of general intelligence* (pp. 41–51). Boston: Pergamon.

- Haier, R. J. (2004). Brain imaging studies in personality: The slow revolution. In R. M. Stelmack (Ed.), *The psychobiology of personality* (pp. 329–340). Boston, MA: Elsevier.
- Haier, R. J., Cheuh, D., Touchette, P., Lott, I., Buchsbaum, M. S., Macmillan, D., Sandman, C., Lacasse, L., & Sosa, E. (1995). Brain size and cerebral glucose metabolic rate in nonspecific mental retardation and Down syndrome. *Intelligence*, *20*, 191–210.
- Haier, R. J., Siegel, B. V., Nuechterlein, K. H., Hazlett, E., Wu, J. C., Paek, J., Browning, H. L., & Buchsbaum, M. S. (1988). Cortical glucose metabolic rate correlates of abstract reasoning and attention studied with positron emission tomography. *Intelligence*, *12*, 199–217.
- Heiphetz, A. & Woodill, G. (2010). *Training and collaboration with virtual worlds*. New York: McGraw-Hill.
- Jung, R. E., & Haier, R. J. (2007). The parieto-frontal integration theory (P-FIT) of intelligence: converging neuroimaging evidence. *Behavioral and Brain Sciences*, *30*, 135–187.
- Lenard, G. (2006). Employers using Facebook for background checking: Is it legal? www.employmentblawg.com. Accessed March 11, 2010.
- Maclim, T. (2006, June 12). Growing number of job searches disrupted by digital dirt. Execunet.com, www.execunet.com/m_releases_content.cfm?id=3349. Accessed March 11, 2010.
- Matarazzo, J. D. (1992). Psychological testing and assessment in the 21st century. *American Psychologist*, *47*, 1007–1018.
- McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence*, *33*, 337–346.
- Mitchell, T. W., & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology*, *67*, 411–418.
- Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *13*, 248–277.
- Oosting, A. (2009, February 24). Hersenscan solicitant is straks vanzelfsprekend [Brain soon to be obvious candidate]. *Algemeed Dagblad*. Over 5 jaar scannen we het brein van elke solicitant [In five years, we will scan the brain of each candidate]. (2009, February 19). *Vacature*.
- Palfrey, J. G., Jr. (2007, June 5). Should Fred hire Mimi despite her online history? *Harvard Business Review*.
- Roehling, M. V. (1999). Weight-based discrimination in employment: Psychological and legal aspects. *Personnel Psychology*, *52*, 969–1016.

- Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*, 1009–1037.
- Schultheis, M. T., Rebimbas, J., Mourant, R., & Millis, S. R. (2007). Examining the usability of a virtual reality driving simulator. *Assistive Technology, 19*, 1–8.
- Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual reality training improves operating room performance. *Annals of Surgery, 236*, 458–464.
- Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*, 470–477.
- www.emotiv.com (2010). EPOC headset. www.emotiv.com/apps/epoc/299/. Accessed March 30, 2010.

Chapter Seventeen

CONCLUDING COMMENTS: OPEN QUESTIONS

Seymour Adler

The chapters in this book present the state-of-the-art of technology-mediated assessment as practiced in leading-edge organizations early in the second decade of this millennium.

As several of our authors note, we are just at the beginning of the journey in using technology to enhance assessment. Our next steps will be determined both by technological developments—largely outside our expertise and control as assessment professionals—and by the progress we make in answering some key open questions about technology-mediated assessment. Those questions can, I believe, be addressed through systematic theory development and theory application, programmatic research, and through informed, open discussion of appropriate practice standards and guidelines.

I see these questions falling into four, increasingly broad, categories: the assessment, the candidate, the organization, and society. My intent is to touch on some of the open questions that relate to each category.

The Assessment

Nuances

Our authors have collectively raised a great number of questions about technology-mediated assessments. These include some fundamental questions such as:

- Are scores on a computer-administered test equivalent to scores on the same test administered in paper and pencil?
- How can you establish test equivalency across languages and cultures?
- What is the impact of proctored versus unproctored test administration on applicant performance?
- What norms should be used to interpret test scores when scores are compared across test versions that differ in language or differ in the items administered?

These and similar basic assessment concerns, to be sure, were not created by technology-mediated testing. In fact, these questions could be and in many cases were raised about the different conditions under which paper-and-pencil versions are administered. However, the rapidly expanding use of technology in assessment has made answering these questions all the more critical to progress.

As we move forward, the way these fundamental questions are posed and answered needs to become far more subtle. The simple comparison of technology- versus paper-and-pencil-administration now seems almost quaint. More nuanced comparisons need to be made between administration across different technologies, say laptops/desktops versus smart phones; between visual versus auditory or even tactile modalities; or across devices that differ in multimedia richness.

Similarly, given the complexities of global testing programs (Ryan & Tippins, 2009), it is naïve to think that the question of test equivalence across languages and cultures can be addressed in *a* study. Like the question of test validity before it (Messick, 1995), the case for equivalence will be made by building an evidentiary base for an assessment tool across a *series* of studies

varying in research design and varying in the forms, formats, languages, and cultures compared, rather than by any one once-and-done study (see Bartram, Chapter 7 in this book, for an example). In addition, this research will need to examine interactions between culture and the medium of assessment. It may well be that *different* media are needed to validly assess the same construct in different cultures, much as we already recognize that items may have to be substantially rephrased and not just literally translated when transported across cultures. For example, written scenario descriptions may produce valid scores on leadership potential in one country, while those scenarios may have to be presented in graphic, cartoon, or video formats to yield similarly valid scores in another country.

Administration

There has been much recent attention paid to issues around unproctored Internet testing. But how much do we really know about technology-mediated assessment in *proctored* settings? To take one proctored setting that has not been studied, today millions of applicants annually take computer-administered high-stakes tests in third-party testing centers. Most of the large third-party testing center networks in the U.S. are actually a combination of company owned and operated sites and small, local “mom-and-pop” sites. These centers vary in the degree of privacy, of quiet, of proctor presence, and of visual stimulation experienced by applicants at their workstations. At any given time, an applicant may be taking a test in the presence of fellow applicants of the same or different race, ethnicity, or gender. What effect do these variations have on social facilitation, identity salience, competitiveness, anxiety, self-handicapping, and other states affecting test performance? More generally, how do these internal states mediate the linkages between test administration conditions and test performance?

Today, when virtually all new devices on which tests can be administered have webcams with increasingly high resolution *facing the applicant* (cameras for communication and not just picture-taking), some of the risks associated with applicant identity in unproctored environments are mitigated. In addition,

these devices capture vocal input that can be used for voice printing to further monitor the identity of the respondent. Keystroke analytics (see Arthur and Glaze, Chapter 4 in this book) add to the ability to monitor applicant identity. The key question then is no longer the simple proctored-versus-unproctored comparison. The questions that we need answers for now relate to the impact on test performance of the *mode* of proctoring (visual, auditory), the *immediacy* (including but not limited to temporal immediacy) of proctoring, the perceived efficacy of monitoring, and the perceived *consequences* of cheating.

There is some research, reviewed by Scott and Mead in Chapter 2 of this book, indicating that retesting generally has a positive effect on assessment scores, with the effect stronger the more similar the forms and the shorter the retest interval. The accessibility of technology-enhanced assessment is likely to increase the frequency of retesting. Applicants today can build in reminders that automatically pop up on their electronic calendars, complete with a link to the testing site, and take the same online test every six or twelve months until they pass. Or, at the same point in time, they may take presumably different tests that nonetheless share substantial content when applying for multiple jobs within an organization or across organizations that use the same vendor's content. The broad question of whether there is an effect of retaking the same test after six months needs to be replaced by more nuanced retesting questions about "item" type (an "item" might be a role-play scenario, not just a traditional multiple-choice item), degree of overlap between test versions, item order effects, modality effects (visual versus auditory), testing circumstances (how more or less desperate the applicant is for a job when taking the test for the second time), test-smartness even if a completely different but parallel item set is administered, and many others.

Design

Looking at the chapters in this book, I am impressed by how creative we have become in leveraging technology to deliver assessment content, especially for high-fidelity simulations (see, for example, Chapter 8 by McNelly, Ruggeberg, and Hall, Chapter 11

by Hartog, and Chapter 10 by Grubb). I am less impressed by our creativity in designing versions of more traditional assessments of ability, personality, or knowledge, which fully leverage the unique added value of technology to measure constructs in more engaging and job-related ways—or to measure different constructs.

For example, ability measures can be designed to administer reasoning skill items or customer service situational judgment items under quiet, baseline conditions and then to present visual or auditory distractors or introduce time or other pressures (for example, the virtual presence of others), to assess test performance *decrement* under circumstances that might reflect real-world situations. I have seen some experimental measures that do this, but rarely in large-scale operational testing programs in commercial settings. There is a huge literature on Implicit Association Test measures of attitudes (Wittenbrink & Schwarz, 2007), with millions of web administrations. The test measures how strongly people unconsciously associate different concepts by assessing how quickly they can categorize objects related to the concept. In attitude measurement, for instance, a racially prejudiced white person might take a few milliseconds longer to associate a black person's picture on the screen with positive adjectives than a white person's picture. The same technique could be applied to evaluate person-organizational fit or leadership preferences or to measure personality in the employment context.

Recently, I saw an interesting approach to measuring job fit. The applicant is presented with a series of very brief (five-to-ten-second) video clips from a Realistic Job Preview. Each clip illustrates how a key task is performed in the target job. Then the applicant is asked to rate the degree to which he or she enjoys or anticipates enjoying actually performing that task on a daily basis. The assessment presents some ten to fifteen task-based realistic video clips and generates a total job-fit score across these key tasks. Prinsloo and her colleagues in South Africa have developed a novel approach to measuring a wide range of facets of cognitive ability with minimal adverse impact by creating a self-paced computer-delivered system that teaches the applicant a novel symbolic language and measures the speed of learning, the types of errors made during the learning process, and the

creativity and fluency in the candidate's use of the symbolic language once mastered, among other facets (www.cognadev.com). Technology can capture keystroke analytics, which, while used today to identify test-taker identity (Arthur & Glaze, Chapter 4 in this book), might in the future and with more sophisticated and diverse keyboards (with controls beyond just letters, numbers, and a touchpad) be used more broadly to measure a range of physical and psychological characteristics.

With accelerated development of technology allowing for the more standardized and reliable scoring of open-ended verbal protocols, we may be able to move away from reliance on multiple- or forced-choice formats, a move already in place for college admissions testing. Nor are we limited to open-ended written responses. Technology already exists that converts speech into verbal protocols with a high degree of accuracy. This allows open-ended responses to be captured orally rather than in writing before they are submitted to content analysis. This may prove particularly valuable in assessing populations with low literacy (see Malamut, Van Rooy, & Davis, Chapter 9 in this book).

More generally, we have to stop limiting our thinking to the tried-and-true item formats that have served us so well since World War I and challenge ourselves to think creatively about the new types of items we can design by leveraging technology. In some cases, these new formats will allow us to assess new constructs that could not be captured in traditional formats. In other cases, when we want to build on the accumulated evidence on our existing tools, we will need to empirically examine whether scores generated through these new formats demonstrate convergence with scores generated from traditional formats.

The Candidate

Demography

There has been some attention given to the impact of applicant demographics on technology-mediated assessment. One early issue, now becoming increasingly moot in the developed world, is access to technology. Virtually all segments within those societies have access—though not necessarily hands-on

experience and comfort—to the technologies used to deliver assessments (see Bartram, Chapter 7 in this book). In the developing world, differential access to technology along demographic lines of class, race, and gender will still play a role for the next decade.

Research on participant demographics should go beyond simply examining demographically related effect sizes on assessment tools or the moderating effect of demographic factors like race and gender on test validities. For a richer understanding of demographic effects, we should capture variables reflecting candidate state that might mediate the relationship between demographic predictors and test performance, variables like attention, anxiety, mood, and test performance motivation. We should be looking at interactive effects on test performance between demographics on the one hand and the medium with which test “items” are presented and with the medium of response, on the other. Technology has created more variance in assessment conditions and hence both a greater need and greater opportunities to study these potentially richer effects.

Even in the developed world, one demographic variable that is likely to have greater impact on technology-mediated assessment than on traditional forms of assessment is candidate age. Research demonstrates significantly greater use of social media and game technology by younger, as compared to older populations (Lenhart, 2009). Today’s teenagers—tomorrow’s entry-level job applicants—have grown up with these technologies as a natural part of their world, their “first language,” so to speak. As one indication, a recent survey of students entering U.S. universities in the fall of 2010 (Beloit College Mindset List, 2010) reported that fewer than 15 percent know how to write using cursive letters (assuming this is a global trend, so much for the future use of graphology for pre-employment assessment in countries like France, Germany, and Israel—and good riddance, at that!). Potential generational differences in reactions to, and performance on, technology-mediated assessments need to be more fully researched, particularly as gaming technology (including the use of avatars) is increasingly employed in assessment.

The Experience

We know that Realistic Job Previews reduce early turnover (Premack & Wanous, 1985). To the extent that technology allows for more high-fidelity experiences during the assessment process, the testing process itself can serve as a Realistic Job Preview in ways not possible in the past. After all, how many jobs involve tasks that require responding to short bits of written information by choosing between four prescribed alternatives?

In recent years, there has been increased use of video to bring situational items to life in Situational Judgment Tests (SJTs), instead of the traditional narrative descriptions (Weekly & Ployhart, 2006). However, most of these video-based SJTs use a multiple-choice format (Hense & Janovics, Chapter 12 in this book); the use of video recordings of open-ended responses described by Cucina, Busciglio, Thomas, Callen, and Walker (Chapter 13 in this book) is a noteworthy exception. The fidelity issue, then, can be framed separately for the stimulus and response sides of a test and it would be interesting to know the relative impact of these two factors on candidate perceptions. Perhaps there is an interaction effect; how do candidates react when there is a great gap between the fidelity of these two components, say when the items are presented vividly in a multimedia format while the candidate has to respond by choosing one of four fixed verbal alternatives. Perhaps making the stimuli more realistic just heightens the unrealism of the response format.

Building on the pioneering work of Gilliland (1993), Bauer, Truxillio, Mack, and Costa (Chapter 6 in this book) provide a justice theory-based lens through which to analyze applicant reactions to testing situations. Their approach makes useful predictions about the impact of the testing experience on the employee's attitude toward the organization, perceptions of the organization's culture, and likelihood of accepting a job offer. Much of what they discuss has not yet been examined empirically in the context of technology-enhanced assessment. In particular, they raise important issues surrounding the nature of assessment feedback and the impact of feedback on the applicant. Just because technically mediated assessment *can* produce immediate

feedback, *should* the applicant be given feedback immediately? Is the “cold” medium of a computer screen the most effective one for providing feedback? Or has our culture evolved to the point that nothing short of instant feedback is acceptable to applicants?

Gessner and Klimoski (2006) provide a different lens through which to analyze the applicant experience—as a “conversation” between the organization and the participant. That conversation takes place within a broad context. For an applicant, the context might include her prior experiences with the organization and its present and former members, including her experiences as a consumer of the organization’s products or services; the messages she received previously in earlier stages of the recruitment and application process; and her prior experience and success with similar assessment tools and media. This context colors how the applicant makes sense of the testing experience and, especially in a high-stakes environment, figures out how to meet the expectations of the organization through his or her behavior during the assessment.

This approach opens up many interesting avenues for research on technology-mediated assessment. Examples include:

- What are the factors related to previous experience in role plays or with games using avatars or in navigating an online in-basket modeled on Outlook or Notes that make an applicant feel advantaged or handicapped relative to other applicants? What are the consequences of those feelings?
- How does the ethnicity, age, gender, appearance, projected temperament, or mood of actors used in video-based assessments influence the candidate’s view of what the organization will evaluate as an effective response on the assessment? How will the actor’s ethnicity, age, or gender serve as identity cues that affect the salience of the candidate’s own social identity during the assessment process?
- Will the situation of taking an assessment in a testing center be different for an applicant if the others taking tests at the same time are construed as competitors for the same job or as a random collection of applicants for multiple positions in multiple organizations? What are the messages and cues that influence that construal?

Extending this approach further, and adopting the paradigm recently proposed by Weiss and Rupp (2011): What if we approached the applicant's behavior in the assessment situation from a person-centric perspective? This would create a new set of interesting questions. What does the experience of taking this assessment feel like to the applicant? What is the applicant thinking about through the experience? Is his mind wandering? Is she ruminating on some memory unrelated to the assessment or daydreaming? What does bored or captivated or "fired up" or fatigued or having the opportunity to shine feel like to the applicant? Does time feel like it is passing slowly or quickly? In explaining the assessment experience to others—including other potential applicants, family, or the recruiter—how does the applicant construct her personal narrative? How does that narrative reflect the personal identity the candidate is trying to build (for example, as tech-savvy or dutiful or "too cool to care" about the assessment outcome)? These and many others are truly psychological questions that as assessment specialists we should be, but generally have not been, deeply interested in.

A key open question that goes beyond behavior in the assessment context: Once applicants accept a job, to what extent do their beliefs about how they were originally treated during the selection process affect their later job performance or long-term tenure? How long does the effect persist and what are the factors that affect that persistence?

The Organization

Image

Individual organizations are expanding their answers to the fundamental question of what makes for a great assessment program. Surely the psychometric fundamentals—reliability, fairness, and evidence for the validity of inferences drawn from scores on the assessment—remain a core part of the answer. But the chapters in this book—especially the case studies—suggest that the design of great assessment programs goes beyond these fundamentals. Administrative ease and cost-effectiveness, to be sure, are part

of the answer, especially during periods of economic stress and resource constraints. But explicitly and implicitly the authors of our case studies have written about the use of technology to enhance the organization's employment brand. The assessment process itself is focused not only on screening out those candidates unlikely to succeed but in building the pipeline of qualified applicants. The process is expected to contribute to the image of the organization in the marketplace.

This image-making has at least two components, each of which requires greater research attention. First, how does the design and delivery of assessment add to or detract from the organization's ability to compete in the talent market? This is where the justice theory approach described by Bauer, Truxillo, Mack and Costa (Chapter 6 in this book) is most relevant. We have assumptions but few empirically supported answers to fundamental questions such as:

- Do applicants actually make inferences from their experiences during the selection process about the technological sophistication, respect for diversity, feedback richness, employee voice, and other features of the organization's climate, policies, and culture?
- What are the messages that applicants share with their social networks about their selection experiences? How are these messages conveyed? How, and how quickly, are these messages spread in an era when applicants can either directly or through their first-level networks communicate a positive or negative experience to thousands of potential recruits through Facebook, LinkedIn, blogs, or Twitter? How do these messages impact on the likelihood that qualified people in the labor market will actively apply for openings in the target organization or respond positively to recruitment overtures?

Second, we need to more systematically investigate the impact of the assessment process, thanks to the web now more visible to the public than ever before, on the image of the organization in the broader society.

- What aspects of the selection process have the strongest influence on organizational reputation?
- How can the assessment process reinforce an organization's corporate social responsibility?
- How does the assessment experience of rejected candidates affect their future patronage of the organization as customers?

Change Management

Each of the case studies presented here summarizes key success factors in the design and implementation of new technology-enhanced assessments. Reviewing these success factors, it is noteworthy how few relate to technical aspects of test design. It seems we've got that down pretty well as a profession. Most deal with the change management aspects of the assessment project: Project planning and management, identifying the key constituencies, stakeholder buy-in, sustained leadership support, creating change champions. However, as Muchinsky (2004) has noted, there is little research on the contextual factors that influence the success of implementing and sustaining newly designed assessment processes in organizations. There are a number of perspectives in the organizational change literature that could fruitfully be applied to better understand the organizational factors that can facilitate or obstruct the introduction of new assessment tools (Adler, Macan, Konczak, Muchinsky, Grubb, & Hurd, 2008). A few of these include:

- Viewing the new selection system as an expression of the organization's business strategy, vision, and/or culture
- Seeing the new selection tool as a way to implement change (if "the people make the place," different kinds of people will make the place different)
- Understanding organizational power and its application in shaping assessment-relevant policies and practices
- Being sensitive to the role of leadership succession in sustaining or terminating existing assessment programs for the sake of "out with the old, in with the new"

- Recognizing the organization—often through the human resources department—as an external sensing entity, imitating competitors or “best practice” companies in the adoption of new selection procedures
- Considering the extent to which values other than equity/meritocracy become a basis for assessment system design, for example, anti-unionism, perpetuating the culture, corporate social responsibility, publicity, diversity outreach, exclusivity
- Recognizing the role of change communication strategies to facilitate or undermine the implementation of new assessment procedures

Many of us have been in situations in which the legal or information technology department has had significantly more say than human resources or the assessment specialists in determining the content, scope, administrative constraints, level of investment, and other aspects of new assessment tools. Given the privacy, access, security, firewall, legal, cross-national, and other issues associated with technology-mediated assessments, the relative influence of these other organizational stakeholders in assessment design decisions is growing. We need to be cognizant of these broader organizational constraints and employ the models and tools that our organizational change colleagues have developed to enable sound professional practice in this area. These change management practices need to be grounded in a body of systematic research that examines the role these macro-level contextual factors play in program implementation and sustainability. At the moment, that research base is quite thin (Muchinsky, 2004).

Society

Sophistication and Visibility

Technology has not only changed the types of assessments delivered. It has greatly increased the magnitude of assessment activity. Using the web, organizations that had relied on interviews in the past now administer psychometrically sound assessment tools to hundreds of thousands of applicants each year. In that sense, technology has made sophisticated assessment more accessible

to organizations large and small, new and established. Tens of millions of applicants around the world are getting exposure to the best assessments our profession has ever produced. To make an obvious point, these applicants are also citizens and customers; technology has undoubtedly raised the public profile of what we do as assessment specialist. This may be a once-in-a-generation opportunity for our profession to be associated with “really cool stuff” that emerges from the combination of solid, credible assessment tools and leading-edge and highly engaging technology.

What will be the impact of this increased exposure to the products of our profession? With a freer, more open, and increasingly global marketplace of assessment solutions, will the test with the most sound psychometrics dominate or will that with the slickest technology become most popular? Perhaps a new hybrid profession will emerge, as it has in the field of learning and instructional design, that combines rigorous psychometric and I/O psychology education with strong multimedia and technology training.

Web Scouring

A recent report (Rosen, 2010) stated that 75 percent of recruiters and human resource professionals responding to a national survey in the U.S. indicated that they conduct online research on job candidates. Social network aggregator search engines can combine data from diverse and public online sources—YouTube videos, LinkedIn profiles and networks, Facebook pages and postings, blog contributions, political contributions, Twitter posts, published letters to the editor, public corporation filings, church bulletins, real estate listings, and much more—to present a comprehensive multi-faceted portrait of a job candidate in a neat portfolio. All this without even pressing the boundaries of privacy by accessing information on the books and films ordered or downloaded from Amazon and Netflix or borrowed from the public library or the music downloaded from iTunes. Rosen (2010) reminds us that the web doesn’t “forget”; over time, and for younger applicants, more and more of their lives will exist on the Internet.

As Zickar and Lake (Chapter 16 in this book) point out, we need more research on the degree to which this personal information on the web can be used to make valid inferences about a person's job-relevant skills, abilities, knowledge, and other personal characteristics. This question can be broken up into several discrete components.

One is whether the information on a person that emerges from a web search really reflects behaviors, choices, expressions of attitude and knowledge, and characteristics of that individual. It is not merely the question of whether a posting attributed to someone was actually authored by that person. To the extent that some social network sites have tight rules about what can be posted or powerful norms about what is acceptable content, those postings may reflect "strong situations" (Mischel, 1984) with limited diagnostic value for assessing individual differences. On the other hand, to the extent that the information found on a person is reflective of that person's preferences and choices, the Internet can be a rich naturalistic setting for data gathering. Gosling and his colleagues (for example, Back, Stopfer, Vazire, Gaddis, Schmukle, Egloff, & Gosling, 2010; Reis & Gosling, 2010) have shown that untrained observers can fairly quickly make valid inferences about someone's personality from the appearance of the person's dormitory room, photograph (for example, narcissists are more likely to wear expensive, flashy clothing and—for females—wear makeup and show cleavage), or Facebook page.

A second component of the question of validity here is whether the evaluator is using a structured set of guidelines that produces reliable assessments that map validly to target constructs. Undoubtedly, in Gosling's research, an important reason that observers are able to validly assess personality traits using data from natural habitats is that these observers are using a common structure to capture their evaluations, a well-validated set of items that much prior research had shown to reflect the target personality traits. Recruiters scouring the web for information on a candidate would need similarly well-constructed guidelines if they hope to extract valid ratings of target constructs.

A third component of the validity question, of course, is whether the constructs assessed by recruiters from information

scoured from the Internet are themselves relevant to criteria—performance, retention, safety, etc.—of interest. This involves classical criterion-related validation.

My point here is that the question of the validity of information scoured on candidates from the Internet needs to be disaggregated and studied in more subtle ways.

Of course, societies may decide and legislate that, valid or not, scouring the Internet for personal information without the explicit consent of applicants is illegal. Out of conviction or image-building (or both), organizations might voluntarily adopt those constraints even in the absence of legislation. Our profession might take a principled stance one way or another on these practices. This is certainly a question that should be debated and discussed at our conferences and meetings. I am hopeful these debates and discussions will be based on a growing body of research and a clear articulation of the competing values in play (for example, privacy versus the satisfaction and success of a better matched new hire).

Answering the Open Questions: Exploitation and Exploration

Bamberger and Pratt (2010) draw on the organizational strategy literature to define two approaches to stimulating interesting and valuable research. I would like to conclude by urging all—practitioners and academics alike—to consider either or both approaches to answering open questions around technology-enhanced assessment. The two approaches are exploitation and exploration.

Use of the exploitation-based approach would involve applying existing theory to “enhance or extend understanding of a given theory’s boundary conditions, to assess the practical implications, or to . . . support theoretical convictions in search of powerful evidence” (Bamberger & Pratt, 2010, p. 668). This book provides several examples of exploitation-based research on technology-enhanced assessment, drawing on justice theory, item response theory, and other theoretical models and approaches. Theories of flow, personal narratives, impression management, social comparison, self-regulation,

and others could all productively use the Internet-based assessment situation as a fertile arena for testing, expanding, and extending theory (Reis & Gosling, 2010). The assessment setting is both naturalistic and at the same time controlled to a degree. Technology-enabled assessment typically generates a great deal of information from thousands of applicants, making such assessment settings a rich source of data with which to test enhancements and extensions of our models.

Exploration-based research, in contrast, looks at unique or different phenomena as they occur in the field and draws implications back to our theories and models. As assessment professionals increasingly leverage advances in technology, they will create new types of assessments, new ways of delivering assessment, new forms of test security, new behavioral contexts within which assessments will be delivered. These in turn can shed new light on current models and force us to revise our models or create new ones. Let us look at just two examples.

Using technologies like Hewlett-Packard's Halo, team members can be located hundreds of miles apart from each other and yet be "seated" around a single conference table, looking at the same document "lying" on that table, maintaining eye contact with life-size images of each other, and converse as naturally as if they were together in the same room. How similar is the behavior of participants in a leaderless group discussion exercise conducted in that environment to what we know of the behavior observed and assessed in such exercises over the decades in "bricks and mortar" assessment centers? Do new or different behaviors emerge and, if so, where do they fit in our competency taxonomies?

Similarly, to take another classic assessment center exercise, how do the strategies used by a participant to address a digital inbox (see McNelly, Ruggenberg, & Hall, Chapter 8 in this book)—with emails, voice mails, instant messages, and other stimuli arriving at irregular intervals throughout—compare to the strategies used by participants in the traditional in-basket? Are these exercises measuring the same constructs? What do these new forms of classic assessment center exercises tell us that is new and different about our models of managerial decision making or our definitions of organization and planning skill?

A second illustration. Technology enables us to gather much more than the responses needed to compute a mean score for each applicant on each scale. Technology can be used to measure response latencies that in some cases may reflect socially desirable responding (Arthur & Glaze, Chapter 4 in this book). Technology can also capture the configuration of responses on a scale. There are many response patterns that can produce a candidate score close to the scale mid-point. Is there meaning to the fact that two candidates, identical in their mean score on Conscientiousness, responded very differently across the twelve items measuring that construct? Are those differences random sampling error or substantive? Employers with large-scale assessment programs are collecting test data on thousands of candidates a day. Given these huge sample sizes, we are now in a position to systematically explore whether there is regularity and job-related meaning to the variations and complex configurations in response pattern.

We certainly do live in interesting times. The increased use of technology raises many fascinating questions. The most valuable answers will come from systematic, theory-guided, nuanced research that examines the mediating processes that link variations in assessment design and administration to outcomes of practical and scientific interest.

References

- Adler, S., Macan, T., Konczak, L., Muchinsky, P., Grubb, A., & Hurd, J. (2008). I meets O: Implementing new selection systems as change management. *Industrial Organizational Psychologist, 45*, 3, 21–26.
- Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science, 21*, 372–374.
- Bamberger, P. A., & Pratt, M. G. (2010). Moving forward by looking back: Reclaiming unconventional research contexts and samples in organizational scholarship. *Academy of Management Journal, 53*, 665–671.
- Beloit College Mindset List. (2010). Beloit College Student Mindset List: Class of 2014. www.beloit.edu/mindset. Accessed August 2010.

- Gessner, T. L., & Klimoski, R. J. (2006). Making sense of situations. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 13–38). Mahwah NJ: Lawrence Erlbaum Associates.
- Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review*, *18*, 694–734.
- Kluger, A. N., Reilly, R. R., & Russell, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology*, *76*, 889–896.
- Lenhart, A. (2009). *Teens and social media: An overview*. Pew Internet and American Life Project. www.pewinternet.org. Accessed July 2010.
- Messick, S. (1985). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- Mischel, W. (1984). Convergences and challenges in the search for consistency. *American Psychologist*, *39*, 351–364.
- Muchinsky, P. M. (2004). When the psychometrics of test development meets organizational realities: A conceptual framework for organizational change, examples, and recommendations. *Personnel Psychology*, *57*, 175–209.
- Premack, S. L., & Wanous, J. P. (1985). A meta-analysis of realistic job preview experiments. *Journal of Applied Psychology*, *70*, 706–719.
- Reis, H. T., & Gosling, S. D. (2010). Social psychology methods outside the laboratory (pp. 82–114). In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 1) (5th ed.). Hoboken, NJ: John Wiley & Sons.
- Rosen, J. (2010, July 25). The end of forgetting. *The New York Times Magazine*, 32–37, 44, 45.
- Ryan, A. M., & Tippins, N. T. (2009). *Designing and implementing global selection systems*. Malden, MA: Wiley-Blackwell.
- Weekley, J. A., & Ployhart, R. E. (2006). *Situational judgment tests*. Mahwah NJ: Lawrence Erlbaum Associates.
- Weiss, H. M., & Rupp, D. E. (2011). Experiencing work: An essay on a person-centric work psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *5*.
- Wittenbrink, B., & Schwarz, N. (2007). *Implicit measures of attitudes*. New York: Guilford Press.
- www.cognadev.com. The Cognitive Process Profile. Accessed August 2010.

Name Index

A

Abraham, J., 132
Adams, J. S., 195
Adler, S., 17, 418, 429
Algina, J., 153
Ali, L., 396
Allen, M. J., 153
Alliger, G. M., 120, 128
Andersen, D. K., 407
Anderson, B., 403
Anderson, C. D., 121, 122
Anderson, D., 55
Anderson, M. G., 121
Anderson, N., 194, 226, 356
Anderson, N. R., 202
Anseel, F., 209
Ansley, T., 178
Arena, D., 406
Armstrong, T., 191
Arthur, W., Jr., 15, 99, 102, 109, 110, 111, 118,
119, 120, 121, 122, 123, 124, 125, 127, 129,
131, 134, 138, 139, 140
Arvey, R. D., 194, 206, 217, 400
Ash, S. R., 217
Ashton, M. C., 131, 327
Avedon, M. J., 192, 199

B

Back, M. D., 399, 400, 432
Bagozzi, R. P., 403, 405
Bailenson, J. N., 406
Baker, R. C., 211
Bamberger, P. A., 433
Bansal, V. K., 407
Barrett, C., 102, 108, 109
Barrett, G. V., 119
Barrett, P., 122, 129
Barrick, M., 193
Bartram, D., 16, 114, 176, 177, 194, 224, 225,
226, 229, 231, 232, 233, 235, 236, 238, 241,
368, 369, 370
Bateson, J., 368, 370
Bauer, T. N., 15, 190, 192, 194, 195, 199, 200,
206, 211, 213, 214, 215, 217
Baydoun, R., 48

Beatty, J., 364
Beaty, J., 107, 175, 204
Beaty, J. C., 56, 128, 191, 208, 380
Beaty, J. C., Jr., 102, 108, 109
Becker, D. F., 48
Behling, R., 93
Bejar, I. I., 32
Bennett, M., 211
Berners-Lee, T., 225
Bertolino, M., 211, 213, 214
Bertua, C., 356
Bevill, S., 401
Bies, R. J., 195
Biga, A., 158, 181, 204
Bikeland, S. A., 121, 122, 125, 127
Biskin, B. H., 48
Blitz, D. L., 49
Blumental, A. J., 190
Boben, D., 225
Bobko, P., 407
Bodner, T. E., 211
Borg, I., 51
Borman, W. C., 132
Borneman, M. J., 126
Bott, J., 122
Bowers, C., 409
Bowler, W., 17, 355
Boyle, S., 47
Bracy, C., 401
Bradley, P., 126
Brannick, M. T., 121, 122, 125, 127
Brennan, R. L., 61, 106
Brice, T., 209
Broadbent, M., 93
Brown, A., 238, 370
Brown, D. J., 128, 190, 217
Browning, H. L., 403
Brysse, K., 208
Buchanan, T., 42
Buchsbaum, M. S., 403
Bugbee, A. C., Jr., 47
Burke, E., 17, 104, 105, 113, 117, 355, 363, 365,
366, 367, 368, 370
Burnett, J., 102, 108, 109
Burns, D. S., 401

Burns, G. N., 131, 327
 Busciglio, H. H., 17, 338
 Butera, H., 131, 132
 Buyse, T., 56
 Byrne, B. M., 45, 46, 51, 52

C

Callen, N. F., 17, 338
 Champion, J. E., 400
 Champion, M. A., 120, 122, 192, 194, 195, 199,
 200, 203, 206, 211, 215
 Candell, G. L., 52
 Canli, T., 404, 405
 Cannon-Bowers, J., 409
 Canoune, H. L., 48
 Cappelli, P., 211
 Carroll, J. B., 182, 345
 Cartmill, J. A., 408
 Cash, T. F., 401
 Cellar, D. F., 121
 Chaharbaghi, K., 71
 Chan, D., 201
 Chang, H. H., 178
 Chang, H-H., 117
 Chang, S-H., 178
 Chernyshenko, O. S., 51, 52
 Cheuh, D., 403
 Chmielowski, T., 125, 129, 141
 Christiansen, N. D., 122, 126, 130, 131, 327
 Cizek, G. J., 101, 102, 103, 104, 106, 107, 114, 137
 Clause, C., 201
 Cober, R. T., 190, 217
 Cohen, R. L., 195
 Colquitt, J. A., 211
 Connelly, B. S., 121, 122
 Constable, R. T., 404
 Converse, P. D., 131, 132
 Corbett, J., 409
 Cormack, B., 406
 Cosman, P. H., 408
 Costa, A. B., 15, 190
 Coyle, B. W., 215
 Coyne, I., 225, 229
 Craig, J., 194, 195, 199, 200
 Cregan, P. C., 408
 Crocker, L., 153
 Cronbach, L. J., 130, 235
 Crowne, D. P., 126
 Cucchiarelli, A., 246
 Cucina, J. M., 17, 128, 131, 132, 338
 Cureton, E. E., 370
 Curtin, P. J., 128

D

Dale, L. R., 401
 Davey, T., 28, 359

Davis, J., 56
 Davis, V. A., 16, 267
 Dawson, C. R., 191, 208, 380
 Day, D. V., 135, 192, 194, 199, 201
 De Ayala, R. J., 153
 De Corte, W., 208
 de Fruyt, F., 356
 Decker, D. L., 120
 Del Vecchio, D., 217
 Delbridge, K., 201
 Delery, J. E., 120
 Deller, J., 122
 DeShon, R. P., 120, 201
 Dew, A. F., 211
 Di Paolo, N. T., 55, 108
 Dickter, D. N., 67
 Dietvorst, R. C., 403, 405
 Dilchert, S., 122
 Dineen, B. R., 217
 Dipboye, R. L., 122
 Do, B., 110, 111, 119, 138
 Dodd, B. G., 178
 Dolen, M. R., 199
 Donovan, J. J., 122, 132, 176
 Dorsey, D. W., 71
 Doverspike, D., 121, 122, 190
 Downey, K., 355
 Drasgow, F., 28, 46, 47, 48, 49, 50, 51, 52, 56,
 105, 106, 107, 110, 111, 112, 117, 118, 119,
 138, 141, 154, 175, 201, 204, 229, 328, 356,
 364, 386
 Drauden, G., 194, 217
 Dulebohn, J. H., 94
 Dumnette, M. D., 121, 127
 Dwight, S. A., 120, 122, 128, 132, 176

E

Eaton, N. L., 121, 127
 Edens, P. S., 120, 122
 Edwards, B. D., 119
 Edwards, J. R., 134
 Egloff, B., 399, 400, 432
 Ehrhart, K. H., 217
 Eisenberg, N., 135
 Ellingson, J. E., 121, 122, 130
 Ellis, A.P.J., 120
 Ellis, B. B., 52
 Embretson, S. E., 158
 Erdogan, B., 213, 214
 Estabrook, L., 57
 Evers, A., 225
 Eysenck, H. K., 127

F

Fallow, S. S., 191, 208, 380
 Fallon, J. D., 102, 108, 109

- Farah, M. J., 404, 405
 Feild, H. S., 193, 385
 Fekken, G. C., 128
 Fernández-Hermida, J. R., 225
 Ferrara, P., 194, 195, 199, 200
 Fetzer, M., 17, 380, 381
 Feurer, R., 71
 Fine, S., 110, 111, 119, 138
 Fitzgerald, C., 140
 Fix, C., 370
 Flanagan, J. C., 341
 Fleishman, E. A., 362
 Fletcher, J. D., 46
 Fletcher, P.A.K., 73
 Fluckinger, C. D., 122
 Foster, C. T., 128
 Foster, D., 112, 116, 117, 119, 364
 Fox, J., 406
 Fox, S., 57
 Frey, S., 48
 Furnham, A., 130
- G**
- Gaddis, S., 399, 400, 432
 Gallagher, A. G., 407
 Gateley, M. T., 408
 Gatewood, R., 193, 385
 Georgiadou, E., 178
 Gerrard, M. O., 55
 Gessner, T. L., 426
 Gibbons, A. G., 75
 Gibbons, R. D., 15, 37, 38
 Gibby, R. E., 153, 158, 181, 204
 Gibson, W. M., 56, 175, 204, 364
 Gillen, B., 401
 Gillespie, M. A., 133
 Gilliland, S. W., 15, 192, 193, 194, 195, 199, 200,
 203, 206, 208, 209, 211, 213, 214, 215, 216,
 217, 364, 425
 Glabeke, K., 225
 Glas, C.A.W., 154
 Glasgow, S., 401
 Glaze, R. M., 15, 99, 102, 109, 110, 111, 118, 123,
 124, 125, 127, 129, 131, 134, 138, 139, 140
 Goffin, R. D., 126, 130, 131
 Goldberg, L. R., 42
 Golden, J. H., 102, 108, 109
 Goldenberg Schoepfer, R. J., 17, 338
 Goodstein, J., 406
 Gosling, S. D., 399, 400, 432, 434
 Gough, H. G., 126
 Granz, J., 396
 Grauer, E., 56
 Gray, J. R., 406
 Graziano, W. G., 121, 135
 Greaud, V. A., 47
- Green, B. F., 47, 156, 157
 Griffith, R. L., 125, 129, 141
 Griggs, R. A., 386
 Groth, M., 211
 Grubb, A., 16, 429
 Grubb, A. D., 293
 Grubb, W. L., 120
 Gueutal, H. G., 75
 Guilford, J. P., 128, 133
 Guilford, J. S., 128, 133
 Guion, R. M., 39, 40
 Guo, J., 105, 106, 110, 112, 117, 118, 138, 356
 Gupta, D., 124, 139
 Guttenberg, J., 225
- H**
- Haas, B. W., 404, 405
 Haier, R. J., 403, 404
 Hair, E. C., 135
 Hale, J., 364
 Hall, C. R., Jr., 253
 Hall, C. R., 16, 71
 Halpert, J. A., 55, 108
 Hambleton, R. K., 25, 153, 166, 226,
 229, 233
 Handler, L., 229
 Hanson, B. A., 106
 Hanvongse, A., 121
 Harms, H. J., 155, 156
 Harold, C. M., 120, 217
 Harris, D. J., 106
 Harris, M., 191, 206
 Harris, M. M., 194
 Hartog, S. B., 16, 307
 Harvey, A. L., 47
 Hausknecht, J. P., 55, 108, 192, 194, 199, 201
 Hayes, S. C., 217
 Hazlett, E., 403
 Hedge, J. W., 93
 Hedricks, C., 131, 132
 Heggstad, E. D., 111, 131
 Heiphetz, A., 408
 Hemingway, M. A., 124, 139, 204
 Hense, R., 17, 102, 108, 109, 324
 Hetter, R. D., 178
 Hezlett, S. A., 345
 Hirsh, J. B., 131
 Ho, C., 121
 Hogan, J., 122, 126, 129
 Hogan, R. T., 122, 126, 129
 Holden, R. R., 128
 Holland, W. H., 40
 Hollenbeck, J. R., 122
 Holtz, B. C., 45
 Honey, P., 314, 316
 Horrigan, J., 57

Horrigan, J. B., 190
 Horvath, M., 51
 Hough, L. M., 121, 122, 123, 126, 127, 129, 130,
 131, 132
 Howard, K., 44
 Huang, L., 121
 Hunter, J. E., 356
 Hunthausen, J. M., 192
 Hurd, J., 429
 Hurtz, G. M., 122

I

Impara, J. C., 140, 364
 Imus, A., 131, 132
 In-Sue, O., 357
 Inwald, R. E., 126
 Ispas, D., 158, 181, 204

J

Jackson, D. N., 126, 131, 327
 Janovics, J., 17, 324
 Jaxtheimer, J. W., 408
 Jensen-Campbell, L. A., 135
 Jodoin, M. G., 32
 Johnson, J. A., 42
 Johnston, N. G., 130
 Jones, A., 132
 Judge, T. A., 122
 Jung, R. E., 403, 404

K

Kacmar, K. M., 120
 Kaminski, K. A., 124, 139, 204
 Kamp, J. D., 121, 127
 Kantrowitz, T. M., 17, 191, 208, 380
 Keeping, L. M., 217
 Kehoe, J. F., 67
 Keil, F. C., 406
 Kemp, C. F., 45
 Kerkvliet, J. R., 176
 Kiesler, S., 49
 Kim, B. H., 133
 Kingsbury, G., 140
 Kisamore, J. L., 121, 122, 125, 127
 Klawnsky, J. D., 121
 Klimoski, R. J., 400, 426
 Kluger, A. N., 132
 Kolen, M. J., 61
 Kolotkin, R. L., 48
 Konczak, L., 429
 Koong, K. S., 190
 Kriska, S. D., 199
 Kristof, A. L., 134
 Kroner, D. G., 128
 Kunce, C., 120, 133
 Kuncel, N. R., 126, 345

L

Lacasse, L., 403
 Laffitte, L. J., 45, 51, 52
 Lake, C. J., 17, 394
 Lance, C. E., 51
 Langdon, J. C., 211
 Lange, S. R., 124, 139
 Lautenschlager, G. J., 45, 49
 Leaman, J. A., 128
 Lemley, R. E., 401
 Lenard, G., 397
 Lenhart, A., 424
 Leonard, J. A., 128
 Leung, K., 241
 Levashina, J., 120
 Leventhal, G. S., 195
 Levin, R. A., 53, 122
 Levine, M. V., 38
 Levy, P., 190
 Levy, P. E., 217
 Leyhe, E. W., 48
 Liang, X., 94
 Liden, R. C., 215
 Lievens, F., 56, 111, 191, 192, 194, 201, 206,
 208, 209
 Ling, J., 217
 Liu, C., 51
 Liu, L. C., 190
 Livingston, G., 57
 Lönnqvist, J.-E., 130, 131
 Lord, F. M., 153, 154, 184
 Lott, I., 403
 Lucke, S. B., 121, 131
 Lyne, R., 135
 Lyon, J. S., 141

M

Macan, T., 429
 Macan, T. H., 192, 199
 McBride, J. R., 155, 164, 165
 McCloy, R. A., 15, 121, 127, 131, 153, 158,
 181, 204
 McDaniel, M. A., 120, 122, 345, 403
 McElreath, J. M., 128, 131, 132
 McFarland, L. A., 120, 121, 122, 126, 131, 132,
 176, 199, 407
 Mack, K., 15, 190
 Maclim, T., 396
 Macmillan, D., 403
 McNelly, T., 16, 253
 McPhail, S. M., 38, 40, 41
 Maertz, C. P., 199
 Mahoney-Phillips, J., 17, 355
 Malamut, A., 16, 267
 Manson, T. M., 121, 122, 125, 127
 Margolis, A., 229

- Marler, J. H., 94
 Marlowe, D. A., 126
 Martin, C., 194, 217
 Martin, C. J., 408
 Martin, C. L., 215
 Matarazzo, J. D., 402–403
 Matesic, K., 225
 Maurer, S. D., 345
 Maynes, D., 140, 365
 Mazzeo, J., 47
 Mead, A. D., 15, 21, 38, 46, 47, 48, 49, 50
 Meade, A. W., 45, 49
 Meijer, R. R., 156, 160
 Messick, S., 358, 419
 Michaelis, W., 126
 Michels, L. C., 45, 49
 Miller, J. L., 53, 122
 Miller, M. L., 121
 Millsap, R. E., 192, 200
 Miranda, R., 73
 Mischel, W., 432
 Mitchell, T. W., 400
 Moag, J. S., 195
 Mol, S. T., 192
 Montgomery, G. E., 131, 327
 Moore, L., 120
 Moreno, K. E., 164, 180
 Morgeson, F. P., 122, 200, 203
 Moriarty, N. T., 55
 Moriarty-Gerrard, M. O., 108
 Morrison, M., 131
 Moscoso, S., 356
 Muchinsky, P. M., 429, 430
 Mumford, A., 314, 316
 Muñoz, J., 225
 Murphy, K., 122
- N**
- Nagao, D. H., 215
 Naglieri, J. A., 229
 Nering, M., 359
 Nering, M. L., 156, 160
 Neuechterlein, K. H., 403
 Neuman, G., 48
 Neustel, S., 55
 Newman, D. A., 141
 Nisbett, R. E., 401
 Novick, M. R., 153, 154
 Nye, C. D., 28, 105, 106, 110, 111, 117, 118, 119,
 138, 356
- O**
- O'Brien, M. K., 407
 O'Connell, M., 122
 Olson-Buchanan, J. B., 154, 201, 328
 Omura, K., 404
- Ones, D. S., 53, 122, 123, 130, 132, 345
 Oosting, A., 403
 Oswald, F. L., 51, 122, 130, 131, 132, 133
 Outtz, J. L., 40
 Overton, R. C., 155, 156
- P**
- Pace, V. L., 132
 Pack, P., 47
 Paek, J., 403
 Paese, M., 192, 199
 Palfrey, J. G., Jr., 397
 Panti, M., 246
 Paronto, M. E., 192, 206, 211, 215
 Parshall, C. G., 28
 Parsons, C. K., 215
 Pashley, P. J., 28
 Pastor, D. A., 178
 Paulhus, D. L., 101, 120, 126
 Paunonen, S., 130
 Pearlman, K., 56, 57, 107, 110, 175, 177, 192,
 200, 204, 364
 Peters, K., 190
 Peterson, J. B., 131
 Philips, H. L., 128
 Pierre, J., 356
 Piotrowski, C., 191
 Pitoniak, M. J., 25
 Pleban, R. J., 408
 Ployhart, R. E., 51, 63, 194, 195, 199, 200, 211,
 217, 425
 Polly, L. M., 211
 Pommerich, M., 180
 Pomplun, M., 48
 Pophan, S. M., 128
 Pratt, M. G., 433
 Premack, S. L., 425
 Prifitera, A., 229
 Prinsloo, 422
 Pulakos, E. D., 93
- R**
- Rainie, L., 57, 190
 Raju, N. S., 45, 51, 52
 Ramsay, L., 133
 Rauschenberger, J., 215
 Rawson, E., 406
 Raymond, M. R., 55
 Reeve, C. L., 111, 131
 Reilly, M. E., 362
 Reilly, R. R., 128, 132, 192, 200
 Reis, H. T., 432, 434
 Reise, S. P., 158
 Reiss, A. D., 130
 Reynolds, D. H., 15, 24, 66, 75, 84
 Richman, W. L., 49, 50

Richman-Hirsch, W. L., 201
 Robertson, I. T., 356
 Robie, C., 121, 128, 130, 135, 141
 Robins, K. W., 135
 Robson, S. M., 132
 Roc, 231
 Roehling, M. V., 401
 Roerch, T., 401
 Rogelberg, S. G., 42
 Rogers, H. J., 153
 Rogers, J., 44
 Roman, S. A., 407
 Rosen, J., 431
 Rosse, J. G., 53, 122
 Roth, P. L., 407
 Rothstein, M. G., 130
 Roy, R., 131, 132
 Ruggenberg, B. J., 16, 253
 Rupp, D. E., 24, 75, 427
 Russell, C. J., 132
 Russell, D. P., 67
 Ryan, A. M., 51, 120, 121, 122, 126, 131, 132, 135,
 194, 195, 199, 200, 211, 419
 Ryan, M. A., 122

S

Sacco, J. M., 67, 199
 Sacher, J., 46
 Sackett, P. R., 56, 121, 122, 130, 194, 201, 206
 Salgado, J. F., 356
 Sanchez, R. J., 194, 195, 199, 200, 217
 Sandman, C., 403
 Sands, W. A., 155, 165
 Satava, R. M., 407
 Saupe, J. L., 106
 Scarborough, D. J., 135
 Schleicher, D. J., 200, 203
 Schmidke, J. M., 120
 Schmidt, F. L., 345, 356, 357
 Schmit, M., 229
 Schmit, M. J., 122
 Schmitt, N., 51, 120, 122, 130, 133, 201
 Schmukle, S. C., 399, 400, 432
 Schuler, H., 194, 209
 Schwarz, N., 422
 Scott, J. C., 15, 21
 Segall, D. O., 28, 37, 38, 56, 107, 164, 175, 176,
 178, 180, 204, 364
 Seymour, N. E., 407
 Shaffer, J. A., 357
 Shaw, J. C., 211
 Shephard, W., 204, 364
 Shepherd, W. J., 56, 102, 107, 108, 109, 175
 Shupe, C., 93
 Siegel, B. V., 403
 Sigmund, C., 176

Silverman, S. B., 135
 Silvester, J., 202
 Sinclair, R. R., 135
 Sireci, S. G., 25
 Slade, L. A., 51
 Smith, B., 130
 Smith, B. D., 130
 Smith, D. B., 122
 Smith, D. E., 192, 199
 Smith, M., 356
 Smith, M. A., 121, 122, 125, 127
 Smither, J. W., 192, 194, 200
 Smits, M., 403, 405
 Snell, A. F., 121, 131
 Snyder, L. A., 75
 Sosa, E., 403
 Spector, C. E., 121, 122
 Spector, P. E., 51
 Stanton, J. M., 42
 Stanush, P. L., 127
 Stark, S., 51, 52, 386
 Stecher, M. D., 53, 122
 Steiner, D. D., 194
 Stelly, D. J., 38, 41
 Stoffey, R. W., 192, 200
 Stone, D. L., 75
 Stopfer, J. M., 399, 400, 432
 Straetmans, G.J.J.M., 155
 Strickland, W., 194, 217
 Subramani, M., 93
 Swaminathan, H., 153
 Sydell, E. J., 121, 131
 Sylva, H., 193
 Sympson, J. B., 178

T

Tatham, N., 363, 367
 Tay, L., 28, 105, 106, 117, 118, 138, 356
 Taylor, J. E., 102, 109, 110, 111, 118, 123, 124,
 125, 127, 129, 131, 134, 138, 139, 140
 Taylor, L. R., 155, 156
 Teachout, M. S., 71
 Templer, K. J., 124, 139
 Tetrick, L. E., 128
 Tett, R. P., 121, 122
 Thibaut, J. W., 195, 203
 Thomas, P. H., 17, 338
 Thomas, S. C., 192, 194, 199, 201
 Tippins, N. T., 1, 56, 57, 58, 75, 107, 136, 158,
 175, 191, 194, 204, 205, 208, 231, 356, 364, 419
 Touchette, P., 403
 Triantafillou, E., 178
 Truxillo, D. M., 15, 190, 192, 194, 195, 199, 200,
 206, 211, 212, 213, 214, 215, 217
 Tucker, J. S., 213, 214
 Tuulio-Henriksson, A., 130

U

Uziel, L., 126

V

Valenti, S., 246

van Dam, K., 194

Van de Vijver, F., 241

Van der Linden, W. J., 153, 154, 166

Van der Lugt, A., 403, 405

Van Hoye, G., 194

Van Rooy, D. L., 16, 267

van Someren, G., 363, 367

Vandenberg, R. J., 51

Vasilopoulos, N. L., 128, 131, 132

Vazire, S., 399, 400, 432

Velasquez, R., 229

Venkataramani, Y., 200, 203

Verbeke, W. J. M. I., 403, 405

Verkasalo, M., 131

Verschoor, A. J., 155

Vessey, J., 44

Vierra, R., Jr., 120

Villado, A. J., 102, 109, 110, 111, 118, 120, 123,
124, 125, 127, 129, 131, 134, 138, 139, 140

Viswesvaran, C., 53, 122, 123, 130, 132

W

Wade, M. W., 107

Wainer, H., 40, 154

Walker, D. D., 17, 338

Walker, L., 195, 203

Wanous, J. P., 425

Wargin, J., 71

Warner, J. L., 121, 122

Waters, B. K., 155, 165

Waters, L. K., 131

Watts, S. M., 408

Waung, W., 209

Weathers, V., 213, 214

Weber, M., 71

Weekley, J. A., 63, 206, 215

Weekly, J. A., 425

Weill, P., 93

Weiner, J., 355

Weiner, J. A., 84

Weisband, S., 49

Weisberg, D. S., 406

Weiss, D. J., 37, 38, 42, 154, 164, 184

Weiss, H. M., 427

West, B. J., 120

Wetzel, C. D., 164

Whetzel, D. L., 122, 345

White, J. K., 215

Whitley, B. E., Jr., 102, 103, 105

Wild, E., 211

Williams, B. A., 38

Williams, D., 190

Witt, E., 57

Wittenbrink, B., 422

Woehr, D. J., 121

Wood, J. L., 120

Woodill, G., 408

Wroblewski, V. R., 131, 327

Wu, J. C., 403

Y

Yant, T. S., 121

Yen, W. M., 153

Yonce, C. A., 211

Yoo, T., 133

Yoon, C., 403, 405

Yoshita, Y., 125, 129, 141

Yun, G., 120

Z

Zaal, J. N., 225

Zbylut, M., 128

Zenisky, A. L., 25

Zhang, J., 117

Zickar, M., 17, 53, 394

Zickar, M. J., 121, 122, 141, 155, 156

Zimmerman, W. S., 128, 133

Zinnes, J. L., 386

Zukier, H., 401

Subject Index

Page references followed by *fig* indicate illustrated figures; followed by *t* indicate a table; followed by *e* indicate an exhibit.

A

- Academy of Management, 411
- Adapted tests: cultural factors to consider for, 231–341; IRT differential item functioning (DIF), 51–52; ITC Guidelines on Test Adaptation for, 233–234; language translation issues for, 50–51, 232–235, 277–278; SEM/MACS for, 51–52
- Adjacent technology systems, 83–84
- Administration: CAT proctored Internet-based testing, 157–177; CAT unproctored Internet testing (UIT), 175–177; consistency of technology-enhanced assessment, 9–10; questions and answers on issues of, 420–421; VBT (video-based test), 341–343
- Age differences, Internet usage rates, 57
- Aggregation of data: comparing scales between/within countries, 237; guidelines for, 235–237, 241; local vs. global norms, 234–237, 241; norm reference groups used for, 232, 235
- American Educational Research Association (AERA), 114, 164
- American National Standards Institute (ANSI), 245
- American Psychological Association (APA), 114, 164
- Americans with Disabilities Act (1990), 405
- APA *Standards*: on choice of performance criterion measures, 40; on equivalence, 43; on reliability, 33; on validity, 38–39
- APA Task Force on Testing on the Internet, 229
- Applicant reactions: future directions for research and practice, 217; Gilliland's model of, 194–195, 199; procedural rules for managing, 196 t –216; video-based test (VBT), 347–349 e . *See also* Candidates
- Applicant reactions procedural rules: consistency, 196 t , 206–208; feedback, 197 t , 208–210; job relatedness, 196 t , 200–202; listed, 195, 196 t –198 t ; non-technology-specific, 197 t –198 t , 214–216; opportunity to perform, 196 t , 203–205; propriety of questions, 197 t , 213–214; selection information and explanations, 197 t , 211–213
- Applicant tracking system (ATS): assessment systems integrated with, 3; Marriott International's use of, 270–271, 279 fig –280; widespread use of, 356
- Armed Services Vocational Aptitude Battery (ASVAB), 155, 170, 172, 178, 179–180
- Assessments. *See* Technology-enhanced assessments
- Association of Test Publishers (ATP), 367
- ATP Guidelines on Computer-Based Testing, 229
- Attributes: stimulus and response, 26 e –27 e ; Universal Competency Framework (UCF) measuring, 369–374 fig . *See also* Personality scales; Skills
- Automated assessment systems: common criteria for effectiveness of, 95 t ; common features of online, 85 t –86 t ; common modes of, 76; employee development programs support by, 77–78; environmental/organizational context of, 71–73, 74; executing implementation plan for, 88, 91–96; implementation framework used for, 67–71; planning for additional applications of, 78–79; supporting high-volume screening, 75–77; supporting hiring decisions, 77; talent systems and assessment context of, 73, 75; technical facilitators and constraints of, 79–84; technical requirements for, 84, 87; user requirements for, 87–88, 89 t –90 t ; vision statement and goals of, 73. *See also* Computer adaptive testing (CAT)
- Automatic Data Processing, Inc., 102

B

- Bandwidth, 80
- Bank branch manager case study: background information on, 324–325; developing new assessment process, 325–335; Global Personality Inventory-Adaptive (GPI-A),

- 326–327*t*; lessons learned and recommendations, 335–337; measuring success during, 335; mini-role-play interview, 332; MMSJT (multimedia situational judgment test), 328–329; organizational/political challenges during, 333–334; Supervisory Potential Index (SPI), 326; technological challenges during, 334–335; validation of new assessment used for, 329–332; video-based coaching STJ, 330*fig*–333
- BFCAT, 38
- Brain imaging; modern-day assessment using, 401–406; practical and ethical problems with, 404–405; techniques for, 402–404
- British Standards Institute (BSI), 245, 246
- BS7988 Code of Practice, 246
- C**
- Candidate pools: economic depression and higher KSOA levels in, 383; Internet access/digital divide impacting, 56–58; technology-enhanced assessments effects on, 6–8; unproctored Internet testing (UIT) effects on, 7–8
- Candidate screening: applicant tracking system (ATS) of, 3, 270–271, 279*fig*–280, 356; automated systems supporting high-volume, 75–77; on candidate experience, 425–427; digging for digital dirt form of, 395–413, 431–433; proctored Internet-based testing for, 104–106, 123–125, 175–177, 230; Realistic Job Previews, 425; user requirements for automated systems, 87–88. *See also* Case studies; Computer adaptive testing (CAT); Tests
- Candidate selection. *See* Technology-based selection
- Candidates: attribute information gathered from, 26*e*–27*e*; CAT format and preparation by, 159; cheating by, 12, 52–54, 100–116, 367, 385; communicating with, 197*t*–198*t*, 214–216; “conversation” between organizations and, 426; demography of, 423–424; economic depression and higher KSOA levels in, 383; executive-level, 253–266, 324–337; experience of, 425–427; Internet access and digital divide of, 56–58; learning styles of, 314, 316–317; low-literacy, 267–280; response distortion by, 52–54, 101, 123–131; retesting, 54–56, 108–111; technology-enhanced assessments and expectations of, 8–9. *See also* Applicant reactions
- Case studies: bank branch manager selection, 324–337; Darden Restaurants, Inc., 253–266; Federal Bureau of Investigation (FBI), 293–306; The Interpublic Group of Companies, Inc. (IPG), 307–323; Marriott International, Inc., 268–292; U.S. Customs and Border Protection (CBP), 338–354. *See also* Organizations; Technology-based selection; Technology-enhanced assessments
- Cheating: actions to take against perpetrators, 113–116; amount of, 102–103; CAT format making it difficult for, 157, 385; definition of, 101; interactive voice response (IVR) and, 12; as malfeasant behavior, 100–101; proctored versus unproctored Internet-based settings, 104–113; response distortion as form of, 52–54; VAT (Verify Ability Test) principles on, 367. *See also* Response distortion; Tests
- Cheating detection: positive evidence of cheating issue of, 114; score comparison and verification testing, 107–112; statistical, 106–107; technological, 112–113
- Cheating deterrence: monitoring for, 116–117; test design characteristics and features, 117–119, 131–135; warnings and threats as, 117
- Cognitive assessments: CAT-based, 384–386; equivalence in, 46–48; unsupervised Internet testing (UIT) for, 358–362. *See also* Knowledge tests
- Communication: applicants and two-way, 198*t*, 215–216; honest, 198*t*, 216; interpersonal effectiveness of, 197*t*, 214–216
- Computer adaptive testing (CAT): advantages of using, 24–25; cheating made more difficult through, 157, 385; considerations for development of, 164–179; considerations for the use of, 155–164; data analysis requirements for, 161–162; description of, 153–154, 193, 381; growth of, 355–356, 357; implementation and maintenance of, 387–390; legal implications for, 163–164; lessons learned from, 391–392; level of interactivity of, 28; National Institute for Educational Measurement development of, 155; notion of “controlled reliability” of, 385; ongoing maintenance required for, 162–163; organizational challenges of implementing, 382–383; relationship between response and item selection in, 153–154*fig*; security gap in, 359; success metrics of, 390–391; “whole person” approach of, 383–387*t*. *See also* Automated assessment systems; Candidate screening; Test design/formats
- Computer adaptive testing (CAT) tests: cognitive ability, 384–386; DoD’s Armed Services Vocational Aptitude Battery (ASVAB), 155, 170, 172, 178, 179–180; Graduate Record Examination (GRE), 155, 155*s*; hard skills, 384; P&G’s Computer-Adaptive Reasoning ASDF Test (CARAT), 155, 171, 178, 180–184;

- personality assessment, 386–387*t*; 16PF Questionnaire, 38
- Computer adaptive testing development: calculating information, 170–171; content type, 165–166; data requirements and item parameterization, 167–168; end-user requirements, 165; IRT model used in, 154–155, 166–167, 168–169; item timing, 172–173; measurement precision, 168–169; movement between items, 173; ongoing evaluation and management, 177–179; quality assurance, 173–174; starting/stopping rules, 171–172; supervised vs. unproctored delivery, 175–177; validation, 175
- Computer-Adaptive Reasoning ASDF Test (CARAT) [P&G], 155, 171, 178, 180–184
- Computer-Based and Internet Delivered Testing Guidelines* (TIC), 43
- Consistency rule: description of, 206; potential benefits of, 206; potential challenges, 206–207; recommendations for practice, 207–208
- Constructed response, 28
- Context, Constraints, and Requirements (CCR) analysis: environmental/organizational context of, 71–73, 74; executing the implementation plan for, 88, 91–96; framework for using assessment technologies for, 69–71; talent systems and assessment context of, 73, 74–79; technical facilitators and constraints, 79–84; technical requirements for the assessment, 84, 87; user requirements for, 87–88
- Cost issues: computerized adaptive testing (CAT), 160–161; technology-enhanced assessments, 4–6; unproctored Internet testing (UIT) and associated, 6; video-based test (VBT), 345–347*e*
- Cultural factors: affecting assessment, 231–232; choice of norm reference groups, 232; comparing scales between/within countries, 237–240; guidance on using local vs. global norms, 234–241; ITC Guidelines on Test Adaptation for, 233–234; Marriott's applicant assessment consideration of, 276–277. *See also* Language
- D**
- Dactyl Nightmare* (game), 406
- Darden Restaurants, Inc.: assessing critical job turns at, 255–256*fig*; continuous process improvements at, 265; lessons learned from experience of, 266; organizational and political landscape of, 253–256*fig*; success metrics and insights learned from, 263–265; virtual assessment center solution used by, 256–261; web-enabled assessment implementation/maintenance at, 261–263. *See also* Olive Garden; Red Lobster Delivery. *See* Test delivery
- Design. *See* Test design/formats
- Det Norske Veritas (DNV) [Norway], 244, 245
- Development. *See* Professional development
- DF (data forensic) algorithms, 364–365
- Digital dirt: cleaning up “dirt” industry, 396; ease and examples of scouring for, 395–396; ethical issues of digging for, 396–399; legality of using, 401; possible legal guidelines on practice of, 433; prevalence of online research for, 431–433; scientific issues of using, 399–401; validity issues of, 432–433
- Digital divide, 56–58
- DIN (national standards institute) [Germany], 244, 245, 247
- E**
- Editorial review, 29
- Elaboration strategy, 133
- Emotional intelligence, conceptualizing, 27
- Empirical keying, 132
- Employee development programs, 77–78
- Employer value proposition (EVP), 272–374*fig*
- Employers from Spying on Your Facebook* (Facebook group), 398
- Employers Using Facebook as a Background Check Is Wrong!* (Facebook group), 397
- Environment. *See* Technical environment
- Equal Employment Opportunity Commission (EEOC), 23, 30, 163, 164
- Equivalence: APA *Standards* on, 43; of cognitive assessments, 46–48; definition of, 41; MEQ (measurement equivalence), 45; of non-cognitive assessments, 48–49; protoctored retesting and issue of, 111; showing, 42–43; summary of issues and recommendations for, 49–50
- Equivalence designs: multiple-groups, 43–45; multiple-measurements, 45–46
- Ethical issues: brain imaging and, 404–405; of digging for digital dirt, 396–399; keeping up with the latest, 413. *See also* Legal guidelines
- European Association of Psychological Assessment (EAPA), 242
- European Federation of Psychologists' Associations (EFPA), 242
- European Test Review Criteria and Test User Standards (EFPS), 242–244
- Executive-level assessment: bank branch manager selection, 324–337; Darden Restaurants' web-based, 256–266;

organizational/political landscape context of, 253–256/fig
 Exploitation-based assessment approach, 433–434
 Exploration-based assessment research, 434–435

F

Face validity, 28–29
 Facebook, 395, 396, 397–398, 410, 431
 Federal Bureau of Investigation (FBI):
 automated assessment design objectives, 297; description of, 293–294; discrimination lawsuit brought against the, 294–295; lessons learned by, 303–306; LSAs (leadership skills assessments) developed for, 297–303; redesigned promotion process used by, 295–296
 Feedback rule: description of, 197*t*, 208–209; potential benefits, 209; potential challenges, 209; recommendations for practice, 210
 Financial Services Authority (FSA), 372–374
 Firewalls, 80
 fMRIs (functional magnetic resonance), 405
 Forced-choice response formats, 131
 Foster Item, 119
 FSA's Treating Customers Fairly (TCF), 373–374

G

Global Personality Inventory-Adaptive (GPI-A), 326–327*t*
 Globalization: of assessment through technology, 225–226; growth of the web interface impacting, 226–227. *See also* International assessment issues
 Graduate Record Examination (GRE), 155

H

Harvard Business Review, 398
 Hiring. *See* Technology-based selection
 Honesty procedural justice rule, 198*t*, 216
 Hourly eHiring System (Marriott International): centers of expertise (COE) used in, 281–283; development and validation of, 275–280; implementation process and challenges of, 280–286; integration of HR systems with, 278–280; key stakeholders of, 281; lessons learned from, 290–292; success measurement of, 286–290
Human Resource Executive magazine, 411
 Human resources: human resources outsourcing (HRO), 267–268; talent management responsibilities of, 356, 357
 Human resources transformation initiative (HRT): hourly (non-management) staffing challenge of, 268–270; Marriott International's, 267–292

I

Implementation framework: adjacent technology systems, 83–84; assessment context/talent systems considerations, 73, 75; Context, Constraints, and Requirements (CCR) analysis using, 69–71; executing the implementation plan, 88, 91–96; goals establishing for, 73; infrastructure characteristics and, 79–81; introduction to, 67–71; organizational characteristics considered for, 72–73, 74; software deployment model and, 81–82; technical requirements for assessment, 84, 87; technical support, 82–83; understanding the environment for, 71–72, 73; user requirements, 87–88, 89*t*–90*t*
 Implementation plan execution: acquiring third-party software for, 92; building custom software for, 91–92; case study on M&A integration assessment, 94; issues to consider for, 88, 91; managing a successful, 93–96
 Intellectual property (IP), 244–245
 Intelligence (IQ) testing, 403–404
 Interactive voice response (IVR): administrative ease and flexibility of, 12; cheating issues of, 12; description of, 193; multiple-choice format of, 2–3; testing materials presented through, 2
 International assessment issues: choice of norm reference groups, 232, 235; comparing scales between/within countries, 237–240; cultural factors affecting, 231–232; guidelines for technology-based, 228–231; ISO standards relating to, 245–247; legal and professional standards and guidelines, 241–244; local versus global norms, 234–241; Marriott International's approach to, 267–292; protection of intellectual property, 244–245; test adaptation guidelines, 233–234. *See also* Globalization
 International Bureau of Weights and Measures, 41
International Journal of Testing (ITC), 229, 242
International Prototype Kilogram, 41
 International Reputation Management, 396
 International Task Force on Assessment Center Guidelines, 257
 International Test Commission (ITC): computer-based tests guidelines by, 76*e*; Guidelines on Test Adaptation, 233–234; *International Journal of Testing* by, 229, 242; international testing guidelines by, 115, 117, 228–231; Test Use Guidelines on standards/legal issues, 241–242
 Internet access, 56–58
 Interpersonal effectiveness, 197*t*, 214–216
 The Interpublic Group of Companies, Inc. (IPG): assessment needs of, 307–308;

- challenges facing, 308–310; Learning Styles Questionnaire (LSQ) used at, 314; lessons learned by, 321–323; MyLead assessment program developed at, 310–319; MyLead results at, 319–321
- I/O psychology: combining technological training with, 431; ethical issues of digital dirt context of, 398–399; utility of digital dirt context of, 401
- IRT. *See* Item response theory (IRT)
- ISO 9126 standard, 245–246
- ISO 10667 standard, 247
- ISO (International Organization for Standardization), 245–247
- ISO23988 standard, 246–247
- ITC Guidelines*: computer-based tests guidelines by, 76*e*; international testing guidelines, 115, 117, 228–231; Test Use Guidelines on legal/standards guidelines, 241–242
- ITC Guidelines on Test Adaptation, 233
- Item format: definition of, 28; editorial review of, 29; Foster Item, 119; pretesting or piloting, 29–30, 31*e*; two types of, 28
- Item response theory (IRT): applied to cognitive and non-cognitive measures, 370; CAT development using, 154–155, 166–167, 168–169; CAT requirement of expertise in, 162; description of, 32, 33; differential item functioning (DIF), 51–52; ideal point paired comparison basis of, 386–387; LOFT model using, 359–360*fig*; to replace single index of reliability, 33–38
- Items: CAT approach to calculating information through, 170–171; CAT development of, 166–173; differential item functioning of, 40; international guidelines on adaptation of, 233–234; P&G's Computer-Adaptive Reasoning ASDF Test (CARAT), 182–183. *See also* Reliability; Tests; Validity
- Iterative item linking, 52
- J**
- Job analysis, 275
- Job relatedness procedural rule: description of, 1964, 200; potential benefits of, 201; potential challenges, 201–202; recommendations for practice, 202
- K**
- Knowledge tests: amount of cheating during, 102–103; CAT-based, 384; cheating in proctored vs. unproctored Internet-based, 104–106; cognitive ability measured by, 101; detection of cheating during, 106–116; deterrence of cheating during, 116–119. *See also* Cognitive assessments
- KSAOs: economic depression and higher levels of candidate, 382; evaluating psychometric validity of measures, 39
- L**
- Language: ITC Guidelines on Test Adaptation of, 233–234; Marriott's applicant assessment consideration of, 277–278; norm reference group used for testing, 232, 235; translation issues, 50–51. *See also* Cultural factors
- LEADdR (Aon), 257
- Learning management systems (LMS), 356
- Learning styles: Learning Styles Questionnaire (LSQ), 314; MyLead program (IPG) application of, 316–317
- Legal guidelines: Americans with Disabilities Act (1990), 405; computer adaptive testing (CAT), 163–164; discrimination lawsuit against FBI, 294–295; international and national standards and, 241–244; keeping up with the latest, 413; for measuring relevant criteria, 23–24; piloting items within context of, 30; possible evolution of digital dirt, 433. *See also* Ethical issues
- Legend Quest* (game), 406
- Lie scales, 125–127
- LinkedIn, 396, 431
- Local access computers, 80–81
- LOFT (linear-on-the-fly-testing) model: description of, 359–360, 366; schematic summary of, 360*fig*
- Low-literacy applicants: applicant tracking system (ATS) used for, 270–271, 279*fig*–280; Marriott's heart-of-house web-based assessment of, 272–275*fig*; Marriott's Hourly eHiring System for, 275–292; Marriott's transformed HR approach to, 267–268
- LSAs (leadership skills assessments) [FBI]: benefits to the FBI, 302–303; description and process of, 297–300; development and validation of, 300–302; lessons learned from use of, 303–306
- M**
- M&A integration assessment, 94
- Malfeasant behavior, 100–101
- Management and Graduate Item Bank (MGIB), 361
- Marriott International, Inc.: applicant tracking system (ATS) used by, 270–271, 279*fig*–280; assessment development and validation process of, 275–280; heart-of-house web-based assessment by, 272–275*fig*; Hourly eHiring System of, 275–292; hourly (non-management) staffing challenge of, 268–270; HRO (HR outsourcing) used by, 267–268; lessons learned by, 290–292; measuring success at, 286–290;

- strategic HR transformation initiative (HRT) by, 267–268; system implementation process and challenges, 280–286
- MCAT, 38
- Mean and covariance structures (MACS), 51–52
- Measurement equivalence (MEQ), 45
- Measurement opportunities, 30
- Measurements: bank branch manager case study, 335; building assessment specifications for, 25–29; CAT need for precise, 168–169; cheating, response distortion, and retesting, 52–56; common criteria for automated assessment effectiveness, 95*t*; computer adaptive testing (CAT) for, 24–25; computer adaptive testing (CAT) success, 390–391; conduct editorial review/pretest the items for, 29–30; Darden Restaurants' web-based executive selection, 263–265; developing assessment plan for, 24–25; noncognitive tests, 119–135; of relevant criteria, 23–24; reliability of, 30, 32–38; SEM (standard error of measurement), 33, 112; standardization and equivalence of, 41–50, 41–52; successful implementation role of, 94–96; validity, 38–41, 56. *See also* Technology-based assessment evidence
- Media inclusion, 28
- Messick's model for validity, 370
- Mini-role-play interview, 332
- Miniwatts Marketing Group, 227
- MMSJT (multimedia situational judgment test), 328–329
- Monitoring testing, 116–117
- MRI (magnetic resonance imaging), 402, 405
- Multiple-groups equivalence designs, 43–45
- Multiple-measurements equivalence designs, 45–46
- MyLead program (IPG): challenges facing IPG leading to, 308–310; design process used to develop, 310–311; innovations of the, 316–319; lessons learned from development of, 321–323; results of the, 319–321; structure of the, 311–316*fig*
- MySpace, 396
- N**
- National Council on Measurement in Education (NCME), 114–115, 164
- National Institute for Educational Measurement, 155
- The Netherlands' Psychological Association, 244
- Noncognitive tests: amount of response distortion in, 121–123; description of, 119–120; deterring response distortion, 131–135; inconsistency responding, 127–128; lie scales, 125–127; measures used in, 119–135; response distortion in proctored vs. unprotected, 123–124; response latencies, 128; score comparison and verification testing, 124–125; statistical detection and control, 128–131
- Norm reference groups, 232, 235
- O**
- Occupational Personality Questionnaire (OPQ), 49
- Office of Federal Contract Compliance, 164
- Olive Garden, 254. *See also* Darden Restaurants, Inc.
- Online assessment. *See* Unproctored (unsupervised) Internet testing (UIT)
- Opportunity to perform (OTP): description of, 196*t*, 203; potential benefits, 203–204; potential challenges, 204–205; recommendations for practice, 205
- Organizational/political challenges: automated assessment systems, 71–73, 74; bank branch manager case study and, 333–334; computer adaptive testing (CAT) and, 382–383; executive-level assessment, 253–256*fig*
- Organizations: automated assessment systems in context of, 71–73; bank branch manager assessment organizational/political challenges, 333–334; change management adopting technology-enhanced assessments, 429–430; characteristics of, 72–73; “conversation” between candidates and, 426; how assessment approach impacts image of, 427–429. *See also* Case studies; Technical environment
- P**
- PC Magazine*, 411
- Personal digital assistants (PDAs), 2
- Personality scales: CAT-based, 386–387*t*; digging for digital dirt form of, 395–413; Global Personality Inventory-Adaptive (GPI-A), 326–327*t*; impact of warnings, verification, and threats on, 132; MMSJT (multimedia situational judgment test), 328–329; Occupational Personality Questionnaire (OPQ), 49; response distortion during testing, 52, 54, 101, 120–125; Supervisory Potential Index (SPI), 326. *See also* Attributes; Scalings
- PET (positron emission tomography), 402, 405
- Pilot test analyses, 29–30, 31*e*
- Political issues. *See* Organizational/political challenges
- Principles for the Validation and Use of Personnel Selection Procedures* (SIOP), 163
- Procter & Gamble (P&G), 155, 171, 178, 180–184
- Proctored Internet-based testing: CAT administration of, 175–177; cheating in

- unproctored versus, 104–106; description of, 104; International Test Commission (ITC) guidelines, 230; response distortion in unproctored vs., 123–125
- Professional development: Darden Restaurants's web-based assessment and, 253–266; FBI's web-based assessment and, 293–306; IPB's web-based assessment and, 307–323; Marriott's web-based assessment and, 268–292; video-based coaching STJ of bank managers, 330*fig*–333
- Profile matching, 133–135
- Propriety of questions rule: description of, 197*t*, 213; potential benefits, 213; potential challenges, 213–214; recommendations for practice, 214
- Psychometric theories, 32–33
- R**
- Racial/ethnic Internet usage rates, 57
- Random sampling theory, 32–33
- Randomized test construction (RTC), 117–118, 176
- Realistic Job Previews, 425
- Reasoning Screen (Procter & Gamble), 155, 171, 178, 180–184
- Reconsideration opportunity, 198*t*, 216
- Red Lobster, 254. *See also* Darden Restaurants, Inc.
- Reliability: APA *Standards* on, 33; establishing, 30, 32–33; item response theory on, 32, 33; notion of CAT “controlled reliability,” 385; random sampling theory on, 32–33; UCF framework, 370; video-based test (VBT), 343–344. *See also* Items
- Remotely delivered assessments. *See* Unproctored (unsupervised) Internet testing (UIT)
- Reputation Defender, 396
- Reputation Management Consultants, 396
- Response action, definition of, 28
- Response distortion: as form of cheating, 101; how to respond to, 128–131; lie scales to detect, 125–127; noncognitive test cheating in, 120–124; problem of, 52–54; proctored versus unproctored testing, 123–125. *See also* Cheating
- Response distortion detection: inconsistency responding for, 127–128; lie scales for, 125–127; response latencies for, 128; score comparison/verification testing for, 124–125
- Response latencies, 128
- Retesting: associated with increased test scores, 108–111; cheating associated with, 54–56; equivalence issue of, 111; *ITC Guidelines* on disclosure of procedures, 115
- S**
- Scalings: incomparable, 52; iterative item linking to make comparable, 52. *See also* Personality scales
- Score of record (SOR), 363, 364
- Scoring algorithm, definition of, 28
- Search for 1,000,000 People who Agree That Facebook Shouldn't Effect* (sic) *Jobs* (Facebook group), 398
- Security: CAP (computer adaptive testing) gap for, 359; CAT advantage for, 157–158; International Test Commission (ITC) guidelines, 230; retesting as detrimental to, 54–56; technology-enhanced assessment test materials, 10–11; UIT and framework for, 364, 365*fig*–366; unproctored Internet testing (UIT), 5, 10; VAT (Verify Ability Test) principle on, 367
- Selection information/explanations rule: description of, 197*t*, 211; potential benefits, 211–212; potential challenges, 212; recommendations for practice, 212–213
- SEM (standard error of measurement): cheating detection through, 112; description of, 33
- SEM (structural equations modeling), 51, 52
- Service-level agreement (SLA), 83
- Situational judgment tests (SJTs), 193
- 16PF Extraversion, 38
- 16PF Questionnaire, 38
- Skills: assessment of candidate, 27*e*; economic depression and higher levels of candidate, 382; evaluating psychometric validity of KSAOs measures, 39; knowledge tests on, 101–119, 384. *See also* Attributes
- SMEs (subject matter experts), 335
- Social networking sites: digging for digital dirt using, 395–396; tracking latest technological advances through, 410
- Society for Human Resource Management, 411
- Society for Industrial and Organizational Psychology, 163, 398–399, 411
- Society for Industrial Psychology Conference (2008), 356
- Software: acquiring third-party, 92; building custom, 91–92; technical support of, 82–83; user requirements for, 87–88, 89*t*–90*t*. *See also* Technical environment
- Software deployment model, 81–82
- Software-as-a-service (SAS), 355
- Speeded tests, 118–119
- Stakeholders: Hourly eHiring System (Marriott International), 281; measurements used to secure buy-in of, 96; successful implementation role of, 93–94; UCF to meet concerns of, 367–368
- Standards for Educational and Psychological Testing*, 114–115, 164

Standards/standardization: American National Standards Institute (ANSI), 245; British Standards Institute (BSI), 245, 246; BS7988 Code of Practice, 246; characteristics of, 42; Det Norske Veritas (DNV) [Norway], 244, 245; equivalence of, 41–50; European Test Review Criteria and Test User Standards (EFPS), 242–244; German DIN (national standards institute), 244, 245, 247; importance of psychological measures, 41; International Bureau of Weights and Measures, 41; *International Prototype Kilogram*, 41; International Test Commission (ITC) guidelines, 76*e*, 228–231; ISO (International Organization for Standardization), 245–247; ISO23988 standard on test delivery, 246–247; *ITC Guidelines* on legal guidelines and, 241–242

Statistical cheating detection, 106–107

Stop Businesses from Facebook Stalking You (Facebook group), 397–398

Stop Nosy (sic) (Facebook group), 398

Supervisory Potential Index (SPI), 326

Sustainability: change management grounded in, 430; of company culture and service standards, 269; P&G's Computer-Adaptive Reasoning ASDF Test (CARAT), 183–184; as software implementation goal, 67, 70

SWOT analysis, 258

T

Talent management systems (TMS), 356, 357

Technical environment: adjacent technology systems, 83–84; automated assessment systems, 71–72; bank manager selection case study, 334–335; common features of online assessment, 85*t*–86*t*; example for automated assessment systems, 74; infrastructure characteristics, 79–81; International Test Commission (ITC) guidelines, 229; software deployment model, 81–82; software requirements for assessment, 84, 87; technical support, 82–83; user requirements and, 87–88. *See also* Organizations; Software

Technical support, 82–83

Technology: combining I/O psychology with training in, 431; globalization of assessment through, 225–226; growth of the Web, 226–227; infrastructure of, 79–81; sophistication and visibility of, 430–431

Technology confidant, 412

Technology infrastructure: bandwidth, 80; firewalls, 80; implementation dependence on, 79–80; local access computers, 80–81

Technology-based assessment evidence: establishing reliability of, 30, 32–38;

establishing validity, 38–41, 56; evaluating job relatedness, 40–41. *See also* Measurements

Technology-based assessment plan: for automated system applications, 78–79; building assessment specifications, 25–29; developing the, 24–25; executing the implementation of, 88, 91–96; flexibility of, 70; pretesting or piloting items, 29–30, 31*e*

Technology-based assessment specifications: building, 25–29; incorporating face validity in, 28–29; item format, 28–31*e*, 119; what to include in, 28

Technology-based selection: automated assessment systems supporting, 77; consistency consideration of, 206–208; examining how job applicants perceive, 191–193; for executive-level, 253–266; how it is perceived by applicants, 190–218; increasing use of, 190–191; job relatedness consideration of, 200–202; opportunity to perform consideration of, 203–205; propriety of questions asked during, 213–214; providing applicants with feedback during, 208–210; providing information/explanations on the, 211–213; technologies available for, 193–194. *See also* Case studies

Technology-enhanced assessment pros/cons: administrative ease and flexibility, 11–12; candidate expectations regarding, 8–9; cheating issues, 12–13, 14; consistency of administration and scoring of, 9–10; costs associated with, 4–6; effect on quality and quantity of candidate pool, 6–8; security of test materials, 10–11

Technology-enhanced assessments: adaptation and language translation issues, 50–52; advancements in, 1–2; brain imaging approach for, 401–406; change management process of adopting, 429–430; cultural factors affecting, 231–232; description and uses of, 24; digging for digital dirt practice of, 395–401; ensuring quality of measurements, 21–56; equivalence of cognitive, 46–48; equivalence of non-cognitive, 48–49; for executive-level selection and development, 253–266; exploitation and exploration questions on, 433–435; fairness of, 56–58; future of, 13–14; implementing, 66–97; international issues related to, 224–248; keeping up with the latest, 410–413; M&A integration, 94; nuances of, 419–420; pros and cons of using, 4–13; sophistication and visibility of, 430–431; types of available, 193–194; underlying measurement principles for, 23–30; virtual reality, 406–410. *See also* Assessments; Case studies; Technology-enhanced assessments

- Test delivery: CAT advantages for, 159; common modes of computer-based, 76; impact of the online revolution on, 228; ISO23988 standard on, 246–247; proctored Internet-based testing, 104–106, 123–125, 175–177, 230; technology-enhanced, 2–4. *See also* Unproctored (unsupervised) Internet testing (UIT)
- Test design/formats: advantages and disadvantages of CAT, 156–164; cheating detection through, 117–119, 131–135; elaboration, 133; empirical keying, 132; equivalence, 43–46; forced-choice response, 131; Foster Item, 119; interactive prompts or cautions, 132–133; International Test Commission (ITC) guidelines, 230–231; profile matching and nonlinear models, 133–135; questions and answers on issues of, 421–423; randomized test construction (RTC), 117–118, 176; speeded tests, 118–119; warnings, verification, and threats, 132. *See also* Computer adaptive testing (CAT)
- Test information function, 33–34
- Test scores: cheating detection by comparing, 108–111; detecting response distortion by comparing, 124–125; fewer scoring errors with CAT, 157; lie scales, 125–127; reconsideration opportunity for applicants to review their, 198*t*, 216; retesting associated with increasing, 108–111; score of record (SOR), 363, 364; video-based test (VBT) reliability and, 343–344
- Testing International* (ITC newsletter), 229
- Testing materials: common delivery modes of computer-based, 76; international guidelines on adaptation of, 233–234; technology-enhanced delivery of, 2–4
- Tests: cheating detection by verification, 107–112; cognitive ability measured by knowledge, 101–119; Computer-Adaptive Reasoning ASDF Test (CARAT), 155; intelligence (IQ), 403–404; international guidelines on adaptation of, 233–234; item format of, 28–30, 31*e*; noncognitive, 119–135; procedural rules for managing applicant reactions to, 196*t*–216; proctored Internet-based testing, 104–106, 123–125, 175–177, 230; reliability of items, 30, 32–38; response distortion results of, 52–54, 101, 120–131; retesting following, 54–56, 108–111, 115; situational judgment tests (SJTs), 193; validity of items, 38–41, 56. *See also* Candidate screening; Cheating; Items; Unproctored (unsupervised) Internet testing (UIT)
- TIC *Computer-Based and Internet Delivered Testing Guidelines*, 43
- Translation/back-translation (TBT), 50
- Treating Customers Fairly (TCF) [FSA], 373–374
- Twitter, 396
- Two-way communication, 198*t*, 215–216
- U**
- Uniform Guidelines on Employee Selection Procedures* (EEOC), 163, 345
- Unit of analysis, 234
- Universal Competency Framework (UCF): competency profile of, 369, 371; EVP (employer value proposition) met by, 272–374*fig*; reliability of, 370; six dimensions of, 369; three performance dimensions of, 371–372; validation of, 372
- Unproctored (unsupervised) Internet testing (UIT): candidate expectations and, 9; candidate pool effects by, 7–8; CAT administration of, 175–177; CAT advantage for secure, 158; cheating issues of, 12–13, 14, 104–113; considering fairness of, 58; consistency of, 196*t*, 206–208; costs associated with, 6; description and advantages of, 100, 380; DF (data forensic) algorithms to verify, 364–365; drive for cognitive ability, 358–362; employer value proposition/candidate experience challenges of, 367–374*fig*; International Test Commission (ITC) guidelines, 230–231; large pool of items required for, 28; LOFT model used for, 359–360*fig*; 366; opportunity to perform (OTP) rule of, 196*t*, 203–205; questions regarding value of, 22; response distortion in proctored vs., 123–125; security issues of, 5, 10, 364, 365*fig*–366; targeted at graduate or campus recruitment, 357–358; UCF (Universal Competency Framework) taxonomy used for, 369–374*fig*; VAT (Verify Ability Test) extension of, 362–367. *See also* Test delivery; Tests
- U.S. Customs and Border Protection: assessment needs of, 338; VBT (video-based test) developed by, 338–354
- U.S. Department of Defense (DoD), 154–155, 170, 172, 179–180
- User requirements, 87–88, 89*t*–90*t*
- V**
- Validity: bank branch manager assessment process, 329–332; CAT development of, 175; definition of, 38–39; effect of retesting on, 56; evaluating job relatedness, 40–41; evaluating psychometric, 39–40; LSAs (leadership skills assessments) [FBI], 301–302; Marriott's applicant assessment approach to criterion, 278; Messick's model for, 370; P&G's Computer-Adaptive Reasoning ASDF Test (CARAT), 183; question of online candidate

- research, 432–433; video-based test (VBT), 344–345. *See also* Items
 - Verification testing: detecting cheating by, 107–112; detecting response distortion by, 124–125
 - Verify Ability Test (VAT), 262–267
 - Verify Verification Test (VVT), 363
 - Video-based assessments: description, 193; Video-based coaching STJ, 330*fig*–333; video-based test (VBT), 338–354
 - Video-based test (VBT): administration of the, 341–343; applicant and test user reactions, 347–349*e*; cost-effectiveness of, 345–347*e*; description of, 339–340; eight steps of, 340–341; examining CBP’s development of, 338–339; lessons learned and suggestions for, 349–353*e*; present use of the, 343; scoring and reliability of, 343–344; validity and adverse impact, 344–345
 - Virtual reality assessment: conclusions about, 409–410; description of, 406–407; research on, 407–409
- W**
- Web patrols, 245
 - Web-based management simulations, 193
 - Wired* magazine, 411
- Y**
- YouTube, 431