Michael Fetzer · Kathy Tuzinski *Editors*

Simulations for Personnel Selection

Springer

Simulations for Personnel Selection

Michael Fetzer • Kathy Tuzinski
Editors

Simulations for Personnel Selection

Springer

Editors
Michael Fetzer
CEB
Alpharetta, GA
USA

Kathy Tuzinski
CEB
Minneapolis, MN
USA

ISBN 978-1-4614-7680-1 ISBN 978-1-4614-7681-8 (eBook)
DOI 10.1007/978-1-4614-7681-8
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013943115

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

As a scientist, practitioner, technologist, and futurist, simulations have held a special place in my world for a long time. In my first week of graduate school, I found myself assigned to write the inaugural paper of my career on the subject of work samples. Though I knew almost nothing about personnel selection at that time, my research led me to Asher and Sciarrino (1974), an article that immediately had a strong intuitive appeal and still remains a foundational element in my philosophy of personnel selection. In this article, the authors introduce the concept of "point to point correspondence," a concept suggesting that prediction is enhanced when predictors are designed to be miniature replicas of the criterion space.

The central concept of giving an applicant a set of tasks that have fidelity with the job for which they are being evaluated not only makes intuitive sense and provides strong face validity, but also sets the stage for superior measurement. For these reasons, work samples (aka simulations) are both our past and our future. They have been around for a long time and are in no danger of extinction anytime soon. Be it 1950 or 2050, simulations provide significant value because they can simultaneously engage candidates in a way that is not possible with traditional tests, provide a realistic job preview, reliably and accurately measure the complex interplay of traits required to perform a job, and provide insights on which targeted individual training and development programs can be based.

Though the core concept that underlies the value of simulations has not changed over the years, the jobs that they are designed to measure certainly have. In the not too distant past, manufacturing dominated the economy and the bulk of jobs tended to draw less on a complex interplay of traits and autonomous decision making and more on monotonous tasks requiring more physical traits such as dexterity, strength, spatial perception, and mechanical aptitude. Such traits are relatively easier to measure reliably and accurately via straightforward tests that reside mostly in the physical realm.

Although many traditional work sample tests are just as useful today as they were 50 years ago, one cannot argue with the fact that technology has led to major changes in the world of work. With each passing decade an increasing number of the manual, physical tasks that were once core to a great number of jobs are now done by machines. These machines include computers, a technology that has had perhaps the biggest influence in history on the nature of industry and the jobs required to support it. In today's workplace, be it a factory, a call center, or a mine, almost every worker must interact with both people and technology in order to perform their jobs effectively.

There is no doubt that the massive changes created by advances in technology have been, and will continue to be, transformative. The overall influence of technology has changed more than just jobs and the nature of work. It is changing society as a whole by creating new levels of efficiency, connectivity, and access to information. In today's world it seems that things we may never have thought possible are becoming a reality on an almost daily basis. The world of work is no exception as technology is driving a major shift in the nature of jobs and the skills required to perform them.

In his remarkable book, *The Physics of the Future*, Michio Kaku, a futurist and world renowned physicist, uses factual scientific information to predict what we can expect in the near future. The author discusses the future of the workforce and suggests that by mid-century (2030–2070), almost all lower-level jobs will be automated. He goes on to suggest that the types of jobs that will not be automated will be those that require "the one commodity that robots cannot deliver: common sense." To me this is a very important message because it suggests that at some point in the not too distant future, the tools we use to make hiring decisions will require much deeper levels of insight to be effective. In short, this prediction suggests that selecting workers in the future will require a newer, more complex and insightful set of tools and clearly represents the writing on the wall for the personnel selection industry. Now is the time for us to begin exploring the predictive tools of the future and I believe that the star of the show will be our old friend, the work sample/simulation.

The good news is that, just as it has with almost every other industry under the sun, technology has also radically transformed the testing industry. The duties of my first job as a personnel selection intern at a high tech telecom company in the mid-90s included receiving faxed test answer sheets, using plastic scoring keys to score penciled in bubbles, and tracking these scores in a homemade Excel file. It took a good deal of my effort to obtain data that was usable by hiring managers for decision making. The progress in 20 years' time has been nothing short of incredible. We are now able to configure, deploy, and analyze predictors in the blink of an eye and to feel quite confident when using the data they provide to make employment decisions.

In just a short amount of time we have arrived at a point where both scientists and practitioners alike are comfortable with the technology-based delivery of assessment tools. This movement has been a huge force in advancing our field because it has opened the floodgates for an ever increasing number of firms to realize the benefits of quality selection science. The testing industry has moved further in the past decade than it did in the time between its inception and the dawn of internet delivery. Simply reading the chapters in *Technology-Enhanced Assessment of Talent*, a book I consider to be a prequel of sorts to this one, shows that we have quickly gained clarity on some key issues related to the use of technology in employment testing (i.e., candidate reactions, cheating, faking, equivalence, access, etc.). The benefits of technology are now a given in our industry and we have begun to see that the most successful personnel selection firms are those that include equal parts technology and science in their practice and products.

Despite all the excellent strides we have made in both practice and research, it is not a stretch to posit the idea that the state of evolution and change both in the world of work and within the testing industry is vastly outpacing the state of personnel selection research. I feel that this gap is actually the largest when it comes to simulations. Currently the state of the art for simulations includes jobs where the work environment is easy to replicate using technology and jobs where interactions can be modeled using avatars. Examples include contact centers where agents are using a series of programs and screens all day, manufacturing where a computer controlled panel can allow adjustments, and some parts of professional level jobs that are easy to replicate with technology such as in-baskets.

There is no doubt that the value of currently available simulations is well understood and that they are providing tremendous benefits for those who use them. However, while animated situational judgment exercises are flashy, provide a more engaging experience, and superior measurement over their paper and pencil counterparts; they are usually constrained to a scripted interaction. Simulations that provide high levels of prediction via point to point correspondence will need to be more complex because job environments and the role humans play within them are becoming increasingly complex. I picture the simulations of the future offering more wide open, "sandbox" environments in which an almost infinite number of interactions and outcomes are possible.

To approach modeling and measuring work interactions in a more realistic way, I believe that we need to look towards the more sandbox-like world of gaming. Gamification is a very real phenomenon because games provide entertainment and have a knack for making even boring tasks more interesting. In 20 years' time almost everyone entering the workforce will have been raised during an era when highly complex and sophisticated games are available on demand, any place, at any time. This will set a level of comfort and even expectation that will drive demand for games to be relevant in an increasing number of places.

I believe it a foregone conclusion that simulations are likely to become increasingly game-like in order to leverage technology to replicate work environments. I am not saying that for simulations to be useful, they must be game-based. Rather, I use games as an example because they provide a realistic window to the future of simulations. One in which all types of jobs can be modeled and the interactions and behaviors that define job performance can occur in a fluid and realistic manner.

Unfortunately, as flashy and interesting as games may be, they also highlight a major gap between simulation technology and research/science. This gap is based on the fact that at present, our ability to reliably draw trait-based inferences from unstructured simulated interactions simply does not exist. In order to make meaningful predictions based on the measurement of constructs required for job performance, we still must create simulations that are highly structured and scripted. Although such simulations still have tremendous value and can offer excellent levels of predictive accuracy, they remain limited in many ways. Truly simulating a work environment requires the ability to review data from a more wide open and realistic game-based simulation and use it to trace behaviors displayed back to specific traits that can be reliably measured. When it comes to this end, the research is just not there to support a practice of selection science that can use the best qualities of games that look and feel like those used for entertainment to drive the creation of realistic simulations. The Holy Grail for simulations is the ability to make such inferences based on wide open and unstructured interactions within complex simulated work environments of all types.

For example, in a simulation (or game), it may be possible to extract reams of data indicating how long it took someone to accomplish something or the specific path they chose while completing a task. At present, many simulations (but few games) track the outcomes of highly structured simulated interactions and the ability to manage simple communications directly back to a set of specific competencies required for job performance. So it seems that at present, the use of simulations forces us to choose between raw empiricism that does not provide sound trait-based measurement and highly structured and less fluid simulations, that while measuring important traits, place limitations on realism and complexity.

I believe that the future lies in bridging this gap. I also believe that the first step requires us to draw from the timeless insight provided by Landy (1986) who espoused that we seek to find ways to see the relationship between predictor and criterion not via one constrained path guided simply by data or via theoretical research alone, but rather through a more balanced, comprehensive, and insightful program of research. Success in the realm of simulations will require a multidisciplinary approach in which science, technology, and psychology all play an equally central role in modeling and understanding the links between predictors and criteria.

Given the rapidly increasing pace of technology and the dearth of research related to selection science and advanced simulations, I am thrilled to be part of the publication of this book. At the very least, the chapters in this book provide a strong foundation for our understanding of the current state of the art job simulations, provide an excellent "how-to" for those interested in developing simulations, and provide justification for the value and existence of simulations. Beyond this, I hope this book provides inspiration for the creation of new streams of research and practice that can occur concurrently and which leverage the core concept of point to point correspondence to new levels of measurement and insight. It is specifically my dream to see selection scientists not as stamp collectors, but rather as collaborators playing a role in forging new pathways for psychological measurement by figuring out how to reliably and accurately draw trait-based inferences from highly realistic and more wide open simulations.

Finally, it is my sincere hope that this book will have an impact not just on the field of personnel selection, but on society as a whole by increasing the ability to help companies hire in a way that will allow them to realize increased productivity and profitability via a happy and engaged workforce.

New Orleans, Louisiana Charles Handler
March 2013

References

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519–533.

Kaku, M. (2011). The physics of the future: How science will shape human destiny and our daily lives by the year 2100. NewYork: Anchor Books.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*(11), 1183–1192.

Tippins, N. T., & Adler, S. (Eds.) (2011). Technology- enhanced assessment of talent. San Francisco, CA: Wiley.

Acknowledgments

This book would not have been possible without the undying enthusiasm, relentless dedication, and invaluable contributions of our chapter authors. I would also like to thank my beautiful wife Erin, devoted daughter Camryn, and awesome son Cade, whose unconditional support is worth more than words can say.

(Michael Fetzer)

A special thank you goes to the contributors to this book. We couldn't have done it without you. I would also like to acknowledge the wonderful, zany support of SU and MTD, the two most important peeps in my life.

(Kathy Tuzinski)

Contents

About the Contributors

Seymour Adler is a Partner in the Performance, Talent, and Reward Practice at Aon Hewitt, and received his Ph.D. in Industrial-Organizational (I-O) Psychology from New York University. For over 35 years, he has directed the development and implementation of talent assessment and talent management programs for major private and public sector organizations, with particular emphasis on the design and validation of simulations. He has published widely in both the scientific and professional literature. In addition to his work as a practitioner, Seymour has served throughout his career on graduate I-O faculties; and currently he is an Adjunct Professor at Hofstra University. He is a Fellow of SIOP, APS, and APA and Past President of the Metropolitan New York Association of Applied Psychology. He is the coeditor of Technology-Enhanced Assessment of Talent (2011) in the SIOP Professional Practices series.
e-mail: seymour.adler@aonhewitt.com

Renee Barr is a Senior Consultant with CEB where she develops and implements client assessment programs to optimize quality of hire and legal defensibility. She earned her Ph.D. from the University of Houston's Industrial-Organizational Psychology program. She has developed a wide range of assessment programs including online selection testing, behavioral interviews, simulations, assessment centers, and executive coaching across a wide range of industries and job levels. Her previous roles included Assessment Solution Manager at First Advantage where she was responsible for updating and implementing software skill simulations.
e-mail: renee.barr@shl.com

Anthony Boyce leads assessment and selection innovation activities at Aon Hewitt, where he is a Senior Consultant. In this capacity, he has been responsible for the creation of more accurate, efficient, engaging, and globally relevant assessment and selection processes. Most recently, his efforts have been focused on building and deploying computerized adaptive tests of both cognitive and non-cognitive constructs, as well as building the capability to efficiently develop and deliver media-rich, web-based job simulations. In addition to his research and development activities, he consults with Fortune 500 clients to help them implement large-scale talent management initiatives. Dr. Boyce received his Ph.D. in Industrial-Organizational Psychology from Michigan State University, has published articles

in several leading journals, and is a regular contributor to the annual conference of the Society for Industrial and Organizational Psychology.
e-mail: anthonyboyce@gmail.com

Valentina Bruk-Lee is an Assistant Professor of Industrial-Organizational Psychology at Florida International University. Her research is primarily focused on applicant reactions to technologically advanced assessments, multitasking ability, conflict in organizations, and occupational health and well-being. Dr. Bruk-Lee has consulted with both private and public sector organizations in the areas of employee selection, job analysis, performance measurement, and competency modeling. She has also led the development and localization of assessments used in global selection programs. Her research has been published in journals such as Human Performance, Journal of Occupational Health Psychology, and Work and Stress. Dr. Bruk-Lee has presented her work at various recognized national conferences, including the annual meeting of the Society for Industrial and Organizational Psychology.
e-mail: Vblee@fiu.edu

Mary Ann Bucklan is the Director of Research at Employment Technologies Corporation, where she participates in the design and development of state-of-the-art selection and development technologies, with a focus on simulation-based products. She was a lead developer on the EASy Simulation® Virtual Customer suite of simulations, which won the Top HR Product award from HR Executive Magazine (2011). She has conducted research in the areas of simulation validity, applicant reactions to assessment, and structured employment interviewing. She has extensive consulting experience in financial services, call center, telecommunications, semiconductor, and manufacturing environments. She earned an M.S. in Industrial-Organizational Psychology at the University of Central Florida. She is coauthor of The Job Analysis Kit (1995, with Sandra McIntire and Deonda Scott).
e-mail: Maryann.B@etc-easy.com

Jacqueline Carpenter is a Consultant with Shaker Consulting Group, Inc. and a Ph.D. candidate in the Industrial-Organizational Psychology program at the University of Akron, where she also earned an M.A. in Industrial-Organizational Psychology. Her research interests include corporate image perceptions, organizational attraction, and applicant reactions, particularly in the contexts of technology-enhanced assessment for selection and web-based recruitment.
e-mail: Jackie.Carpenter@shakercg.com

Christine Corbet is a Senior Consultant with Right Management's Talent Management Practice in New York City. She has worked across a broad array of industries to help organizations select, retain, and develop high-quality talent at all job levels. She has considerable experience in the areas of pre-employment selection test development and implementation, executive coaching, training facilitation, and competency modeling. Her work has also included the delivery of large-scale organizational change initiatives and the redesign of performance management systems. She holds a B.S. in Psychology with a Citation in Honors from Marywood University where she earned the Psychology Department Medal. She did her graduate work in the

Industrial-Organizational Psychology Ph.D. program at the State University of New York's (SUNY) University at Albany; it was there that she received the prestigious President's Award for Excellence in Teaching. She is an active member of the Society for Industrial and Organizational Psychology.
e-mail: christine.corbet@gmail.com

Christopher Coughlin is a Research Scientist on the Product Development and Innovation team at CEB. In this role, he leads the development, validation, and implementation of software and computer skill simulations. Prior to joining CEB, he worked on organizational development initiatives at Spherion, a Randstad company. He earned his B.S. in Psychology from the University of Georgia and his M.S. in Industrial-Organizational Psychology from Valdosta State University. He is a member of the Society of Industrial and Organizational Psychology, Society for Human Resource Management, and the American Psychological Association.
e-mail: Chris.Coughlin@shl.com

Christie Cox Brodbeck is a Senior Consultant with Shaker Consulting Group, Inc. and earned her Ph.D. from the University of Akron. She has worked with clients to develop and implement a variety of talent management solutions, including performance management tools, executive assessments, large-scale surveys, and web-based pre-employment assessments and simulations. Prior to joining Shaker Consulting Group, Christie was the program and research manager for the global employee selection program at a Fortune 500 hospitality company where she was responsible for job analysis, employee survey initiatives, and the development and implementation of online pre-employment assessment and interviewing programs across the enterprise.
e-mail: Christie.Brodbeck@shakercg.com

Erica Drew is a Doctoral Candidate in Industrial-Organizational Psychology at Florida International University, where she received her Master's degree in 2011. Her research focuses on legal issues in personnel selection, applicant reactions to technologically advanced assessments, as well as selection procedures to enhance employee safety performance. Her applied experience extends to employee selection, performance appraisal, and leadership and employee development. Erica's work has been presented at national conferences and recognized for competitive student research awards.
e-mail: edrew001@fiu.edu

Jared Ferrell is a Consultant with Shaker Consulting Group, Inc. and is currently a Ph.D. candidate in the Industrial-Organizational Psychology program at the University of Akron. He earned his M.A. in Industrial-Organizational Psychology from the University of Akron in 2011. Jared has presented at numerous professional conferences and his work is published in multiple peer-reviewed journals. His research interests include organizational attraction, applicant reactions to assessments, training in virtual settings, social networking websites, and selection and training issues facing older job seekers.
e-mail: Jared.Ferrell@shakercg.com

Christopher Frost is a Consultant with Shaker Consulting Group, Inc. and is currently a Ph.D. candidate in the Industrial-Organizational Psychology program at Central Michigan University. He earned his M.A. in Industrial-Organizational Psychology from Central Michigan University. His research interests include personality assessment, automatic scoring of qualitative data, and competition in the workplace.
e-mail: Chris.Frost@shakercg.com

Brett Guidry is a Doctoral Candidate at Purdue University and currently holds an M.S. degree in Industrial-Organizational Psychology from Purdue. Brett has done research in the areas of self-regulation, assessment centers, and judgment and decision making, and he has presented research findings at several conferences. He has served in a consulting role for projects related to assessment centers, hiring and selection systems, and licensure exams. The projects Brett has worked on include clients such as the United Nations, the Emirates Group, the Kansas City Police Department, and the City of New York. He has also recently been a recipient of the Douglas W. Bray and Ann Howard Research Grant, as well as a Purdue Research Fund grant in order to further his work in judgment and decision making within assessment center simulations.
e-mail: bguidry@psych.purdue.edu

Sara Gutierrez is a Senior Research Scientist in the Product Development and Innovation team at CEB. In this role, she leads the design and development of CEB's multimedia simulations, computer adaptive ability tests, and assessments for mobile delivery. Her primary areas of expertise relate to application of item response theory (IRT), validation methods and procedures, and test security. Sara holds a Ph.D. in Assessment and Measurement from James Madison University and an M.A. in Industrial-Organizational Psychology from George Mason University. Dr. Gutierrez has presented at national conferences on topics such as efficiency and applicability of innovative item types, computer adaptive testing in personnel selection, unproctored internet testing, and the use of mobile devices to deliver pre-employment tests. She is a member of the Society for Industrial and Organizational Psychology, the International Association for Computer Adaptive Testing, and the American Psychological Association.
e-mail: Sara.Gutierrez@shl.com

Charles Handler is the President and Founder of Rocket-Hire and is a thought leader in the area of technology based assessment systems. Throughout his career he has specialized in developing effective, legally defensible employee selection systems that leverage technological innovation. He has taken what he learned while developing recruitment and selection solutions for a wide variety of organizations and combined it with his love of technology to help clients develop new models for employee selection. His philosophy focuses on combining sound science with innovation and practicality to create online hiring strategies that provide ROI and demonstrate the value of human capital. He works with a number of start-up companies as an advisor on matters related to measurement and technology. Charles has a Master's degree and a Ph.D. in Industrial-Organizational Psychology.
e-mail: chandler@rocket-hire.com

Ben Hawkes is the Global Director of Assessment Research and Development at Kenexa, an IBM Company. Graduating from the University of Surrey and the University of North Texas, he has worked in the field of Organizational Psychology in the US and Europe for over 15 years. He specializes in the use of assessment for selection and development, ranging from unproctored internet testing to interviewing and assessment centers. His research interests include the deployment of assessments to mobile devices, the use of video game techniques within assessment, and test-takers' perceptions of animated characters, or 'avatars', in simulation-based assessments. He has presented at the Society for Industrial and Organizational Psychology, the Association of Test Publishers and the UK's Association for Business Psychologists. He also delivers courses for the UK's Chartered Institute of Personnel and Development in the area of human resource strategy.
e-mail: ben.hawkes@uk.ibm.com

Brent Holland is Vice President, Research and Consulting at FurstPerson, Inc., and an authority on various aspects of employee selection, including validation, personality assessment, cognitive ability testing, and the design and use of multimedia simulations. He completed his Ph.D. in Industrial-Organizational Psychology at the University of Tulsa. Brent's current work involves designing innovative assessments and custom pre-hire solutions that enable companies to improve the performance and retention of front-line, customer-facing employees. His research interests focus on the validity and structure of multimedia simulations, alternative validation, causes of employee attrition in customer-facing occupations, and strategies to improve the accuracy and reliability of structured interviews.
e-mail: brent.holland@furstperson.com

Theodore Kinney is a Consulting Manager at Select International. He has extensive experience researching, developing, implementing, and validating unique selection solutions in organizations. He has experience managing selection efforts across industries and positions. Dr. Kinney also has vast experience training organization members at all levels on topics, such as behavioral interviewing, turnover reduction, and effective selection strategy. In his role, he manages a team of consultants in the implementation of selection projects and building strong client relationships. He also serves as Select's Executive Assessment Practice Lead, where he is responsible for all aspects of the Executive Assessment business. Dr. Kinney regularly presents his research in various academic journals and conferences. His primary research focus is in predicting performance in applied multitasking environments. Dr. Kinney completed his Ph.D. in Industrial-Organizational Psychology at Pennsylvania State University, along with his B.A. in Psychology. He received his M.S. in Industrial-Organizational Psychology from Colorado State University.
e-mail: tkinney@selectintl.com

Dawn Lambert is Director, Selection and Assessment at FurstPerson, Inc. and holds a Ph.D. in Industrial-Organizational Psychology from the University of Tulsa. Dawn is responsible for client research and consulting at FurstPerson, and works with numerous Fortune 500 organizations to help them find, hire, and keep the right

employees through the use of customized pre-hire solutions. She has helped to design, develop, and validate a variety of high-fidelity simulations, and has conducted more than 200 validation studies involving the use of simulations for selecting qualified candidates for customer-facing jobs. Her research interests include best practices and future directions in the use of alternative validation strategies, factors underlying early-stage attrition, drivers of employee success in remote work environments, and ongoing opportunities for improvement in the use of multimedia simulations for customer facing roles.
e-mail: dawn.lambert@furstperson.com

Martin Lanik is the President and Founder of Global Assessor and Affiliate Faculty in the Master of Organizational Leadership program at Regis University. He holds a Ph.D. in Industrial-Organizational Psychology from Colorado State University and consulted for organizations in multiple industries, including healthcare, interpreting, mining, financial services, food services, transportation, education, information technology, and government. His almost ten-year expertise in assessment centers and simulations spans across Africa, Asia, Europe, and North and Central America. Recipient of the Bray/Howard Award from SIOP, he is a frequent speaker at international conferences with over 40 invited and peer-reviewed talks and papers. Most recently, he was invited as the keynote speaker at The Assessment Center Study Group of South Africa and as a panelist at SIOP, wrote a feature article for the Chief Learning Officer magazine, and spoke for the European Personnel Selection Office.
e-mail: martin.lanik@globalassessorpool.com

Joe Latorre is the Director of Product Design and Consulting at Employment Technologies Corporation (ETC). He is responsible for designing and directing the development of all ETC brand, and custom simulation offerings. He also directs the development of all custom projects, products, and services offered by ETC. He has participated in the development of multiple video-based and computer-based simulations for a variety of domestic and international clients in Canada, Japan, Europe, India, and the Philippines. He holds the distinction of developing several nationally recognized award winning simulations. These include five HR Executive Magazine, "Product of The Year" awards within the last 8 years. He holds an M.S. degree in Industrial-Organizational Psychology from Valdosta State University. He also has over 20 years of national and international consulting experience and has presented at several national conferences related to simulation based technology for employee selection.
e-mail: Joe.l@etceasy.com

Amie Lawrence is the Manager of Product Development at Select International. She is an expert in the design, development, and validation of psychological assessment tools. She has extensive experience leading the development of many different kinds of psychological assessments including online competency-based, in-baskets, job simulations and motivational fit. She also spent several years as a Senior Consultant leading the implementation of assessment systems at all organizational levels for companies around the world. Her project management experience spans a variety of industries, including automotive, retail, manufacturing, distribution, utility,

telecommunications, international aid, and finance. She is a member of the Society for Industrial and Organizational Psychology (SIOP) and the American Psychological Association (APA). Dr. Lawrence maintains an active role in conducting applied research and is a frequent presenter at academic conferences. She is an IPMAAC paper award winner. She received her M.A. and Ph.D. in Industrial-Organizational Psychology from the University of Akron and a B.S. in Psychology from Virginia Tech.
e-mail:alawrence@selectintl.com

Jolene Meyer is a Research Manager in the Product Development and Innovation group at CEB. In this role, she leads a team of Industrial-Organizational psychologists who are responsible for the development of CEB's line of computer-adaptive cognitive ability tests, new product research and development, and test security initiatives. In her tenure at CEB, Jolene has managed complex assessment product lines, directed cross-functional teams, developed a variety of assessments, including cognitive ability, personality, and simulations, and led consortia research across a variety of clients. Jolene holds an M.A. in Industrial-Organizational Psychology from Minnesota State—Mankato. She has published in industry journals and presented at conferences on topics including range restriction in cognitive ability assessments, testing using mobile devices, and front line manager simulations. She is also a member of the American Psychological Association and Society for Industrial and Organizational Psychology.
e-mail: Jolene.Meyer@shl.com

Matthew O'Connell is co-Founder and Executive Vice President of Select International. Over the past 20 years he has been a leader in designing, evaluating, and integrating selection tools into systems that meet the needs of Global 2000 organizations. He is the coauthor of the bestselling book, Hiring Great People and is author or coauthor of more than 100 articles or book chapters on selection and assessment. Dr. O'Connell is actively involved in applied research, a frequent presenter at professional conferences, and an Adjunct Professor of Psychology at San Diego State University. He is a fellow of the Society for Industrial and Organizational Psychology (SIOP) and a member of the American Psychological Association (APA). He is a Best Paper award winner for the Academy of Management and SIOP. He received his M.A. and Ph.D. in Industrial-Organizational Psychology from the University of Akron and his B.A. in Psychology from Earlham College. He has lived and worked in both Mexico and Japan and is fluent in Japanese and Spanish.
e-mail: moconnell@selectintl.com

Deborah Rupp is the William C. Byham Chair in Industrial-Organizational Psychology at Purdue University. She received her Ph.D. from Colorado State University, and was previously an Associate Professor of Psychology, Labor/Employment Relations, and Law at the University of Illinois. Her assessment-related research focuses on validation, developmental assessment centers, and the use of technology to enhance assessment and development. She is a two-time recipient of the Douglas Bray/Ann Howard Award for research on leadership assessment and development.

She has co-chaired the International Congress on Assessment Center Methods, and co-led the taskforce that published a revision to the Guidelines and Ethical Considerations for Assessment Center Operations. Her research has been cited in US Supreme Court proceedings, and she worked with myriad organizations around the world (e.g., UNICEF, the Emirates Group, the South Korean government, SK, Doosan, Beijing University, IACBE). Rupp is a Fellow within the Society for Industrial and Organizational Psychology, the Association for Psychological Science, and the American Psychological Association. She is currently serving as a SIOP representative to the United Nations and Editor-in-Chief of Journal of Management.
e-mail: ruppd@purdue.edu

Eric Sydell is a Founding Partner and Director of Research and Innovation at Shaker Consulting Group, Inc. He earned his M.A. and Ph.D. at the University of Akron. His work interests include finding new ways to leverage technology to enhance the pre-employment assessment and selection process, and using large data sets to extract ever deeper insights in to what makes job candidates successful in new roles. Eric also consults with various Fortune 500 clients, conducts research on candidate response distortion and reactions, and speaks on various selection-related topics.
e-mail: Eric.Sydell@shakercg.com

About the Editors

Michael Fetzer is the Global Director of Advanced Assessment Technologies at CEB, the leading global provider of pre-employment assessments and talent measurement solutions. In this role, he is responsible for leading teams of I-O psychology and simulation/multimedia experts that design and develop leading-edge simulations and multimedia-based assessments that are utilized across many industries around the world. In addition, Dr. Fetzer is responsible for developing and implementing new and innovative methods for multimedia-based talent measurement, including serious games and other cutting-edge technologies.

Dr. Fetzer holds a doctoral degree in I-O psychology and is a contributing member of the Society for Industrial and Organizational Psychology (SIOP), the Association of Test Publishers (ATP), the International Test Commission (ITC), the International Association of Applied Psychology (IAAP), and the American Psychological Association (APA).

Kathy Tuzinski is a Principal Consultant at CEB. She works with a broad spectrum of clients to provide guidance on the application of CEB products to address their talent-related needs. In addition to her work with clients, she is also a product development expert with over 14 years of experience creating innovative and engaging products for talent measurement, incorporating the latest animation and video technologies. She specializes in translating customer requirements into sellable products, managing cross-functional teams from IT and R&D, and adapting existing methods for test development to create entirely new ways of testing people's skills and capabilities. Her simulations have earned 'Product of the Year' awards from Human Resources Executive and Customer Interaction Systems magazines.

Prior to her current role at CEB, she was a research scientist for Personnel Decisions, International (now PDI Ninth House, a Korn Ferry Company) and Inscape Publishing, where she provided research consulting to Fortune 100 clients and developed the widely used DiSC-based INDRA tool for employee development. Kathy received her M.A. in Industrial/Organizational Psychology from the University of Minnesota and a B.A. in Psychology from St. Catherine University. She is a frequent presenter at industry conferences (SIOP, ATP), and has published several articles in industry journals.

Chapter 1
Simulations for Personnel Selection: An Introduction

Kathy Tuzinski

Multimedia simulations stand alone in their ability to elicit, capture, and measure the behaviors that are most similar to actual job performance. This book surveys the current landscape of multimedia simulations for personnel selection, with accessible chapters written by those who have shaped the landscape through their pioneering efforts to merge new technologies into the practice of industrial and organizational (I-O) psychology. It will be of great interest to students, researchers, and practitioners who are looking for guidance in developing and implementing multimedia simulations for employee selection. It is also a valuable source of information about the wide range of simulations in use today, and is designed to provide inspiration, ideas, and lessons learned to novice and expert simulation developers alike.

One can find books that address various concepts related to simulations, such as computer automated scoring and the use of technology in assessment, but this is the first book to focus entirely on multimedia simulations. Other books about simulations are geared towards the implementation, development and scoring of credentialing and licensing simulations (exams) that primarily measure field-specific knowledge and highly technical skills, but not necessarily the constructs that are more common in the workplace, such as personality, judgment, and other job-related 'soft' skills. This book treats the world of multimedia simulations for personnel selection as a discipline in its own right, incorporating wide-ranging issues such as implementation, scoring, development, and validation. With helpful chapters containing the best technology for developing simulations, step-by-step instructions and lessons learned, as well as latest research on candidate reactions and group differences, this book is likely to be the best resource available.

Today's multimedia simulations take a variety of forms. Easy to recognize, but difficult to define exactly, these multimedia simulations are numerous and challenging to categorize. The one common thread that runs through all simulations is their design for one purpose: to capture work-relevant performance, either while performing a task, interacting with another person, or working with systems. They also share an ability to keep candidates more engaged relative to other types of assessments.

K. Tuzinski (✉)
CEB, Minneapolis, MN, USA
e-mail: kathy.tuzinski@shl.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_1, © Springer Science+Business Media New York 2013

From that common thread, however, simulations diverge. Organizing frameworks are lacking. An organizing framework would help not only to bring greater definition to the term "simulation" but also, may serve as a basis for prescriptive simulation design, laying the foundation of standards for simulation development and processes for validity, reliability, and scoring. It may also help the reader understand how the different examples in the book fit within the larger simulation space. Before offering a view of what an organizing framework might look like, it is helpful to make a brief visit to the roots of today's modern simulations.

1.1 Multimedia Simulations in Context

> Hopefully, these first few attempts are the beginning of a whole new technology of behavior sampling and measurement, in both real and simulated situations. If this technology can be realized and the consistencies of various relevant behavior dimensions mapped out, the selection literature can cease being apologetic and the prediction of performance will have begun to be understood – Wernimont and Campbell (1968, p. 376).

Written over 40 years ago, the above quote is prescient. Simulations have come a long way since their formal introduction into the military during the 1940s, and a decade later, when they figured prominently in the managerial selection and development programs at AT&T.[1] As technological innovations advanced, simulation developers found ways to capitalize on the opportunities. Today's simulations incorporate all forms of multimedia, including audio, video, and 3D animation, and automation in the delivery and scoring. Simulations are now used everywhere and for a wide variety of positions.

Simulations are best characterized as measurement methods, rather than a type of test or construct. In its various forms, simulations measure hard and soft skills, personality, task performance, job knowledge, and cognitive ability. Furthermore, simulations are used in a variety of applications, including certification testing (e.g., ophthalmic technicians), licensure, training, and personnel selection. Simulations are rooted in three categories of tools, which by themselves show similarities and overlap: (1) assessment centers, (2) work samples, and (3) situational judgment tests. At the risk of oversimplifying, the multimedia simulations contained within this book reflect these tools with technology added for the purposes of improving and automating assessment delivery, data capture, and scoring.

1.1.1 Assessment Centers

The assessment center method grew out of the research labs of the modern twentieth century psychological measurement movement and the field of military selection.

[1] Some good resources on the history of testing: reaching back to China 3000 years ago (Oakland 2006); the progression of testing from the birth of psychological science in the late nineteenth century up through modern testing (Waldron and Joines 1994); assessment centers (Lievens and Schollart 2009); situational judgment tests (Whetzel and McDaniel 2009); and early multimedia assessments (McHenry and Schmitt 1994).

Psychologists developed precursors to the modern assessment center method in their research (e.g., Henry Murray's Harvard Psychological Clinic Study in the 1930s) and applied testing in military contexts (German Officer Selection in 1920s to 1942; British War Office Selection Boards, and the United States Office of Strategic Services (US OSS) Program in World War II). Eventually these methods made their way into managerial selection and development in the business world. These methods were first applied to the civilian sphere in the British Civil Service Assessment in 1945 and AT&T Management Progress Study which started a decade later. The typical assessment center incorporates some or all of the following components: multiple measures, observations, and assessors; behavior capture; trained assessors; and the integration of behaviors by pooling information (ratings or impressions) from raters.

Assessment center exercises are simulations of major aspects of performance for a given role. Assessment centers have traditionally been used for assessing managers, but have also been used for sales people, and roles for public safety, for both selection and development. Exercises such as the in-basket, leaderless group discussion, and the role play are all meant to test out a candidate in a real life situation to look for samples of performance. Although they can be used for measuring many different types of job related skills and abilities, they tend to measure competencies—a mixture of knowledge, skills, abilities and other characteristics (KSAOs)— such as interpersonal skills, communication skills, planning and organizing, and analytical skills (SIOP 2013).

More than 20 years ago, Waldron and Joines (1994) made predictions about things to come in assessment centers, such as multiple-choice in-baskets, the increased use of low fidelity simulations, and remote assessment. Furthermore, they predicted an increase in automation, including simulated email systems, data integration, and exercise scoring and reporting. Today's managerial assessment centers in fact do incorporate the latest in technology that is reflected in today's jobs, providing a more realistic twenty-first century "day in the life" experience relative to traditional brick and mortar assessment centers. Text messages and emails have been incorporated into simulations, allowing for information to arrive non-sequentially, just like the way it is in the work day of the typical manager of today (McNelly et al. 2011). Demand for an alternative to the in-person assessment center has also increased in response to budgetary constraints and the increasing affordability of remote assessment and technology-enabled assessment. For those interested in getting a sneak peek into the future of assessment centers, the chapter by Guidry et al. on novel techniques for tracing decision-making process in managers will provide ample food for thought.

1.1.2 Situational Judgment Tests

Situational judgment tests (SJTs) appeared on the scene at about the same time as assessment centers. In the 1920s, a widely used SJT with response options was likely a subtest from the George Washington Social Intelligence Test (Whetzel and McDaniel 2009). Army psychologists used the SJT format for measuring the judgment of soldiers in World War II, and a number of SJTs were developed in the 1940s for assessing supervisory potential. By the 1950s and 1960s, organizations were using SJTs for managerial selection (Whetzel and McDaniel 2009).

SJTs present situations to candidates that might be encountered on the job and ask candidates to respond in one of two ways, what they would do or what they should do, given the situation. Situational judgment tests have been used for measuring a number of different constructs, such as interpersonal (Lievens and Coetsier 2002) and leadership judgment (Bergman et al. 2006), and conflict resolution skills (Olson-Buchanan et al. 1998).

SJTS are preferred over higher fidelity simulations, such as assessment center role plays, because of their ease in administration and scoring. SJTs have grown in popularity for a number of reasons. First, their high face validity provides more favorable candidate reactions, relative to standard personality or knowledge-based multiple-choice or Likert-scaled formats. Second, smaller subgroup differences (i.e., minority-white, or female-male) have been found for some SJTs relative to traditional cognitive ability tests (Clevenger et al. 2001), thus providing an opportunity for greater validity while minimizing the risk of adverse impact. Third, large-scale studies indicate that SJTs have substantial criterion-related validities (McDaniel et al. 2001).

The contact center simulation featured in the Holland and Lambert chapter in this volume uses an SJT item format to gather candidate responses on the most effective response to say, given the situation, where the "situation" can be any combination of what the caller just said, the information provided in the simulated agent software, and other pertinent information that was provided to the candidate during the assessment. The coaching simulation described by Gutierrez and Meyer in this volume is an example of an SJT that uses a video-based format. In moving from a low-tech, in-person assessment center exercise (role play) to a high-tech remote situational judgment test (coaching simulation), the result is a lower-fidelity assessment that is lower in cost for organizations, has higher availability, standardization, and ease of use.

1.1.3 Work Samples

Definitions for work samples have varied in the literature and there appears to be some disagreement over what constitutes a work sample. Simply stated, this disagreement is over whether work samples include a wide range of measures at the high and low ends of the (primarily) physical fidelity spectrum, or whether this label is reserved only for those measures at the high end of the spectrum. The broader definition would include low-fidelity measures such as "talk through" interviews and situational judgment tests in addition to high-fidelity measures, such as cockpit simulators, performance tests, and assessment center exercises.

The most literal definition of a work sample would be a hands-on performance test in which a job applicant is required to actually perform a job-related task under the same conditions as those required on the job. Measures that are classed under the heading of work samples can be organized according to the degree by which they are removed from the two features of actual hands-on performance and a real work setting (Callinan and Robertson 2000).

In the broader definition, there is room for work samples to include both situational judgment tests as well as assessment center exercises if level of fidelity is specifically called out in the definition of a work sample (as in “the contact center simulation is a low-fidelity work sample”). The narrower definition of work samples, on the other hand, would only include measures in which the applicant performs a selected set of actual tasks that are physically and/or psychologically similar to those performed on the job (Roth et al. 2005). This difference in opinion has created some confusion in the literature in terms of when a test is a work sample vs. something else, and how to interpret previous research findings, particularly surrounding the validity of work sample tests.

1.2 The Fidelity Continuum: An Organizing Framework

Simulations vary in their ability to replicate the physical and psychological fidelity of a work task. Physical fidelity is the extent to which a test itself involves the actual tasks performed on the job (Truxillo et al. 2004), whereas psychological fidelity represents the extent to which the relevant knowledge, skills and abilities (KSAs) are called forward in the process of completing the task (Goldstein et al. 1993). In both cases, the degree of fidelity is a continuum rather than a dichotomy.

Physical and psychological fidelity are related concepts: although some psychological fidelity comes along whenever there is physical fidelity, it is possible to have psychological fidelity without physical fidelity (think, for example, of a paper-and-pencil SJT). However, Goldstein et al. (1993) held that physical fidelity is less important to content validity than psychological fidelity, particularly if the job requirements do not involve physical-based tasks, such as operating machinery, fixing equipment, manipulating physical objects, and so on.

Fidelity maximizes the point-to-point correspondence between the simulation and the task it is meant to represent (Asher and Sciarrino 1974) and should therefore have a direct impact on the validity of the measure. It increases face validity, which can improve candidate’s perceptions of the assessment as well as provide the benefit of a realistic job preview. Poor face validity has been suggested to reduce the candidate’s desire to perform well, possibly leading to biased test scores (Arvey et al. 1990). Fidelity also aids content validity by matching the KSAs brought out by the simulation to the requirements of the role, which is important if this is the main strategy for justifying the use of the assessment (i.e., a selection procedure can be validated by a content-oriented strategy if it is representative of the important aspects of performance on the job, according to the Uniform Guidelines).

According to Lievens and De Soete (2012), the logic of maximizing point-to-point correspondence between the predictor and criterion is conceptualized differently for high-fidelity versus low-fidelity simulations. In high-fidelity simulations (assessment centers and work samples), assessors observe and rate actual on-going candidate behavior, which shows true point-to-point correspondence with the criterion. On the other hand, low fidelity simulations (such as an SJT) sample applicants' procedural knowledge about effective and ineffective courses of action in job-related situations. There is not the same level of point-to-point correspondence in low-fidelity simulations because the behavior of choosing among alternatives is not the same as constructing and actually demonstrating the behavior one wishes to make in response to complex interactions with other humans (Thornton and Rupp 2006). However, it is important to note that moving to low-fidelity forms does not necessarily in all cases harm the criterion-related validity (Lievens and Patterson 2011).

Many researchers have suggested that stimulus and response fidelity should be considered separately, particularly when interpreting research findings (Lievens and De Soete 2012; Truxillo et al. 2004). For example, Funke and Schuler (1998) found that moving towards higher fidelity on the stimulus side (from orally-presented questions to video) had little effect on the validity but moving to higher fidelity on the response side (multiple-choice vs. written vs. orally-given replies) did affect validity. Response fidelity appeared to put a ceiling on the gains in validity that could be achieved by increasing the fidelity of the stimulus. Given the proliferation of simulation types, with varying levels of fidelity on the stimulus and the response sides, treating stimulus and response fidelity separately is good advice. What started out as a suggestion, is now likely critical when interpreting research findings. It is expected that more research in this area will continue, with findings such as those recently found by Lievens et al. (2012), demonstrating that response fidelity may not only affect validity but also have modest effects on reducing the saturation of cognitive ability, increasing the saturation of certain personality traits, as well as improving candidate perceptions.

To provide an organizing framework around work samples, performance tests and competency testing, Truxillo et al. (2004) grouped these assessments into three main categories based on level of physical fidelity. The first group of tests included those that were physically just like the job. The second group of tests represented those that closely sampled the tasks performed on the job, such as the physical ability tests often used for selecting into public safely positions. The third group of tests were those that closely resembled the job (in that they present applicants with a work-related situation through video), but rather than applicants showing what they would do, applicants described what they would do in a given situation. This framework has been adopted here for grouping simulations into different fidelity categories, treating the stimulus and responses separately. Tables 1.1 and 1.2 represent a framework for simulations, based on where the stimuli (Table 1.1) and responses (Table 1.2) fall on a continuum of fidelity. For the purposes of simplicity, physical fidelity and psychological fidelity are not shown, but it may be assumed that psychological fidelity exists across all forms of stimuli, increasing from left to right, and physical fidelity is primarily present in the last column only.

This framework is useful as a starting point for organizing research findings on validity, applicant reactions, and group differences, and may be useful for a few different applications: to categorize and understand the differences among the possible simulations described in this book and elsewhere; to inform R&D spending by

Table 1.1 Continuum of stimulus fidelity

	Lower			Higher
Presentation	Written or oral delivery	Video, audio or animation (linear)	Video, audio or animation (branching)	Live action
Description	Computer, paper-and-pencil, or orally delivered situation	Stimuli (e.g., action, dialogue, or computer system screens) that do not adapt in response to the candidate's actions	Stimuli that change in response to the candidate's actions	Stimuli that maximize point-to-point correspondence between the task and the simulation and change in response to the candidate's actions
Example assessment types	Situation-based interviews, written SJTs	Call center simulations, interpersonal simulations, static role plays and in-basket exercises, static software simulations	Serious games, dynamic (computer adaptive) simulations	Work samples, flight simulators, role plays, in-baskets, job tryouts

Table 1.2 Continuum of response fidelity

	Lower			Higher
Description	Closed-ended (linear)	Closed-ended (branching)	Open-ended declarative	Open-ended behavioral
Response types	Multiple choice questions	Multiple choice questions	Written or oral response	Demonstrated behaviors

highlighting areas of greatest return on investment; and to provide a means for the practitioner to evaluate different simulation options when making decisions about which to use. For example, if research demonstrates that increasing response fidelity greatly improves candidate reactions and validity and decreases group differences, this would provide the practitioner with the evidence needed to build the case within the organization to invest in response-gathering technology, such as webcams, that can be provided to candidates during the assessment process (Oostrom et al. 2011). The technologies described by Guidry et al. in Chap. 11 for collection and analysis of free-form behavioral responses would likely get increased attention as a result of this research, especially if it can be coupled with the new scoring technologies described by Sydell et al. in Chap. 5.

The investment in technologies to increase fidelity is more easily justified for jobs where mistakes could be costly, dangerous, or fatal (such as pilots, air traffic controllers, and surgeons). However, many organizations in the civilian sector do not hire for such mission-critical roles. Should the same level of stringency be applied in the case of a front-line manager or call center agent? For these roles, costs of implementing the simulation are considered along with the cost of making a bad hire. In Chap. 2, Boyce et al. address fidelity in greater depth and discuss the benefits (and drawbacks) of fidelity depending on the intended purpose.

Level of fidelity is just one factor that differentiates simulations. Simulations vary in the constructs measured (or tasks represented), comprehensiveness (degree to which the entire job performance domain is represented by the tasks that make up the measure), job role, purpose (i.e., credentialing, training, or selection), and difficulty. As research findings on different simulations accumulate, there will be greater opportunities to assemble these findings in a framework that highlights the effects that various simulation facets have on important outcomes such as applicant perceptions, group differences, and validity.

1.3 Preview of Chapters in the Book

Creating a new simulation from scratch is as much an artistic as a scientific endeavor. An apt metaphor is that of trying to write a symphony or putting together the pieces of a 3D puzzle. There are many layers to consider: content, look-and-feel, flow, scoring, and multimedia elements. The final product needs to be psychometrically sound as well as provide an experience to the candidate that feels authentic and coherent and also has the ability to draw out the relevant performance behaviors.

It can be challenging for even the most talented divergent thinker to get his or her mind around the process. Many of the authors in this book are pioneers in this new area of test development. They have had to rely on their own ingenuity and skills in measurement, psychometrics, and storytelling without the benefit of how-to manuals or best practices for simulation development. This is why this book is so needed at this point in time.

1.3.1 Section I: Simulations in the Selection Context—Broader View

Section I introduces some important topics when considering the use of simulations for employee selection. The chapters in this section address a wide variety of issues, including challenges and opportunities facing simulations in the selection context, current and emerging tools and technologies available to simulation developers, methods for scoring simulations, and current and future research directions for candidate reactions to simulations.

Kicking off Section I, Boyce et al. in Chap. 2 discuss the considerations, challenges, and opportunities of simulations in the employee selection context. They cover the organizational and individual issues relevant to the use of simulations during the attraction and recruitment stages and beyond. The authors explore the existing research on each issue and provide practical guidance, highlight areas in need of additional research, discuss the ideal conditions for simulation use, and share strategies for effective implementation within organizational selection contexts.

Bruk-Lee et al. in Chap. 3 review the research to date on candidate reactions to simulations and media rich assessments. They begin broadly, with a historical overview of candidate reactions to assessment and then focus specifically on candidate reactions to technologically advanced assessments through the lens of procedural and distributive justice. Furthermore, they propose that such research should distinguish between administration mediums and media types, as well as consider the impact of individual differences. Finally, the concept of the uncanny valley (candidates' perceiving animations as 'eerie' or 'unhumanlike') is explored. They highlight the paucity of research in candidate reactions to media rich assessments in general and call for an increased emphasis on this line of research.

Hawkes in Chap. 4 provides an overview and evaluation of the tools available to assessment simulation creators. Written for the non-technical reader, this chapter provides a how-to guide for selecting the most appropriate tools for simulation projects. It addresses each stage of the simulation production process, covering technologies for content creation, authoring, and deployment. Hawkes provides additional insight into the important things to consider when deciding on which technology to choose, including ease-of-use, flexibility, limitations, accessibility, and cost. Real-world examples of these technologies are sprinkled throughout the chapter, and a comprehensive list of links to additional resources is provided.

Closing out Section I, Sydell et al. in Chap. 5 review the scoring of simulations. They believe that harnessing the power of larger data sets and advancements in the ability to combine items and item types into scoring algorithms will bring a shift in the predictive capabilities of simulations. They also discuss the use of automatic scoring and branching logic, as well as new technologies for automating the scoring of qualitative (i.e., open-ended) assessment responses. They assert that a much deeper understanding of who a person is and how they will behave can be developed by examining interactions between simulation components and other sources of information.

1.3.2 Section II: Simulations in Action

Section II provides real world examples of simulations in action. The chapters in this section provide examples and discuss simulations for service roles, manufacturing, contact centers, and managerial selection. Simulations for assessing computer proficiency and leadership and decision-making are also discussed.

Barr and Coughlin in Chap. 6 describe simulations that measure computer skills, from more generic computer proficiency to the specific (e.g., Microsoft Office). These simulations have the potential to be more ubiquitous as they are relevant across a wide variety of job roles/levels and industries. In fact, they are relevant for any job where computer skills are required. Unlike other simulations that only need to resemble the psychological fidelity of workplace systems (call center software comes to mind), software simulations need to represent the interface (physical fidelity) as close to the real thing as possible. The authors provide a helpful discussion surrounding the many decisions that have to be made when designing and scoring computer/software simulations.

Holland and Lambert in Chap. 7 focus on the use of multimedia simulations in contact centers. Contact centers are challenging work environments that require the employees to provide a high quality customer service or sales experience to customers while interacting with multiple computer programs, under significant time pressure, and often under the watchful eye of performance management systems that track their communication style, reliability, and performance. Contact center simulations not only provide a realistic job preview but create more a engaging candidate experience. They have evidence showing that contact center simulations may be the single best predictor of many job-specific metrics.

O'Connell et al. in Chap. 8 describe the use of interactive simulations in manufacturing settings. They mention several examples of interactive simulations that are currently in use by manufacturing organizations. Manufacturing is an interesting case in which 'simulation' can just as easily mean a multimedia-based assessment measuring targeted competencies, as a complex multi-workstation setting involving the actual equipment and processes encountered on the job. As manufacturing roles change to include more decision-making, multitasking, and collaboration, the use of multimedia simulations in these jobs is only expected to increase.

LaTorre and Bucklan in Chap. 9 review simulations for service roles. Service roles include positions in retail sales, customer service, and banking, where interpersonal effectiveness is often a top job requirement. Focusing on the application of the assessment center methodology to high volume, non-managerial positions, they describe elements of a best-in-class assessment program, along with lessons learned and important pitfalls to avoid. The chapter also describes research approaches to establish the return on investment (ROI) of simulations, and techniques for communicating this value to the hiring organization.

Gutierrez and Meyer in Chap. 10 describe the use of multimedia simulations for selecting managers and front-line supervisors. Simulations for these positions have gained in popularity in recent years as organizations have learned that they can be a cost-effective alternative to in-person assessments. Two different simulations were developed simultaneously, a coaching skills role play and an inbox assessment. Careful attention was paid to representing the types of scenarios that managers encounter on a daily basis, such as coaching direct reports, prioritizing one's work and that of others, monitoring employees, and making decisions under pressure with limited information. They have found these simulations to be well-received by candidates and human resources (HR) recruiters alike.

Guidry et al. in Chap. 11 focus on simulations that reflect the realities of the postrecession economy, the unanticipated and complex situations that have become the exception rather than the norm for business leaders today. Their focus, as they say, is to 'stretch the boundaries of virtual simulations' functionality'. They propose novel ways to leverage technology to measure previously unobservable decision strategies to bring out into the open normally concealed thought processes of candidates. They predict that detecting and measuring what has been up to this point difficult to assess using more traditional technologies will inform both the science and the practice of simulations for assessing and developing leaders.

Fetzer in Chap. 12 takes a peek forward at the future of simulations for employee selection. He believes that much can be done to advance the technology to develop new simulations that will render simulations of tomorrow to look more like games from the entertainment industry than traditional tests of yesterday. "Serious games," developed primarily for the military and government, are picking up momentum in the civilian sector, and will provide greatly enhanced levels of user engagement, measurement opportunity (and complexity), test security, and positive business outcomes for the organizations who utilize them in their hiring processes. Fetzer predicts that the future of game-inspired assessments (GIAs) will take two distinct forms: those that are more like the casual games of today, and those that are as realistic as the then-current technology will allow. Either way, he has no doubt that these GIAs will become the standard for personnel selection in the not-too-distant future.

1.4 Concluding Thoughts

The chapters in this book reflect the latest thinking and research in multimedia simulations and should be of interest to experts as well as students and a general business audience, and represents an important step towards normalizing the presence of simulations in the context of personnel assessment. We suspect that in only a few years what is being represented in this book as novel will become more commonplace, and may even appear outdated. The field continues to move forward as simulation developers push the envelope by incorporating new technologies, adopt methods from other fields outside of I-O psychology (like credentialing, licensure, and education), work with a diverse set of talented people from marketing, technology, and multimedia production, and improve on scoring processes. It is personally a very gratifying time to work in this field.

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519–533.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14,* 223–235.

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8,* 248–260.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86,* 410–417.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6,* 115–123.

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis- content validity process. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (pp. 2–34). San Francisco: Jossey-Bass.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment, 10,* 245–257.

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *Oxford handbook of personnel assessment and selection* (pp. 383–410). New York: Oxford University Press.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96,* 927–940.

Lievens, F., & Schollart, E. (2009). Assessment centers: Overview of practice and research. Online readings in testing and assessment. http://www.intestcom.org/Publications/ORTA/The%20assessment%20center.php. Accessed 28 Feb 2013.

Lievens, F., De Corte, W., & Westerveld, L. (2012). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management.* doi:10.1177/0149206312463941.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 193–232). Hillsdale: Lawrence Erlbaum Associates, Inc.

McNelly, T., Ruggeberg, B. J., & Ray Hall, C. Jr. (2011). Web-based management simulations: Technology-enhanced assessment for executive-level selection and development. In N. T. Tippins & S. Adler (Eds.), *Technology-enhanced assessment of talent* (pp. 253–266). San Francisco: Jossey-Bass.

Oakland, T. (2006). Introduction: 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: Issues and advances* (pp. 1–11). Chichester: Wiley.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51,* 1–24.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology, 10,* 78–88.

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.

SIOP (2013). Types of employment tests: Assessment centers. https://www.siop.org/workplace/employment%20testing/testtypes.aspx. Accessed 7 Mar 2013.

Thornton, G., & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah: Lawrence Erlbaum Associates.

Truxillo, D. M., Donahue, L. S., & Kuang, D. (2004). Work samples, performance tests, and competency testing. In M. Hersen (Ed.), *Comprehensive handbook of psychological assessment* (pp. 345–370). Hoboken: Wiley.

Waldron, B., & Joines, R. (1994). Introduction to assessment centers: Assessment centers 101. Workshop conducted at the IPMAAC conference on public personnel assessment, Charleston, SC. June 26, 1994. http://www.ipacweb.org/files/ac101.pdf. Accessed 7 Mar 2013.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372–276.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19,* 188–202.

Part I
The Broader View

Chapter 2
Simulations in the Selection Context: Considerations, Challenges, and Opportunities

Anthony S. Boyce, Christine E. Corbet and Seymour Adler

The use of simulations for assessing and selecting employees has a long and distinguished history in both government and commercial organizations, in the United States and globally. Over the past few decades, however, simulations have been increasingly used as "free-standing" elements of selection processes, outside their traditional role embedded in full-blown assessment centers. Today, simulations are used in selection processes in varying forms in conjunction with other assessment tools such as structured interviews, cognitive ability tests, and personality inventories. As with each component of a selection process, the simulation designer must consider the incremental value that the simulation adds to the process, relative to the other elements. There are many potential advantages for employing simulations for selection, but potential drawbacks exist as well.

In this chapter, we begin with a discussion of the advantages and disadvantages of using simulations. Next, we highlight some of the key considerations involved in the design and implementation of simulations, providing practical guidance and recommendations for areas in need of future research. We then examine the role of simulations in the candidate life-cycle from the perspective of both organizations and candidates. Finally, we conclude with a brief exploration of how changes in technology are likely to impact the nature of simulations and how candidates interact with them in the future.

A. S. Boyce (✉) · S. Adler
Aon Hewitt, New York, NY, USA
e-mail: anthony.boyce@aonhewitt.com

C. E. Corbet
Right Management, New York, NY, USA
e-mail: christine.corbet@right.com

S. Adler
e-mail: seymour.adler@aonhewitt.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_2,

2.1 Advantages and Disadvantages of Simulation Use

2.1.1 Validity

Work samples, including simulations, have consistently demonstrated strong criterion-related validities (e.g., Schmidt and Hunter 1998). Indeed work samples typically are among the most valid tools for employee selection. Positive validation evidence has emerged across a wide range of settings and target jobs, from entry-level to management positions. Of course, the meta-analytic publications blend together in a single broad category labeled "work samples" or "simulations" a variety of specific procedures that differ a great deal in their design (as is true for meta-analyses of the validity of structured interviews, for example). Nonetheless, there are solid reasons for expecting simulations to be valid predictors of job performance.

To begin, as performance-based tests, simulations are less subject to the threats of faking and social desirability-based response distortion than what has been found to be true of personality inventories and interviews. In addition, during a simulation, as on the job, the participant *behaviorally* demonstrates within a representative situational context the underlying capacities—technical knowledge, prior learning, personality traits, and cognitive ability— that other selection tools only measure in a nonbehavioral format. In the classical formulation articulated by Wernimont and Campbell (1968) close to a half-century ago, simulations measure behavioral *samples*, not *signs*. Thus, if designed properly, the content of simulations mirrors the content of the target position. Indeed content-oriented strategies are often appropriately and successfully used to validate simulations. Furthermore, beyond actually being valid for selection decisions, simulations have another advantage—they *look* valid given their greater degree of fidelity and similarity to the target position.

2.1.2 Credibility

Over the past 20 years, research and practice in industrial-organizational (I-O) psychology has begun to recognize the importance of ensuring that elements of the assessment process appear credible and fair to candidates. Academic research, starting with Gilliland's (1993) landmark study, has largely drawn on justice theory to describe the factors that affect candidate perceptions of the fairness of various assessment procedures. Practitioners similarly are concerned with the impact of how candidates are treated during the pre-employment process on the organization's employment brand (Yu and Cable 2012). Indeed, in some businesses, especially mass-market retail, there is a growing realization that job candidates are also consumers. How people are treated as candidates can affect how prospective customers perceive the organization's brand as a provider of goods or services and whether those perceptions are shared by candidates with their respective social networks.

While simulation fidelity can vary to a great extent, simulations as a whole have higher fidelity than all other assessment techniques (we will address the question of how much fidelity is required in a simulation later in this chapter). Only during a simulation can job candidates be confronted with vividly representative real-life job challenges. They can be given the freedom to respond in ways—by speaking, writing, and making decisions—that reflect how they would actually respond in the target position. Research has demonstrated that perceived job relevance, ability to perform to the best of one's abilities, and the perceived value of potential feedback are all factors that affect the perceived credibility of assessment procedures and the perceived fairness of outcomes coming from those procedures. Not surprisingly, meta-analytic findings (e.g., Hausknecht et al. 2004) have found simulations to be consistently perceived as more favorable by candidates than all other selection techniques, with the exception of interviews.

2.1.3 Flexibility

In many ways, an organization looking to add a simulation to the selection process has a larger range of design and delivery choices than it would have with alternative assessment methods. For instance, traditional psychometric tools measuring cognitive ability, job knowledge, or personality can be administered in paper form or on a computer. Length typically ranges from 15–90 minutes, rarely shorter and rarely longer in operational selection situations. In contrast, as the chapters in this volume amply illustrate, simulations take on many forms. They can last 10 minutes or 10 hours and can have candidates interact with a single role-player or with a team of role-players. They can be technology-enabled or rely on print materials. They can require the candidate to behave naturally (e.g., creating and delivering a sales presentation to a live audience) or in highly structured ways (e.g., selecting a behavior from a list of alternative responses). Candidate behavior can be recorded or not. Correspondingly, if recorded, assessment of candidate performance can occur immediately following participation in the simulation or at a later point in time.

The many simulations described in this book illustrate the range of complexity that simulations today take on. Some are stand alone, relatively brief exercises (as short as 2–3 minutes), while others are day-long simulations with multiple components and multiple interactions. Even the more complex simulations vary in the timeframe they *represent*; some are day-in-the-life experiences and others may have a year's worth of events unfold over the course of a day. Moreover, simulation exercises can be delivered face-to-face, over the telephone, through personal computers, over the web with or without media-rich enhancements, and with or without live visual or auditory interaction with role-players. Simulations can be simultaneously administered to individual candidates or to groups of candidates, and if to groups, individuals in a group can be allowed to interact with each other or not. This range of sizes and shapes affords the designer a great deal of flexibility when inserting a simulation into a selection process.

Finally, simulations—certainly, if they are custom-designed for a particular application—can accommodate the assessment of a very wide range of constructs.

Simulations can be designed to assess leadership or customer service skills. They can tap decision-making under stress or the ability to adapt to change. Simulations can measure the depth of a candidate's technical knowledge and the ability to apply that knowledge in job-relevant situations. This flexibility means that in designing a selection process, the practitioner can identify those critical job requirements not well assessed by other elements of the process and devise a simulation to specifically focus on those gaps.

2.1.4 The Downside

Historically, the key negative to implementing simulations has been the relatively high cost of administering them. An associated negative for adopting simulations has been the need to bring candidates to a facility and have trained, skilled assessors evaluate the candidates' performance on the simulation. In some cases, the simulation further required a special apparatus, for example, for ergonomic measurement in simulations used for firefighters or cable technicians.

At this point in time, the development of a customized simulation is still relatively expensive and time-consuming, even, or especially, if technology is part of the design. It is not unusual for the design of a simulation, particularly if it is technology-enabled, to take 4–6 months from inception to deployment. Urgent selection challenges may require quicker answers. We should point out, though, that the cost of designing and validating a simulation may not be materially different from the cost of designing and validating more traditional measures of aptitude, personality, or job knowledge.

Offsetting the cost of investing in a technology-enabled solution is often the more efficient delivery of simulation content on an ongoing basis. Thus, for example, creating a system that analyzes and scores written verbal content (e.g., inbox e-mail responses) can virtually eliminate professional labor costs associated later with scoring. Costs need to be carefully analyzed in terms of the specialized skills and associated labor rates required at the design and delivery stages. In particular, the calculation of projected returns has to take into account the number of people who will be administered the simulation—the amortization base—and the financial and less tangible risks associated with a bad hire or a bad candidate experience.

2.2 Design and Implementation Considerations

Implementing job simulations can be a relatively simple endeavor if an existing off-the-shelf (OTS) solution is chosen, or it can be a complex process that spans many months and requires substantial input from relevant stakeholders and subject matter experts (SMEs). Where a particular job simulation falls along this continuum is influenced by many situational constraints, including budget for development and ongoing administration; people factors related to stakeholders and the candidate pool; the breadth and depth of the knowledge, skills, abilities and other characteristics (KSAOs) targeted; and the degree to which technology is leveraged to aid in delivery and scoring. Each of these considerations places constraints upon key features (e.g., fidelity, medium, and scoring method) of the job simulation itself as well as the development and implementation processes. In turn, these features impact the outcomes of the simulation (e.g., validity, candidate reactions, and generalizability). While the interrelationships between these factors are somewhat complex, we have broken our discussion into seven key areas so that interested readers can target their review to some or all of the components. These areas are: fidelity, KSAOs, selecting an OTS simulation tool, approach to scoring, proctored versus unproctored, use of technology, and managing the interests of simulation stakeholders.

2.2.1 Fidelity

One of the most salient features of a job simulation is *fidelity,* the degree to which the simulation authentically reflects the targeted job in terms of both stimuli and responses (Motowidlo et al. 1990). Although these two components of fidelity are related, each can be considered to vary along a continuum. Assessment centers and work samples generally reside at the high end of stimuli fidelity in that they include actual (or close approximations of) job materials, equipment, and people to enhance the realism of the assessment and more closely replicate features of the target job. Computer-based simulations often fall into the middle range of stimuli fidelity, replacing actual job materials and equipment with less realistic or more general mock-ups and substituting video, avatar, or audio representations for live people. The low end of stimuli fidelity is anchored by traditional situational judgment tests (SJTs) where the stimuli are limited to text-based descriptions of the task situation. Likewise, the degree to which a simulation affords candidates the opportunity to perform the range of behaviors that would be available to them on the job (i.e., response fidelity) varies greatly and impacts overall fidelity. For example, work samples and role-plays can allow candidates to actually perform the relevant behaviors as they would on the job, whereas computer- or paper-based simulations often limit candidate responses to a set of multiple-choice options.

Among the chief advantages of high fidelity simulations is that they generally exhibit high face validity. Due to the obvious overlap between the simulation and the job, candidates and stakeholders (e.g., recruiters and hiring manages) can easily see how performance on the simulation relates to expected performance on the job. High face validity has been shown to enhance candidate reactions (Hausknecht et al. 2004), may reduce legal challenges (Terpstra et al. 1999), and is likely to enhance stakeholder buy-in and support for a simulation. Fidelity also impacts the degree to which the simulation can serve the function of a realistic job preview (RJP), providing candidates with high quality information on which to self-evaluate whether the job fits with their interests, skills, and preferences (Downs et al. 1978).

In theory, the higher degree of overlap between the simulation and the job present in high fidelity simulations should also result in greater criterion-related validity (McDaniel et al. 2006). However, meta-analytic research examining the validity of simulations ranging from low fidelity (e.g., SJTs) to high fidelity (e.g., assessment centers and work samples) indicates roughly equal validity, in the range of 0.33–0.37 (SJTs: McDaniel et al. 2001; assessment centers: Gaugler et al. 1987; Schmidt and Hunter 1998; work samples: Roth et al. 2005), with one outlying estimate of the validity of work samples being as high as 0.54 (Schmidt and Hunter 1998). Unfortunately, there has been limited research directly comparing low and high fidelity simulations in the same study (see Lievens and Patterson 2011 for an exception where approximately equal validity was observed for an SJT and an assessment center).

Comparative research varying stimuli and response fidelity is also quite limited. Most of the validity research varying stimuli fidelity has focused on video-based versus text-based stimuli for SJTs (e.g., Chan and Schmitt 1997; Lievens and Sackett 2006), and a recent meta-analysis suggests that video-based stimuli exhibit superior criterion-related validity over text-based stimuli (Christian et al. 2010). However, research examining the intermediate ranges of stimuli fidelity allowed for by computer-based simulations using avatars, virtual worlds, and response-dependent stimuli presentation (i.e., branching), is largely lacking.

Research explicitly examining ranges of response fidelity is even more limited. In one of the few such studies, Funke and Schuler (1998) found that the validity of SJTs increased as response fidelity increased, with oral constructed responses (similar to a situational interview) exhibiting greater validity than written constructed responses which had greater validity than multiple-choice responses. Technological advances, such as webcams for recording oral responses to be scored later and computer-based scoring of constructed written responses (Rudner et al. 2006), are making it more economically and administratively feasible for simulations to include intermediate ranges of response fidelity. The current research literature does not provide sufficient guidance in helping the practitioner determine how much fidelity, what kind of fidelity, and under what conditions fidelity results in enhanced criterion validity. Thus, it is our recommendation that other considerations (e.g., stakeholder interests, costs, and simulation purpose) drive fidelity decisions, for the time being.

Aside from a lack of clear evidence that high fidelity simulations have correspondingly high validity, higher fidelity simulations may also bring with them some key disadvantages. For example, they can require greater investments both to develop and administer (e.g., Motowidlo et al. 1990; Patterson et al. 2009), although this depends to an extent upon the specific type and degree of fidelity sought. A high fidelity simulation for retail employees conducted in a store, with trained role-players, and requiring the use of actual computer systems to perform job tasks (e.g., looking up product information and placing an order on a company website), may cost less to initially develop but significantly more to administer than a mid-fidelity computer-based simulation involving video-based stimuli, a simulated company website that tracks user inputs, and multiple-choice responses. Of course, a text-based SJT would result in the lowest development and administrative costs. Given their greater degree of overlap with the target job, high fidelity simulations are also less generalizable across jobs than their low fidelity counterparts. To the extent that a simulation is focused on a narrow set of technical skills required for only one job or a subset of jobs, generalizability may be less of a concern. However, for simulations focused on broader skill sets (e.g., teamwork or managerial skills), some sacrifice in the fidelity of the stimuli (e.g., using a generic organizational context rather than a marketing, finance, or other department-specific context) may be justified to enhance generalizability across jobs, units, and departments.

2.2.2 Knowledge, Skills, Abilities, and Other Characteristics

As noted previously, the flexibility of simulations allows assessment of an immense range of KSAOs. Simulations can be designed to assess very technical (e.g., welding, financial analysis, and specific job knowledge) or more general and broadly applicable KSAOs (e.g., written communication, problem solving, and teamwork). As with any selection tool, the choice of exactly which KSAOs to target should be based on job analytic data. However, in the context of simulations, the measurement of particular KSAOs may both impact and be impacted by several other factors, including the amount and type of job information necessary for development and the fidelity requirements of the simulation. In addition, traditional test development concerns like validity and subgroup differences (e.g., Black–White score differences) may contain an additional layer of complexity when examined within the simulation context. Each of these issues is described later.

The KSAOs targeted in a particular simulation can influence both its fidelity and the type of job information required for development. Simulations focused on the assessment of highly specific or technical skills may require higher fidelity to the extent that the behaviors evincing these skills can only be produced and evaluated in a high fidelity environment. For example, to determine a candidate's ability to perform a complex welding task, the candidate must be given the opportunity to use the appropriate equipment and demonstrate completion of the appropriate welds. For highly job-specific and technical skills such as these, a high fidelity simulation is warranted. At the other end of the continuum, however, candidates for a management position may have their teamwork skills assessed in multiple ways including role-plays, SJTs, or situational interviews. In this instance, because the skills being measured are more broadly generalizable, lower fidelity simulations are an option. While the focus on technical KSAOs and the degree of fidelity required is not wholly confounded, in general, a focus on more technical KSAOs and/or higher fidelity will require more thorough job and task analysis to generate the detailed task data necessary for simulation development. When focusing on more general KSAOs, a competency-based or critical incident approach may be appropriate. More detailed discussions of the types and amount of information required for the development of simulations targeting different KSAOs can be found in Felker et al. (2007) and Whetzel et al. (2012).

Although job simulations have often been stated to have high criterion-related validity and relatively small subgroup differences (e.g., Cascio and Aguinis 2005; Hoffman and Thornton 1997; Schmitt et al. 1996), recent research focusing explicitly on the constructs being measured by job simulations is beginning to paint a more nuanced picture. For example, Roth et al. (2008) conducted a meta-analysis and found that simulations saturated with cognitive ability (e.g., those assessing job knowledge or having high information processing requirements, like many inbox simulations) have much larger Black–White differences than those saturated with social skills (e.g., role-plays or presentations), giving us some insight that both the constructs being assessed and the method by which we are assessing them are jointly contributing to the outcomes of the assessment. Whetzel et al. (2008) observed similar results in the context of a meta-analysis examining subgroup differences as a function of the cognitive loading of SJTs. This is not to say that blanket efforts should be made to reduce the level of cognitive loading in simulations, as doing so has the potential to mitigate subgroup differences while also reducing validity (Ployhart and Holtz 2008). However, practitioners and organizations may be advised to consider modifying simulations slightly to achieve the best of both worlds. Whetzel et al. (2012) provide some suggestions here, including asking assessment center participants to describe their responses verbally as opposed to writing them down. To complicate matters further, two recent meta-analyses have demonstrated some differences in the criterion-related validity of assessment centers and SJTs focused on different constructs, though the results are less compelling than those found for subgroup differences (Arthur et al. 2003; Christian et al. 2010). As researchers have only just begun to explicitly distinguish between constructs and assessment methods when looking at the validity of available tools (Arthur and Villado 2008), this remains an important area for future research. In order for such research to be maximally useful in practice, we ultimately need to determine which constructs are best measured by which methods and at which level of fidelity. Until we know more about how constructs and methods mutually impact simulation performance, practitioners should be aware that changes to either one may result in unintended consequences.

2.2.3 Selecting an Off-the-Shelf Simulation Tool

There are many well-developed commercially available OTS simulations that one can choose to use rather than build a simulation from the ground up. However, OTS options are, for the most part, targeted toward broad job families such as call center, clerical, manufacturing, or managerial. As a result, they may require some sacrifice in stimuli fidelity because they will not be leveraging the context and job materials applicable for a specific job and organization. To mitigate these concerns, those wishing to use an OTS simulation should be cautious to ensure it is focused on the measurement of those KSAOs identified as critical in a local job analysis. That said, the advantages of using an OTS solution can include rapid implementation, established procedures for scoring, established validity, a reduction in the internal resources needed for development (e.g., I-O psychology expertise, SME input), and lower implementation costs. To help determine whether an OTS solution is right for you, we recommend that practitioners consider the following questions:

- What KSAOs do I need to measure? How does the OTS simulation capture these?
- Has the OTS simulation been developed by test experts (e.g., I-O psychologists) from a reputable firm?
- What is the test developer's history of legal challenges to assessments they have developed?
- What kind of validation support does the test developer provide, if needed? Litigation support?
- What technical documentation is available for the OTS simulation? What were the characteristics of the development and normative samples?
- What level of fidelity is present in the OTS simulation and does it align with stakeholder interests and desires?
- How will the OTS simulation be administered? If it is technology-based, does the administration provider have the necessary resources (e.g., computer servers) to meet the demands of my candidate flow? How will candidate questions, concerns, and complaints regarding the testing process be handled? Can it be delivered in an unproctored setting?
- How will qualification standards be established? Does the test developer provide recommended norms?
- How will simulation data be stored and shared with my organization? What data security protocols are in place?
- What are the upfront (i.e., customization and implementation) and ongoing (i.e., administration) costs associated with the use of this solution compared to those of similar tools?
- Can the test developer share case studies or return on investment analyses from previous uses of the tool?

While there is no one right set of answers to the questions above, giving careful thought to each should enable practitioners to make better decisions as to which OTS solution may best meet their needs.

2.2.4 Approach to Scoring

The nature of a particular simulation will, in many cases, determine whether it is best scored by human raters, automated methods, or a hybrid of the two, so it is important to consider any necessary scoring requirements (e.g., "We don't have the resources to use human raters on an ongoing basis") during the planning phase of the development process. In general, simulations with higher response fidelity (e.g., role-plays, physical performance of a task, and orally constructed responses) will require human scorers, while those with lower response fidelity (e.g., responding in a multiple-choice format) can leverage automated scoring, though there are some exceptions to this rule (e.g., where simulation content and automation intertwine, such as simulations to assess computer skills; see Barr and Coughlin 2013, this volume for additional details). The exact KSAOs being assessed can also influence which type of scoring is most appropriate, as some constructs requiring qualitative evaluation (e.g., oral communication or coaching skills) require inferences and judgments that, at this point, technology is not capable of automating. Costs may also be a significant consideration, as human scoring obviously requires a greater investment on an ongoing basis than does the execution of automated scoring. As automated scoring is covered in great detail elsewhere (see Sydell 2013, this volume), we will focus on some of the considerations that go into human scoring of simulations.

In executing human scoring of simulations there are a number of things that must be considered and planned for. First, detailed rating guidelines or checklists need to be created as part of the development process to guide the scoring. These guidelines should focus on observable behaviors and not rely on inferences about internal thought processes of the candidates. Assessors must also be familiar with the simulation and well calibrated in using the rating guidelines. We recommend including samples of different performance levels to be reviewed and discussed during the training. These samples serve as anchors that help calibrate assessors and are useful as references in making ratings in the live simulation setting. Choice of assessors (i.e., psychologists, managers, and job experts) is also important to consider. Research suggests that psychologists are better at discriminating among different constructs measured in a single simulation (Lievens and Conway 2001), but for some simulations requiring familiarity with a task that can only be gained with years of experience or formal education (e.g., welding or computer programming) it will not be an option to use psychologists as assessors. One way to mitigate this concern is to use a hybrid approach where psychologists observe technical assessors and assist them in making their ratings. Regardless of who the assessors are, the total number of constructs assessed should be limited to no more than five, as several studies have demonstrated that assessors are typically unable to make reliable distinctions among a greater number of dimensions (e.g., Thornton and Byham 1982; Shore et al. 1990). If multiple assessors will be observing and scoring candidate behavior, which is recommended, then a procedure for combining scores across assessors must also be established. Mechanical combination (e.g., averaging scores across assessors) has been shown to be superior to consensus or clinical-based methods (e.g., discussing scores and coming to consensus) in forming overall candidate scores (e.g., Feltham 1988; Russell 1985). However, if a consensus discussion is desired for other compelling reasons (e.g., stakeholder concerns), then a formal procedure should be established and followed to obtain final scores (see Jansen 2012, for an example of an effective consensus meeting procedure).

2.2.5 Proctored Versus Unproctored

When determining whether a simulation requires a proctor, the concerns to be addressed are largely the same as for most other selection tests, revolving around the security of test content, the potential for cheating, and administrative efficiency. While the decision to use a proctor may, in some cases, be driven by the design of the simulation itself, there is some flexibility within these constraints. For example, a simulation requiring candidates to appear on site and use an expensive piece of equipment such as a flight simulator will by its very nature require a proctor. Other types of simulations, such as telephone-based role-plays, may be amenable to either proctored or unproctored settings. While organizations may have concerns that the person engaging in the role-play is not the actual job candidate, advances in voiceprint and streaming video/webcam technology now allow us to verify the candidate's identity during the simulation (e.g., by asking the candidate to hold up his or her identification), or at a later date (e.g., by ensuring the candidate looks like the person in the video).

As technology continues to evolve, we are seeing increases in the types and volume of simulations being administered in an online, unproctored format, even those that include a cognitive ability component. Key benefits to this approach include lower administration costs and greater ease with which candidates can apply for the position. Disadvantages include the necessary diligence to implement advanced security features such as creating parallel forms, item randomization, response-dependent branching, and making it more difficult for candidates to capture test content and share it with others. In large part, the decision to administer simulations in an unproctored setting will depend on the level of comfort an organization has with such an approach. Practitioners should consider the necessity of a proctor as early in the simulation design phase as possible, so that they can confirm the necessary resources will be available when the simulation is put into use with live candidates.

2.2.6 Use of Technology

The use of job simulations in the selection context warrants some additional discussion regarding the use of technology beyond that which has been noted earlier. Specifically, modern technology, regardless of how advanced it is, will carry with it some limitations regarding availability, bandwidth, flexibility, and ease of use. Such limitations can be experienced during the design phase (e.g., how the test will be programmed and set up), during the administration phase (e.g., whether the necessary technology is available to candidates, particularly in the global context of developing countries), and in the post-administration environment (e.g., how simulation scores will be databased and stored). While it may be tempting to defer the management of such challenges to a test development firm, even organizations looking to implement an OTS solution will often be faced with these issues as they begin to integrate simulation technologies with their existing human resource information systems (e.g., linking simulations with an applicant tracking system, ATS). While complete coverage of the technological implications for simulation programming, implementation, and data management is beyond the scope of this chapter, we also touch on some of these topics in Sect. 2.3, Role of the Simulation in the Candidate Lifecycle, and we refer the reader to Hawkes (2013, this volume).

2.2.7 Managing the Interests of Simulation Stakeholders

As with any wide-reaching organizational intervention, the stakeholders impacted by simulations each bring with them a distinct set of motives, opinions, and goals. In addition, each group likely has differing levels of knowledge and expertise with regard to the assessment process. To help practitioners navigate needs of each group, we have outlined some of them below, along with recommendations regarding how they can be addressed. It will become apparent that simulations are better suited to addressing some stakeholder concerns than others. Our hope is that by exploring these issues, practitioners will be able to make more educated decisions regarding whether and how to use simulations for selection purposes.

To begin, recruiting and staffing functions are challenged by the availability of qualified labor, cultural changes, and increasingly diverse candidate pools (Ployhart 2006). However, as noted earlier, research indicating that simulations exhibit less adverse impact than more traditional assessment methods (Schmitt et al. 1996) was telling only part of the story (e.g., Roth et al. 2008). As the labor market evolves and candidate demographics continue to change, recruiting and staffing functions should be careful to consider the extent to which cognitive abilities are measured in simulations, just as they would for any other assessment methodology. As aptly pointed out by McDaniel et al. (2007), this should also include giving thought to the way in which simulation instructions are given to candidates. Because knowledge instructions ("Please select the best possible response to this situation") are more closely correlated with measures of cognitive ability than behavioral instructions ("Please select how you would respond in this situation"), reductions in adverse impact realized by moving to a simulation format may be mitigated if thought is not given to the instructions in addition to test content and medium.

A second key set of stakeholders are the hiring managers and their respective lines of business, who are often facing pressures to reduce the cycle time to fill empty requisitions. In this case, simulations may provide an important benefit. Consider, for example, an online and catalogue-based clothing retailer who wishes to hire call center representatives for sales positions. The position requires successful employees to communicate with customers, make decisions regarding upselling opportunities, search through a computer database to check the availability of items and sizes, and enter customer and order information via a keyboard quickly and accurately into an ordering system. If we were to imagine what this process might look like without using simulations to measure candidate skills and abilities, we might envision a multiple-hurdle process that includes a resume screen and reference check to assess keyboarding skills and experience using customer databases, a cognitive ability test to measure decision-making ability, and an interview to assess communication skills. On the other hand, an organization may instead decide to implement a single-hurdle process consisting of a well-designed simulation for call center customer sales positions. The simulation might involve listening and responding to audio files of "customers" calling into the center, or even live assessors role-playing the part of customers. In this scenario, candidates would listen to the "customers" while simultaneously entering information provided into a database and interacting with a website to search for available products and sizes. Candidates would also have to respond to customer requests and engage in upselling behaviors, either by speaking with a live role-player or by choosing a response option on the screen. In this simulated environment, the organization is now able to more efficiently screen candidates on a variety of skills and abilities. One test is able to measure multiple KSAOs, and scheduling challenges like the ones faced when setting up interviews can be reduced or avoided. As a result, when they are well designed and implemented, simulations can help organizations shorten the selection process and reduce cycle time without sacrificing the quality of successful candidates.

A third important group to consider when implementing simulations for selection is human resources (HR) personnel looking to understand the strengths and areas of opportunity in their talent base. Often, HR must balance the benefits provided by simulations with the potential for resistance to assessment from current employees and hiring managers, particularly those employees who may be applying for internal positions. As such, HR personnel, and often I-O psychologists, must "sell" the assessment process to an audience without assessment expertise. While the higher level of fidelity and increased face validity of simulations can make it easier for HR teams to obtain buy-in and support from end users like hiring managers, it is quite evident that some organizational stakeholders still struggle with the idea of simulations in general, for example, the hypothetical nature of role-plays.

To address these concerns, HR teams must first understand the motivations of simulation end users. Surprisingly, however, there has been a lack of research on internal stakeholder reactions to assessment, and we recommend that additional studies be completed in this area. In the meantime, a recent report on barriers to simulation implementation may be of some help. In a survey of federal agencies completed by the United States Merit Systems Protection Board (2009), the most frequently cited barriers to using job simulations for selection included insufficient time to develop assessments, lack of knowledge on the part of stakeholders such that they are not comfortable determining whether simulations are beneficial, insufficient resources to develop and/or administer simulations, and insufficient resources to train test administrators. Addressing these concerns will be critical for HR to successfully convince end users of the benefits simulations can provide. Further, to the extent that HR can take a more consultative approach and facilitate honest dialogue with hiring managers and simulation end users, this report found that they may be able to promote greater openness to simulation use.

A final and perhaps most important group of stakeholders is comprised of the job candidates themselves. While it can be assumed that the motivated test-taker has selection for the target job as his primary goal, his reaction to the recruiting and selection process itself can impact his attraction to the organization, even if he is not selected (Chapman et al. 2005). Because we know that candidates tend to prefer assessments they perceive as job-related (i.e., face valid; Gilliland and Cherry 2000; Richman-Hirsch et al. 2000), and because simulations are, by their very nature, more face valid than their traditional counterparts, simulations can lead to more positive candidate reactions. In addition, the wide availability of the Internet and social networking sites means that increasing numbers of candidates are sharing links and information about jobs and job simulations. This can translate to more widespread knowledge of a particular organization and potential interest (or disinterest) in working there. As a result, modern organizations can and should consider simulations to be an important branding and recruitment tool when looking to attract new talent

(Yu and Cable 2012). It should also be noted that when considering job candidate reactions to selection procedures, not all candidates are the same. While some researchers have broken out candidate groups based on key demographic information such as race and gender (see Ryan and Ployhart 2000 for an overview), little work to date has been done on the motivations and reactions of specialized candidate groups (e.g., candidates who are members of labor unions, internal versus external candidates). Additional work in this area is needed. We direct the reader to Bruk-Lee et al. (2013, this volume), for a more comprehensive look at candidate reactions to simulations in the selection context.

When choosing to implement simulations in lieu of or in addition to more traditional assessment methods, it is important to give due consideration to each of these critical stakeholder groups. While we have provided some tactical approaches to responding to their unique needs, we also strongly advocate the use of a comprehensive change management approach that includes a robust communication strategy when rolling out a new selection procedure of any kind. We concur with Higgs et al. (2000), who note that "a selection process is a product to be marketed in an organization," and we recommend that internal and external practitioners carefully consider the following questions before using simulations in the selection context:

- Why is the simulation being implemented? What value will it bring?
- How can we ensure high-level organizational support for the new process?
- Who will be impacted by the simulation (e.g., external candidates, internal candidates, hiring managers, and recruiters)? How will they be impacted?
- How will the simulation be positioned and "marketed" to candidates, recruiters and hiring managers?
- Will training be required for test administrators? Hiring managers? Others? How will this training occur?
- How will simulation results be stored and communicated? What level of detail will be provided and to whom?

While complete coverage of the change management approach needed to successfully implement a simulation for selection purposes is beyond the scope of this chapter, the interested reader is advised to review Muchinsky (2004) for more information.

2.3 Role of the Simulation in the Candidate Lifecycle

Like other assessment methods, simulations can play an important role before, during, and after their administration. However, given their greater fidelity and the fact that they are often technology-enabled, simulations make unique contributions to the experiences of organizations and individuals that are worthy of discussion here. This section will take a chronological look at the issues facing individuals and organizations when simulations are used for selection purposes—discussing their impact

Table 2.1 Role of the simulation in the candidate lifecycle

	Individual factors and concerns	Organizational factors and concerns
Pre-Administration	Perceptions and expectations regarding the organization and its brand Perceptions and expectations regarding assessments Ease of access to the assessment (e.g., technology requirements and location)	Differential candidate access to the assessment (e.g., technology requirements and location) Development and maintenance costs, including translation concerns where applicable Opportunities to use assessments as a means of candidate recruitment and attraction Test security issues (e.g., the extent to which simulation content is available online)
Administration	Attitudes toward the assessment process (e.g., candidate reactions and test anxiety) Obtaining a realistic preview of the job to help evaluate person-job fit	Providing a realistic preview of the job so that undesirable candidates self-select out of the process Providing reasonable test accommodations Administration costs/efficiency Cheating (e.g., candidates using helpful confederates in an unproctored setting)
Post-Administration	Candidate reactions to the assessment process Post-assessment behaviors that impact the candidate (e.g., accepting or declining a job offer)	Post-assessment candidate behaviors that impact the organization (e.g., accepting or declining a job offer, promoting or badmouthing the organization, and likelihood of challenging assessment results) Subgroup performance differences and adverse impact Decisions regarding how simulation data will be stored, used, and communicated (e.g., what level of detail will be shared with hiring managers and providing feedback as a means of on-boarding new employees)

before, during, and after their administration—with the goal of equipping practitioners with the information they need to make more informed decisions regarding the use of simulations in their or their client's organizations. It will become evident that some of the issues described below (e.g., candidate reactions and test security concerns) are covered elsewhere in depth, as they are not unique to simulation use. In an effort to provide the reader with a complete list of such concerns, we have mentioned them here, but they will only be reviewed at a high level; we will focus our discussion on those topics we believe to be distinguishing with regard to simulations as opposed to other assessment methods. An overview of the issues mentioned is provided in Table 2.1.

2.3.1 Pre-Administration

Concerns experienced by individuals prior to simulation administration primarily fall into two buckets: candidate perceptions/expectations and ease of access to the assessment. While research has explored a wide variety of factors that can impact candidate attraction outcomes (see Chapman et al. 2005 for an overview), we are concerned primarily here with how candidates perceive the assessment (simulation) process, and their attraction to the organization and its brand. As increasing numbers of organizations look to their customer base as a feeder for their candidate pool (consider, for example, the quick service restaurant or retail industries), it is likely that candidate perceptions of an organization will have an increased impact on job pursuit intentions. Candidate reactions and the impact of simulations on candidate perceptions are discussed in greater depth in Bruk-Lee et al. (2013, this volume).

With regard to ease of access to the assessment, the challenges faced by individuals are largely the same for simulations as for more traditional test types. That is, proctored tests require candidates to travel to a test site. Unproctored tests require candidates to have access to a test administration method, most often the Internet, and any necessary supporting technology such as a computer and a webcam. For those without the economic means to easily access the Internet from home, poorer test performance may result given the greater cognitive demands associated with using an unfamiliar testing medium while simultaneously being asked to perform in a high-stakes environment. This issue may be further exaggerated in the case of a novice job seeker as she participates in a job simulation for a position in which she has little to no experience. Some have even argued that lack of easy access to the Internet should be viewed as a disability requiring a reasonable test accommodation (Naglieri et al. 2004).

Differential candidate access to the assessment is a concern for organizations as well. As technology continues to evolve and technology-enabled simulations become more sophisticated, we might expect to see larger differences in the level of access candidates have to unproctored tests, such that more advanced operating systems and internet browsers are required to participate in the assessment process. To the extent that different candidate demographics have differential access to these technologies, the diversity of candidate pools may be impacted, potentially causing an increase in the frequency of legal challenges. As a result, organizations looking to leverage cutting-edge technology for the administration of an unproctored simulation should carefully consider the downstream impact to their candidate pool and be prepared to offer test administration alternatives, such as taking the test at a location with the required equipment available.

A second organizational concern in the pre-administration phase is the cost required to develop and maintain simulation content. Like their traditional counterparts, simulations will cost more to create and sustain to the extent that they are customized, highly technical, and require advanced technology to deliver. However, given the speed with which modern jobs and technology are advancing, technology-based simulations may require more frequent updating than traditional assessments or telephone-based role-plays. While research has consistently indicated more favorable candidate reactions to simulations and work samples than to paper-and-pencil measures (Chan and Schmitt 2004), we are unaware of any research that has examined the extent to which an outdated simulation can potentially negatively impact candidate reactions. While it is evident that simulations which are outdated in the sense that they no longer reflect critical aspects of the job have the potential to increase litigation risk, the extent of the impact of outdated simulation style elements (e.g., referring to a pager message as opposed to a text message) remains unclear. In other words, is using an outdated simulation the same as or potentially worse than a more traditional assessment method from a candidate reactions perspective? While the answers to these questions are yet to be discovered, organizations should at a minimum consider the extent to which the target jobs are changing and could require the updating of simulation content. For this reason, when making an investment to develop a technology-based simulation, we recommend the use of a future-oriented job analysis whenever possible, along with making plans to revisit and potentially update simulation content on a regular basis.

For global organizations, the need to translate simulation content will also impact cost, and under some circumstances, translating simulation content for global use can require some additional diligence beyond that required for more traditional assessment types. Simulations are often designed to look and feel like the target job. This means that a computer screen may actually look and feel like a website, database, or other job tool. When translating content from one language to another, unlike a paper-and-pencil assessment, the look and feel of the words on the screen become important characteristics of the test itself. Thus, challenges can arise if the number of characters/words it takes to say something in one language is significantly different than in another language. To ensure an equivalent look and feel in different languages, words may need to be adjusted so that they fit on the screen and/or are laid out in a similar manner in both versions of the test. However, as is well known, changes of this sort may create issues with regard to the psychometric equivalence of the test across multiple languages. As a result, to the extent possible, practitioners are urged to even more carefully consider word choice when developing simulation content in its original language. Eliminating jargon and idiomatic expressions may no longer suffice. Thought should be given to keeping text as simple as possible so that translated versions of the simulation more readily mirror the content of the native language and so that the look and feel of the simulation is not negatively impacted after translation. In addition, practitioners should use appropriate back translation services by experienced translators to mitigate potential test equivalence issues. If concerns regarding test equivalence still exist, studies to examine the psychometric properties of both versions may be warranted. To identify more specific recommendations in this regard, additional research is needed regarding the physical appearance of tests in the selection context (Ryan and Ployhart 2000).

This leads us to the third concern organizations face during the pre-administration phase: using simulations to attract and recruit candidates. As candidates continue to desire the ability to conveniently look for jobs close to home or across the globe, the popularity of using the Internet for this purpose is likely to remain high. Organizations have responded in kind, increasing their use of the Internet to attract and assess candidates. This creates a critical branding and recruitment opportunity for organizations looking to hire top talent. As such, organizations who choose to share job simulation content online (e.g., to unproctored test-takers who are able to access the test simply by visiting the company's website) should consider the speed with which this information can be communicated to a large audience, i.e., how quickly something can "go viral," and the message that it will send to those who see it. One recommendation in this area is that organizations partner their branding and marketing teams with test developers during early phases of simulation development to ensure that assessments are psychometrically sound while simultaneously portraying the organization in accordance with its brand standards.

A fourth and final consideration for organizations is the security of simulation content. For many companies, there is a strong desire to keep their pre-employment assessment methods under lock and key. Understandably, they are concerned that answer keys will become public, enabling candidates who do not possess the KSAOs required for success in the position to make it through the selection process. This often requires organizations to keep not only their answer keys, but also the assessment content itself secure. Unlike some traditional assessment methods (e.g., cognitive abilities tests or measures of personality), however, simulations often contain content that informed job candidates could reasonably be able to predict. For example, a candidate for a position as an airline pilot likely understands that, if hired, he will one day have to fly a plane. Thus, requiring job candidates to successfully operate a flight simulator would be no surprise to anyone applying for the position. As such, an airline might even publicize the fact that they use flight simulators to select highly qualified candidates, resulting in several key benefits. First, this may discourage unqualified candidates who cannot perform the task from applying. Second, making simulation content available to the public can provide potential candidates an opportunity to practice the skills required for success. Third, top candidates may in fact be more attracted to organizations that use engaging, cutting-edge selection processes, viewing them as leaders in their field. Fourth, if the simulation is developed well and adequately represents the organization's culture and brand, candidates who enjoy the assessment process may also be a better fit for the company and recommend it to others. We caution, however, that the benefits to making simulation content available online are not universal. Like any assessment method, the possibility that candidates will cheat and "game" the test still exists. As simulation content continues to improve and we are able to achieve closer to 100 % overlap between simulation content and the target job, we will be able to move from simulating job activities to enabling a realistic job tryout in a safe environment. In this case, a candidate's ability to "game" the job tryout would essentially translate to his ability to "game" or do well in the actual job—in other words, his ability to be an effective performer if selected for the position.

2.3.2 During Administration

While in the midst of any pre-employment assessment experience, candidate performance can be impacted by a variety of attitudinal and individual differences, including things like test anxiety, locus of control, and level of motivation/desire to be selected for the target position. Because the outcomes of these factors are not unique to the pre-employment simulation experience, we will not explore them in detail here. However, some thought should be given to whether the level of such variables, e.g., test anxiety, will be increased or decreased in the simulation environment as compared to a more traditional pre-employment assessment method. As noted previously, the candidate reactions chapter (Bruk-Lee et al. 2013) of this volume can provide some guidance in this regard.

A second factor impacting individuals during simulation administration is the extent to which the simulation provides a realistic preview of the target job. If candidates can use the simulation experience to learn more about the position, they will be better able to evaluate how well they might fit the role and choose to self-select out of the process if the position is deemed undesirable. In addition, to the extent that simulation content is made publicly available, savvy candidates may recognize the simulation as an opportunity to practice the skills required for success in the job, essentially using the simulation as a method of pre-employment training. Using simulations as a preview of the job can provide lasting benefits for organizations as well, as RJPs have been shown to lead to lower turnover (Earnest et al. 2011).

During the administration phase, organizations should also consider the need to accommodate persons with disabilities. As noted by Naglieri et al. (2004), internet-based tests—and by extension, simulations delivered in electronic form—are too new for us to understand all of the possible potential accommodations. To date, organizations have often implemented accommodations that were useful for paper-and-pencil-based assessments (e.g., providing candidates with extra test time), and some have offered basic technological accommodations like increasing font size, or allowing the use of screen-reader software (e.g., Job Access With Speech (JAWS) or Window-Eyes). While we may be able to readily identify potential accommodations for some of the demands of technology-enabled simulations (e.g., listening to audio files, watching videos, and clicking on a screen), it is likely that there are still other accommodations yet to be identified. With regard to internet testing, Tippins et al. (2006) recommend that organizations consider accommodations that allow candidates to take tests in non-computer formats, but this becomes more difficult particularly for more intense and psychologically involved online simulations. Future research remains to be done in these areas, and practitioners are urged to keep a close eye on relevant legal cases.

The remaining organizational factors that become relevant during simulation administration parallel the pre-administration issues noted above: cost and security. With regard to cost, organizational stakeholders will be concerned with the level of human and technological resources needed to administer assessments, and how efficiently the simulation can meet the demands of the volume of candidates taking the test. The security concerns at this stage of the process surround the opportunity for candidates to cheat on the test, especially in unproctored settings (e.g., by enlisting the help of a talented confederate to take the test for them). As noted above, these concerns are not unique to the simulation environment, so we will not go into further detail here, though the interested reader may wish to review Ryan and Tippins (2009) for more information.

2.3.3 Post-Administration

Once the simulation is complete, we can classify the concerns of the individual into two broad categories: perceptions (i.e., candidate reactions) and behaviors. The behaviors we refer to are those relevant to the job application and selection process that will have a direct impact on the candidate, specifically the decision to accept or reject a job offer. As before, we encourage the interested reader to review the Bruk-Lee et al. (2013) chapter of this volume for a deeper dive on the current state of research regarding candidate reactions and subsequent outcomes for more information.

Organizational considerations during the post-assessment phase begin with those candidate behaviors that have a direct impact on the organization. These include accepting or declining a job offer, promoting or badmouthing the organization, and the likelihood of challenging assessment results. While the first two of these are also addressed in Bruk-Lee et al. (2013), there is a lack of research on the third issue. To date, little work has been done to evaluate the extent to which simulations differ from their traditional counterparts in terms of the likelihood of a legal challenge, though there are some initial findings that assessment centers and work samples tend to be viewed favorably by the courts (Terpstra et al. 1999). In addition, research on candidate reactions that has examined litigation intentions suggests that simulations may lower the chances of a legal challenge (Hausknecht et al. 2004). As a result, until this hypothesis can be more fully explored, we recommend that organizations take steps to ensure candidates view simulations positively.

A second consideration for organizations at this stage is the potential for adverse impact as a result of subgroup performance differences. As noted above, subgroup differences are greater for simulations to the extent that they require higher levels of cognitive ability and information processing. Correspondingly, we would expect subgroup performance differences to be exaggerated if one subgroup has less familiarity with the testing modality, as members of that group will have to focus greater levels of attention on the mechanics of the test (e.g., scrolling with a mouse and clicking on a computer screen) than those already familiar with such activities. For example, despite the fact older generations are using the Internet more often and for a greater variety of activities (Jones and Fox 2009), generational differences may exist with regard to the level of comfort candidates feel in using the technologies necessary to support some types of simulations. Additional research is needed to examine whether candidates from different generations and socioeconomic statuses perform differently on technology-based simulations as opposed to more traditional assessment types.

A final issue for organizations to consider is how simulation data will be stored, used, and communicated, that is, what level of detail will be shared with candidates, hiring managers, and HR. Because simulations have the ability to closely mirror, or even replicate, the conditions of the job, they provide a unique opportunity for organizations to provide feedback to newly hired employees and their managers for the purposes of onboarding and training. Doing so can help to get employees up to speed more quickly than otherwise possible. In addition, in this modern age of "Big Data," savvy companies may use simulation results—for selected and non-selected candidates alike—to study the defining characteristics of their candidate pool such as skill strengths and gaps and how these are changing over time. Companies looking to take a more strategic look at their workforce and sources of talent may find the information simulations can provide to be invaluable in this regard.

In short, individuals and organizations have a variety of concerns before, during, and after simulation use. At times, the concerns of the individual and the concerns of the organization are similar, e.g., how the candidate will react to the organization's brand and the fairness of the process and whether particular groups of candidates will have differential access to the test. However, organizational concerns regarding simulation use also extend to familiar areas in the field of personnel selection such as test content, security, and adverse impact.

2.4 The Future of Using Simulations for Selection

In projecting toward the future, we will share a quote from novelist William Gibson, "The future is already here—it's just not evenly distributed." The two "projections" about simulation design and application that we share in this concluding section are already part of today's landscape in some organizations for some populations in some parts of the world. Our prognostication is that these will simply become ubiquitous in the next 3–5 years.

2.4.1 Gaming

The design of simulations will change. Members of the generation entering the labor market in the middle of the second decade of the twenty-first century have played video games all their lives. These games are engaging and increasingly they are social, involving interaction with other players. Consistent with a desire that assessment tools promote the employment brand, engage candidates, and produce valid scores reflective of motivated test-takers, simulations are likely to take on additional, game-like features. The United States Military has been at the forefront of these efforts (see their Special Team Challenge or Patriot Missile System games at http://www.goarmy.com/downloads/games.html). There are conferences around gaming in HR applications, including assessments (e.g., http://www.GSummit.com) and organizations like Reckitt Benckiser have embraced whimsical, fun-oriented games that get at person-environment fit and other job relevant constructs as part of the employee selection process (e.g., http://www.InsanelyDriven.com). The simulations typically used for pre-employment assessment today may look in a few years like "Pong" looks to those playing today's online games.

2.4.2 *What Test? What Candidate?*

The boundary between an online game played for fun and challenge and a preemployment assessment that influences hiring decisions is going to get increasingly blurry. There are recruiting organizations that post challenging games on the Internet for which undergraduate or professional school students register in order to receive prizes. These simulations are used to measure such constructs as decision-making under stress and creative problem solving. These organizations then identify for hiring organizations those students on each campus who had the highest scores. These students become the priority candidates approached during recruiting visits on campus.

Over the course of time, simulations will increasingly be embedded within company websites for a wide range of potential audiences. At one end are customers who want to know more about what it feels like to work at the target company and might become candidates if sufficiently intrigued by the simulation experience. On the other end, there are traditional, focused job candidates who are submitting to what they know to be a pre-employment assessment. Our key point is that these ambiguities surrounding who is really and consciously a candidate and when an assessment is a test are particularly likely to arise in deploying simulations rather than other assessment techniques. As a result, there will be an increased need to carefully consider the ethics associated with allowing for informed consent of test-takers regarding assessment purpose and use.

2.5 Conclusion

Simulations have a long and strong track record of validity as selection tools in high-stakes environments, though as in the case of all selection tools, there are multiple and complex factors that may, in specific settings, impact optimal simulation design, implementation, and delivery. We have discussed a wide array of these factors in this chapter. Where possible, we have attempted to not only highlight the complexities but to provide research- and practice-oriented guidance for practitioners. However, it should be clear that there remain many open research questions regarding the use of simulations for selection purposes. Some of these questions are fundamental:

- When and how does greater simulation fidelity produce stronger validities? What is the relative importance of stimulus versus response fidelity?
- What are the design elements of simulations that most strongly contribute to candidate perceptions of employer branding and procedural fairness?
- As simulations evolve into increasingly game-like experiences, what generational and other subgroup performance differences will emerge and how will these differences affect simulation validities?

We conclude, then, by urging continued academic–practitioner collaboration to answer these and many other questions raised here about the use of simulations in selection through rigorous, theory-guided empirical research.

References

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435–442.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–154.

Barr, R., & Coughlin, C. (2013). Software simulations. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Bruk-Lee, V., Drew, E., & Hawkes, B. (2013). Applicant reactions. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Cascio, W. F., Aguinis, H. (2005). *Applied psychology in human resource management (6th ed.).* Upper Saddle River: Pearson/Prentice Hall.

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. (2004). An agenda for future research on applicant reactions to selection procedures: A construct-oriented approach. *International Journal of Selection and Assessment, 12,* 9–23.

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology, 90,* 928–944.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63,* 83–117.

Downs, S., Farr, R. M., & Colbeck, L. (1978). Self appraisal: A convergence of selection and appraisal. *Journal of Occupational Psychology, 51,* 271–278.

Earnest, D. R., Allen, D. G., & Landis, R. S. (2011). Mechanisms linking realistic job previews with turnover: A meta-analytic path analysis. *Personnel Psychology, 64,* 865–897.

Felker, D. B., Curtin, P. J., & Rose, A. M. (2007). Tests of Job Performance. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 319–348). Mahwah: Lawrence Erlbaum.

Feltham, R. (1988). Assessment centre decision making: Judgmental vs. mechanical. *Journal of Occupational Psychology, 61,* 237–241.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6,* 115–123.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18,* 694–734.

Gilliland, S. W., & Cherry, B. (2000). Managing customers of selection. In J. F. Kehoe (Ed.), *Managing selection in changing organizations* (pp. 158–196). San Francisco: Jossey-Bass.

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.

Hawkes, B. (2013). Simulation Technologies. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Higgs, A. C., Papper, E. M., & Carr, L. S. (2000). Integrating selection with other organizational processes and systems. In J. F. Kehoe (Ed.), *Managing selection in changing organizations* (pp. 73–122). San Francisco: Jossey-Bass.

Hoffman, C. C., & Thornton III, G. C. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50,* 455–470.

Jansen, P. G. W. (2012). How to apply a dimension-based assessment center. In D. J. R. Jackson, C. E. Lance & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 121–140). New York: Routledge.

Jones, S., & Fox, S. (2009). Generations online in 2009. Pew Internet & American Life Project. http://pewinternet.org/Reports/2009/Generations-Online-in-2009.aspx. Accessed 2 Jan 2013.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86,* 1202–1222.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96,* 927.

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91,* 1181–1188.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N., & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In R. Ployhart & J. Weekley (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 183–204). San Francisco: Jossey-Bass.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640.

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.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., et al. (2004). Psychological testing on the internet: New problems, old issues. *American Psychologist, 59,* 150–162.

Patterson, F., Baron, H., Carr, V., Plint, S., & Lane, P. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Medical Education, 43,* 50–57.

Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management, 32,* 868–897.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85,* 880–887.

Roth, P. L., Bobko, P., & McFarland, L. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58,* 1009–1037.

Roth, P., Bobko, P., McFarland, L., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. *Personnel Psychology, 61,* 637–662.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. The Journal of Technology, Learning and Assessment, 4. http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1651/1493. Accessed 15 Dec 2012.

Russell, C. J. (1985). Individual decision processes in an assessment center. *Journal of Applied Psychology, 70,* 737–746.

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., & Tippins, N. T. (2009). *Designing and implementing global selection systems*. Malden: Wiley-Blackwell.

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

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 115–139). New York: Wiley.

Shore, T. H., Thornton III, G. C., & McFarlane Shore, L. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43,* 101–116.

Sydell, E. (2013). Simulation Scoring. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Terpstra, D. E., Amin Mohamed, A., & Kethley, R. B. (1999). An analysis of federal court cases involving nine selection devices. *International Journal of Selection and Assessment, 7,* 26–34.

Thornton III, G. C. & Byham, W. C. (1982). *Assessment centers and managerial performance*. New York: Academic Press.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., et al. (2006). Unproctored internet testing in employment settings. *Personnel Psychology, 59,* 189–225.

United States Merit Systems Protection Board. (2009). Job simulations: Trying out for a federal job. http://www.mspb.gov/netsearch/viewdocs.aspx?docnumber=452039&version=453207&application=ACROBAT. Accessed 25 Nov 2012.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372–376.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21,* 291–309.

Whetzel, D. L., McDaniel, M. A., & Pollack, J. M. (2012). Work Simulations. In M. A. Wilson, W. Bennett, S. G. Gibson, & G. M. Alliger (Eds.), *The handbook of work analysis: Methods, systems, applications and science of work measurement in organizations* (pp. 401–418). New York: Routledge.

Yu, K. Y. T., & Cable, D. M. (2012). Recruitment and competitive advantage: A brand equity perspective. In S. W. J. Kozlowski (Ed.), *The Oxford handbook of organizational psychology: Vol. 1* (pp. 197–220). New York: Oxford University Press.

Chapter 3
Candidate Reactions to Simulations and Media-Rich Assessments in Personnel Selection

Valentina Bruk-Lee, Erica N. Drew and Ben Hawkes

Over the past three decades, interest in the reactions of candidates to selection procedures has continued to grow with early research focused on understanding candidate preferences for different types of selection assessments (see Cascio and Phillips 1979). While much of this work addressed questions of practical influence, widely cited theoretical frameworks did not emerge until the early 1990s (see Gilliland 1993; Schuler 1993). Consequently, studies began to focus on the personal and situational influences affecting a candidate's reaction to a selection procedure, and moreover, the relevant outcomes that these reactions could impact. Today it is widely recognized that the employment process offers a venue for an exchange of information between the candidate and the hiring organization; candidates demonstrate their knowledge, skills, and abilities (KSAs) and organizations communicate information about their values and culture (Hausknecht et al. 2004). As such, candidate perceptions of assessment procedures are an important consideration for organizations displaying best practices in employee selection.

As the desire for innovation continues to grow, so does the need for understanding how technology impacts candidate reactions to such advancements (see Tippins 2009). Changes in the medium of delivery (computer, Internet) and media content of assessments (two dimensional (2D) or three dimensional (3D) animation, video) have taken place at a faster rate than research in this area has evolved. In this chapter, we present an overview of findings specific to the use of media-rich assessments driven in large by unpublished work, introduce the application of the uncanny valley

V. Bruk-Lee (✉) · Erica N. Drew
Florida International University, Miami, FL, USA
e-mail: vblee@fiu.edu

Erica N. Drew
e-mail: endrew9@yahoo.com

B. Hawkes
Kenexa Limited, London, UK
e-mail: Ben.Hawkes@Kenexa.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_3,

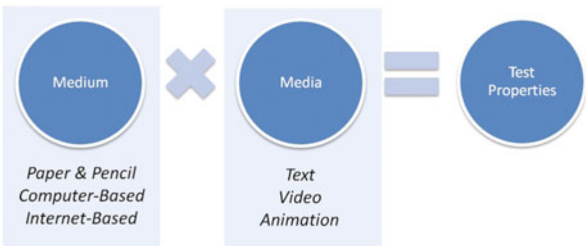
to candidate reactions, and describe areas in need of further consideration by both practitioners and scientists.

Several candidate reaction models have been proposed (see Arvey and Sackett 1993; Schuler 1993), however, the most cited of these is Gilliland's (1993) justice-based model. A brief overview is warranted as a means of framing the relevant findings, in the sections to follow, of studies using advanced technologies. The model describes distributive justice rules that focus on outcome fairness with a special emphasis on feelings of equity (i.e., match between candidate expectations and outcomes), equality, and consideration for special needs. Much of the research, however, has centered on the procedural justice rules, which are grouped into formal characteristics of the selection system, explanations, and interpersonal treatment received during the assessment process. The extent to which an assessment process appears to be valid or content related, known as job relatedness, has received the most attention. Specifically, the focus has been on measuring candidate perceptions of content or criterion related validity, although some have also included candidate perceptions of face validity in its assessment (see Smither et al. 1993). Bauer et al. (2011) present an overview of the potential benefits and challenges posed by technology across each of the procedural justice rules.

3.1 Technology and Its Influence on Candidate Reactions

The introduction of computers and related technology has had a profound impact on personnel selection (Viswesvaran 2003). Modern technologically advanced assessments include a combination of *administration mediums* (paper-and-pencil, computer- and Internet-based testing) and *media types* (text, video, and animation; see Fig. 3.1). Administration mediums and media types are not mutually exclusive; in fact, simulations typically leverage combinations of both. Thus, following a description of administration mediums and associated issues, research regarding candidate reactions to media types across various administration methods will be presented. We concentrate on describing research considering the use of technology in job-centered simulations, including assessment centers and situational judgment tests (SJT) because these assessments are designed to emulate actual job situations, and include situational descriptions that are easily customized with novel innovations in technology. Unfortunately, published research on candidate reactions to technologically advanced simulations remains scant.

Fig. 3.1 Test format as a function of medium and media characteristics



3.1.1 Administration Mediums and Media Types

Before the advent of computers, the ability to create novel and engaging assessments was extremely limited. Presently, the growing capabilities and utilization of computers has provided endless possibilities for enhancing the way assessments are developed and delivered. Technology has introduced interactive features, graphics, video, and animation to assessments that were once limited to simple lines of text and multiple-choice answers. Due to these innovations, *computer-based testing* (CBT) has quickly become a popular medium for administering assessments.

Widespread use of CBT has resulted in a proliferation of technological advancements including innovative items and computer adaptive testing (CAT), each providing insight into candidate capabilities not readily measured by traditional formats. Item innovations have aided in several simulation developments including: item presentation structure, methods of responding, inclusion of media, assessment interactivity with examinee responses, increased item complexity, and more efficient scoring methods (see Parshall et al. 2010 for extensive review of these elements). In general, innovative items are perceived favorably by candidates (Pommerich and Burden 2000; Wendt et al. 2008) and provide a more interactive and engaging test experience (Parshall et al. 2010). Preliminary findings in regard to response actions, such as allowing candidates to drag graphics to form item responses, suggest that such item innovation improves candidate perceptions of job relatedness, engagement, and test ease, as compared to simple computerized multiple-choice formats (Gutierrez 2011a). However, the majority of the research regarding candidate reactions to innovative items has focused on the inclusion of media, which is discussed in a later section.

Another related advancement, CAT, tailors the testing experience to the abilities of the referent examinee by selecting questions based on candidate response accuracy. By administering questions that are progressively harder or easier according to the correctness of answers indicated before, the assessment is able to accurately pinpoint the examinee's standing on the focal ability. Existing evidence considering candidate acceptability of CAT is inconsistent, necessitating the need for more research in this area. For example, a recent study found that incumbent reactions to CAT were more favorable after they took the exam than after they were provided information about how CAT works. Further, 40% reported that using CAT assessments to make hiring and promotion decisions would be unfair even though only 20% thought that the adaptive testing nature of the test itself was unfair (Gutierrez 2011b).

Internet-based testing (IBT) builds upon CBT by allowing assessments to be administered online with or without a proctor. Unproctored Internet testing allows candidates to choose where and when to complete selection assessments, a characteristic shown to elicit favorable user-friendliness perceptions, procedural justice perceptions and intentions toward the organization (Giumetti et al. 2010). In addition to ease of administration, organizations benefit from increased efficiency as online testing provides for more accurate tracking of data, decreases the time spent scheduling on-site visits for candidate testing, and permits candidates to interact with online customer service functions that offer real-time resolution of technological issues or questions (Anderson 2003; Bauer et al. 2006; Tippins 2009). In general, candidate reactions to elements of Internet-based psychological testing have been consistently positive (Broderson and Murphy 2010; Reynolds et al. 2000).

Despite its benefits, IBT has initiated concerns regarding the security and use of personal information provided in online applications and assessments. Within the IBT context, candidate idiosyncrasies relating to concern over the privacy of personal information submitted online hold consequences for candidate fairness perceptions (Eddy et al. 1999). Further, research demonstrates that concern for privacy is related to important candidate and organizational outcomes, through its strong relationship with procedural justice factors (Bauer et al. 2006; Sanderson et al. 2012; Wrenn et al. 2011). As such, IBT elicits higher test taking motivation, greater organizational attraction, and more favorable intentions towards the organization from candidates with low concern for privacy (Bauer et al. 2006; Sanderson et al. 2012). Further, candidates with low concern for privacy tend to believe that online tests are more resistant to faking and that test experiences across candidate administrations are consistent and equitable (Wrenn et al. 2011).

Innovations in CBT have allowed for the inclusion of media in item stems, response options or both. Several types of media have been utilized in simulations and SJTs to illustrate job-relevant situations, including 2D animation (cartoon characters), 3D animation (avatar characters), and video (live-action, human characters). Multimedia components are particularly useful in capturing rich, ambient detail of scenarios that are typically lost in text-based versions of the same content. Further, multimedia simulations promote candidate engagement and perceptions of job relatedness by helping candidates to visualize the problem or situation they are being asked to evaluate.

As depicted in Fig. 3.1, each media type can be applied to paper-and-pencil tests, CBTs or IBTs. For example, a candidate could watch a video of a customer interaction and indicate responses via paper-and-pencil, computer, or Internet. Though CBT and IBT offer a variety of benefits, research suggests that candidate reactions to multimedia assessments are generally more favorable than reactions to computerized text-based and paper-and-pencil simulations (Richman-Hirsch et al. 2000). As such, computerized text-based tests typically serve as a point of comparison, used to demonstrate the advantages of multimedia simulations of the same content. The remainder of this chapter focuses on the benefits of CBT and IBT multimedia simulations, from the candidate's perspective. We make an effort, where possible, to specify the test properties in relation to the medium and media used, although researchers have not consistently reported this information.

3.1.2 Candidate Reactions to Multimedia Simulations

A considerable amount of research has been devoted to substantiating the use of multimedia simulations covering various topics such as format, scoring, validity, and adverse impact. Candidate reactions research has been particularly fruitful, showing that on average, candidates prefer assessments that include multimedia components (e.g., Bryant and Malsy 2012; Chan and Schmitt 1997; Drew et al. 2012b; Gutierrez 2010; Motowidlo et al. 1990; Richman-Hirsch et al. 2000; Wrenn et al. 2011). The literature points to five important features of candidate reactions: (1) candidate perceptions of procedural justice; (2) candidate perceptions of distributive justice; (3) organizational outcomes of candidate reactions; (4) individual differences in candidate reactions; and (5) candidate preference for media types.

3.1.2.1 Candidate Perceptions of Procedural Justice

Multimedia simulations provide candidates with a more realistic picture of on-the-job tasks, which mimic feelings of work pressure associated with work performance. Further, multimedia assessments engage candidates in more job-relevant social interactions, providing additional content and information about the job (realistic job preview) and increasing the appearance that the test is measuring what it is purported to measure. In fact, candidates perceive computerized video-based tests to be more content valid and more likely to predict future job performance than paper-and-pencil and computerized text-based formats (Chan and Schmitt 1997; Motowidlo et al. 1990; Richman-Hirsch et al. 2000). For example, an interactive managerial simulation that included both audio and video components elicited higher perceptions of face validity than a text-based version of the same content. Similarly, significant gains in job relatedness perceptions have been shown for an online managerial assessment center including video components as compared to online text-based options (Bryant and Malsy 2012). Online avatar-based SJTs have also received higher job relatedness ratings than online text-based versions of the same content (Drew et al. 2012b; Wrenn et al. 2011).

While perceptions of job relatedness are dependent on the referent job, opportunity to perform represents perceptions that are specific to the individual candidate. This distinction is important because the two procedural justice factors are not mutually exclusive. For example, a candidate may be aware that a work sample test was relevant to a financial analyst position and still may feel that the selection process denied them the opportunity to demonstrate other critical KSAs such as knowledge of portfolio management or skill in budgeting. Feelings of such missed opportunities are often used as a justification for poor test performance, especially after experiencing unfavorable outcomes (Schleicher et al. 2006).

We cannot yet conclusively determine whether advances in technology help candidates to perceive assessments as giving them more opportunity to perform. When engaged in computer- or Internet-based assessments, candidates are more dependent on their abilities to operate computer functions and hardware, which may impact perceptions. For example, ability to control a computer mouse has been found to be more impactful on opportunity to perform perceptions when engaged in a point-and-click innovative item than a computerized-multiple choice item (Gutierrez 2011a). Opportunity to perform may also be affected by characteristics of IBT (Bauer et al.

2011). While candidates may feel that the freedom to choose where they take the test allows them to perform their best (Beaty et al. 2009), the perceived ease of cheating, connectivity issues, and speed may negatively impact opportunity to perform perceptions.

Due to the ambient detail and additional social interaction information provided in multimedia assessments, it would be reasonable to expect candidates to report greater ability to demonstrate KSAs. Research using a student sample, however, failed to demonstrate support for this claim. Drew et al. (2012b) found that opportunity to perform ratings of an online avatar-based SJT did not differ from those of an online text-based SJT. In a high-stakes situation where passing is a salient concern, differences in candidate perceptions of opportunity to perform across media types may be more detectable.

Multimedia assessments are typically just one of many types of assessments administered within a larger test battery. For example, a candidate may complete a biographical data questionnaire, personality test, and SJT in one sitting. Thus, it is important to know how multimedia components affect candidate perceptions of procedural fairness toward other assessments included within the same hiring procedures. Smither and colleagues (1993) proposed that candidate perceptions may have a spillover effect, in that favorable reactions generated from one assessment may engender favorable reactions to the hiring organization and other assessments within a test battery. By this definition, perhaps candidates would react more favorably to assessments that typically garner less favorable reactions (e.g., cognitive ability and personality tests) after taking an assessment that includes multimedia components. One study has supported this idea by demonstrating that candidate perceptions of their opportunity to perform and knowledge of results were higher for a personality and cognitive ability test, respectively, when these followed an online avatar-based SJT than a computerized text-based SJT version of the same content (Drew et al. 2012b).

3.1.2.2 Candidate Perceptions of Distributive Justice

Candidate perceptions of distributive justice are predominantly determined by outcome favorability (Gilliland 1993), which refers to information provided to the candidate about test performance (pass or fail) or job offers (hired or rejected). In general, candidates who fail assessments or are rejected for employment typically have more negative reactions (Hausknecht et al. 2004; Ryan and Ployhart 2000). Similar conclusions can be made with regards to multimedia simulations although caution is recommended given the extremely limited number of studies available. Bryant and Malsey (2012) reported that negative performance feedback significantly altered previously favorable job relatedness perceptions of a multimedia simulation. Further, Drew et al. (2012a) suggests that providing explanations for performance is critical when using innovative assessments, as candidates may be more affected by outcome favorability. Specifically, distributive justice was lower after candidates received information that they had failed an avatar-based SJT than after receiving failing information on a computerized text-based SJT of the same content.

3.1.2.3 Organizational Outcomes of Candidate Reactions

The selection process serves as an important source of information for candidates, and organizations want to know that their investments in assessment technology will result in favorable and measurable results (e.g., Murphy and Davidshofer 1998). Past research suggests that an organization's policies, practices, and procedures can be inferred from the selection process (French 1987), and may affect a candidate's desire to pursue further interactions with that company. When candidates perceive that selection procedures are fair and equitable they are likely to develop positive attitudes toward the hiring organization resulting in more favorable organizational outcomes (Gilliland 1993). In general, avatar-based simulations are more likely to promote greater intentions to accept an offer for employment and increase the favorability of overall company perceptions than text-based forms (Drew et al. 2012b; Drew et al. 2013; Wrenn et al. 2011). Video-based assessments would also be expected to garner similar reactions, although research is needed to support this claim.

3.1.2.4 Individual Differences in Candidate Reactions

Certain individuals may be more predisposed to react positively to selection instruments, according to: (1) demographic characteristics, (2) attitudes, and (3) experience with computers and the Internet. When considering technologically advanced simulations as selection tools, the inherent idiosyncrasies of the candidate pool must be taken into account in order to ensure favorable candidate reactions.

As seen in Chap. 2, demographic differences are a large concern for organizations and test developers, as subgroup differences have the potential to lead to adverse impact. Early investigations considered how the use of multimedia could decrease demographic differences in candidate reactions and performance. For example, Chan and Schmitt (1997) demonstrated that face validity and test performance was less associated with race when a SJT included video components. This is not surprising given that the use of media in assessments greatly impacts the capabilities candidates use to interpret information provided during testing. While paper-and-pencil and computerized text-based tests require reading, comprehension, and interpretation, candidates viewing scenarios depicted in video or animation generate responses by interpreting verbal and nonverbal cues. Since cognitive tests have the highest potential for subgroup differences, several authors have suggested decreasing cognitive components in assessments as a viable strategy to mitigate risk of adverse impact and increase legal defensibility (e.g., Ployhart and Holtz 2008; Ryan and Tippins 2004).

Recent investigations have considered how candidates attend to the race and gender of those they are "interacting with" while viewing multimedia simulation scenarios. Interestingly, the similarity in race between the candidate and scenario actors has been found to influence candidate perceptions of procedural justice, such that job relatedness and opportunity to perform perceptions were higher when the race of the candidate was similar to the race of the actor (Golubovich and Ryan 2012).

Further, test taker attitudes (e.g., motivation and belief in tests) that reflect a candidate's attitudinal and motivational disposition towards a referent assessment have been shown to influence procedural justice outcomes and intentions toward organizations in general selection contexts (Arvey et al. 1990; Hausknecht et al. 2004). However, very little is known about how advances in assessment technology will affect test taker attitudes. In regard to IBT, Bauer and colleagues (2006) report that test-taking motivation was significantly related to procedural justice factors. More recently, Drew et al. (2013) reported that an online avatar-based SJT promoted test taker motivation and belief in tests, which in turn, influenced subsequent perceptions of procedural justice, job relatedness, intentions toward the hiring organization, and overall company perceptions. Other researchers have considered more general candidate attitudes such as engagement, enjoyment, and satisfaction with the testing process. In general, candidates feel that multimedia assessments are more engaging (Tuzinski et al. 2012), more enjoyable and shorter in length than both paper-and-pencil and computerized formats (Richman-Hirsch et al. 2000). These positive attitudes lead to a more satisfactory experience for the test taker, ensuring favorable outcomes for examinees and organizations.

Lastly, experience with computers has been found to influence individual perceptions of the selection process (Wiechmann and Ryan 2003) and intentions toward the organization (Bauer et al. 2006) when using CBT and IBT, respectively. Differences in the salience of particular testing characteristics may be a function of candidate experience with the Internet and previous work experience, such that candidates with less Internet experience may be more liable to be negatively influenced by system speed or glitches. Specifically, Sinar et al. (2003) found that the perceived degree of efficiency and user-friendliness of an online application was more salient, and thus more impactful, on subsequent evaluations of the hiring organization for candidates with less Internet experience. Further evidence suggests that candidates with less job experience are more likely to attend to the technical characteristics of the test and functionality of the online application while also placing more emphasis on initial interactions with the employing organization (Sinar et al. 2003). Investigations considering differences in reactions to multimedia simulations, however, have failed to replicate these findings with individual differences in Internet familiarity, as well as time spent playing video games (Tuzinski et al. 2012). Nevertheless, a high amount of anxiety associated with computer use has been found to negatively impact face validity perceptions of a video-based SJT (Oostrom et al. 2010).

3.1.2.5 Candidate Preference for Media Types

It can be concluded that, in general, candidates prefer media-rich simulations over text-based assessments, but what about differences between video and animation types? When creating a simulation that is going to present human characters—for example, a multimedia SJT taking place in a customer service environment—a choice that developers must make early on is whether to use live action or animation to depict those characters. From an implementation standpoint, video has the highest level of realism, while animation tends to have lower development and localization costs, shorter production times, and greater ease in customizing characters and backgrounds. The criteria that simulation developers might use to choose between these are explained in detail in Chap. 4, and are summarized in Table 3.1 below.

If animation is chosen over live action, there are further choices to be made. Animation, after all, presents a blank canvas to the creator, offering an infinite range of potential styles. As in Chap. 4, we can broadly classify animation into three broad categories. Table 3.2 summarizes the differences between three primary styles of animation.

The extant literature is only starting to gather evidence to inform how candidate reactions differ according to media type. Tuzinski and colleagues (2012) considered differences in candidate perceptions and preference for three types of multimedia SJT formats (video, 2D, and 3D animation) and a text-based version of the same content.

Table 3.1 Criteria for selection of live action or animation

Production costs and time	For simple single location simulations, the costs of creating live action versus animated scenes are comparable. As the simulation becomes more complex, relying on multiple locations, characters, and events, then the costs of live action increases at a greater rate than animated footage
Practicality	Some simulations require footage of characters in specific situations that are not feasibly recreated with live action video
Flexibility	One disadvantage of video over animation is that any subsequent changes to the script, or recreating in another language, will require a potentially expensive reshoot, whereas animation can be changed relatively quickly
Aesthetics and brand	Sometimes the choice of live action versus animation comes down to the 'look and feel' that is desired by the user of the simulation and whether it aligns with the organization's brand

Table 3.2 Styles of animation

Style		Description
Two dimensional (2D)		Characterized by flat characters, viewed from only a limited number of angles. Examples: *South Park* and *The Simpsons*
Three dimensional (3D)	Caricatured	Emphasizes and accentuates the characters' features, making them appear less-lifelike. Example: *The Incredibles*
	Realistic	Seeks to make the characters look as realistic as possible. Examples: *Polar Express* and *Mars Needs Moms*

In general, video was the most preferred media type, followed by 3D animation, 2D animation, and text. Video was perceived to be more job related and engaging than all other forms, while 3D animation was viewed as more job relevant than 2D animation and more engaging than text. This pattern of candidate preferences for media types suggests that determinations may be related to how media differentially influences the representation, movement, and realism of characters depicted in simulation scenarios. If we can understand *why* candidates prefer some media types over others, perhaps advancements can be made to increase the efficacy of animation to elicit favorable candidate reactions.

3.2 The Influence of Technology on User Acceptability: The Uncanny Valley

About a decade ago, when simulation developers began to experiment with a variety of 3D animation techniques, they noticed that some of the 3D characters that were designed to be more realistic-looking evoked in some viewers a feeling of unease or eeriness. This reaction was specific to 3D realistic characters, and rarely seen with 3D caricature or 2D cartoon-like characters. Subsequent research into the art and aesthetics of animation led us to the concept of the uncanny valley proposed by Japanese robotics professor, Masahiro Mori. Mori (1970) contends that viewers feel more warmth or empathy with characters that are, up to a point, human-like. However, once that point is reached (when characters appear 80–85% human-like) viewers become more aware of its imperfections or not-quite-humanness and the creepier it appears. Beyond this point, when the character becomes even more human-like, the viewer once again feels warmth and empathy towards the character (see Fig. 3.2).

Researchers have established that viewers have similar negative reactions towards certain styles of animated humans in animation, video games, and simulations

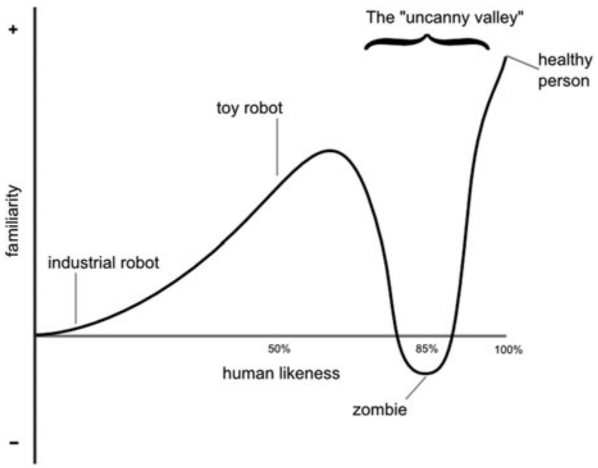


Fig. 3.2 The Uncanny Valley. (Adapted from: Mori 1970)

(Hawkes 2012a; MacDorman et al. 2009; Tinwell et al. 2011). Two commonly cited examples of the uncanny valley in the popular media are the motion pictures *Polar Express* and *Mars Needs Moms* (Nakashima 2011; Pavlus 2012). While the uncanny concept is well established, it is less clear whether the shape of the valley is as Mori (1970) described. Some have replicated the distinctive shape of the valley as depicted earlier (Flach et al. 2012), but others have found the relationship between humanness and familiarity to be less well-defined (Tinwell and Grimshaw 2009; Tinwell et al. 2011).

It is apparent that some styles of animation elicit negative reactions from viewers, but why does this happen? Some researchers have sought to explain why uncanny robots and animations evoke this negative reaction. MacDorman and Ishiguro (2006), for example, state that an uncanny character reminds viewers of death and therefore "elicits culturally-supported defense responses for coping with death's inevitability" (p. 297). Rozin and Fallon (1987) and Moosa and Ud-Dean (2010) suggest that our reaction to uncanny characters stems from an evolved instinct to avoid danger or disease. Tinwell (as cited in Parr 2012) looked at subtle facial characteristics and found that unrealistic eyelid and eyebrow movements led viewers to suspect that the character had aggressive tendencies, leading to possible perceptions of a psychopathic propensity. What does this tell us about candidate reactions to animated characters in assessment simulations? Two questions arise from this. First, could uncanny characters in simulations create negative candidate reactions? And secondly, could candidate reactions to uncanny characters actually affect candidate responses?

Research into the uncanny in the context of simulations and assessment is scant, but studies on video games and multimedia learning suggest that they can impact candidate reactions. Tinwell (2009) presented subjects with a range of animated characters along with a live action video. Interestingly, while the live action video was rated as most human-like, it was not rated as the most familiar. Characters that were rated as more enjoyable to interact with within a video game were more frequently described by words such as "happy," "excited," and "fondness" while those with the lowest ratings were described by the words "irritated," "fear," and "confused". Extrapolating these findings to media-rich assessments suggests that the choice of animation style could well elicit positive or negative feelings from candidates. Where a choice of animation style is available, test developers should consider trialing different styles with the target audience in order to appraise their reactions to them.

However, while negative candidate reactions are clearly an undesirable outcome, impairing the validity of the test is an even more serious concern. For example, what if a candidate's reaction to the animated characters distorts how they respond to an assessment item? As an example, consider a multimedia SJT item that presents candidates with an animation recreation of a customer service interaction. Perhaps the correct response to this item depends upon the candidate demonstrating empathy towards the animated customer. However, if the animated customer is perceived as eerie or uncanny, then might the candidate be less likely to choose a response that demonstrates the desired emotion?

In one of the first studies relevant to this question, MacDorman et al. (2010) presented subjects with live action and animated reconstructions (3D) of the same ethical dilemma in which the character, Kelly, explained why her husband should not be told that she had tested positive for a sexually transmitted disease. Subjects were then asked how they would respond, and were given the choice of agreeing to or refusing Kelly's request. Further, both live action and 3D animation were presented as either fluid (i.e., smooth-moving) or jerky. The findings support a difference in how subjects responded to a live action versus an animated character. Male subjects were more lenient towards the request in the live action and smoothly animated conditions. This difference was not apparent in female subjects.

The implications of this work are not just that the choice of animation or live action may have an effect upon candidates' responses to test items, but that there may be gender differences in the size of this effect. If this finding were to be replicated in a multimedia situational judgment or cognitive item, then we may have inadvertently stumbled on a previous unconsidered form of adverse impact. To investigate this possibility, Hawkes (2012a) presented multimedia SJT and memory items to subjects. Each item was presented in four ways, including live action, 2D, 3D caricatured, and a 3D realistic style deliberately designed to be perceived as uncanny (see Fig. 3.3).

While the 3D realistic style of animation was rated as more eerie, unattractive, and less human-like than live action, it did not significantly affect subjects' responses to cognitive and noncognitive SJT items. The same research also failed to replicate significant gender differences in responses to different styles of animated characters.

Fig. 3.3 Images of a single scenario using multiple media formats

Follow-up research (Hawkes 2012b) sought to establish whether presentational style could interfere with the measurement of a specific trait, namely empathy, within the setting of a multimedia SJT designed to assess customer service skills. As before, while subjects were sensitive to the uncanny traits of the realistic 3D animation, response options were not affected by style of presentation.

3.3 The Influence of Media-Rich Assessments on Candidate Reactions: Future Considerations for Scientists and Practitioners

The literature on candidate reactions consistently tells us that candidates perceive work samples or simulations to be more fair than paper-and-pencil tests (e.g., Robertson and Kandola 1982; Schmidt et al. 1977) and that some tests are viewed as being more job related than others (Huffcut 1990). Nonetheless, studies focused primarily on the use of technologically advanced assessments are still in their infancy. In fact, many of the findings cited in this chapter are based on unpublished work and conference presentations. As has been noted by previous reviews on candidate reactions (see Chan and Schmitt 2004; Ryan and Ployhart 2000) we continue to see measurement issues among studies focused on simulation use. For example, the large majority of them rely on student samples. In some cases, this potential weakness was mitigated by using entry level retail simulations with students who either have an interest in pursuing retail jobs or previous retail experience. We also uncovered a significant need to distinguish between mediums of presentation and media type. Although distinctions between paper-and-pencil and computer- and Internet-based testing have been made with regards to technology's impact on assessments, the literature on candidate reactions continues to confound medium of assessment delivery and media type used. It is important for both practitioners and researchers to consider the implication of these two elements on their findings given that each presents unique concerns.

We recognize the significant value of existing candidate reaction models in advancing this field. Yet, we consider it a timely need to revise these models in accordance with current technological advances. We agree that issues related to concern for privacy and measurement equivalence of simulated versus paper-and-pencil tests are critical, however, we have raised several novel issues for consideration. It is apparent that different styles of characters—animated or live action—are capable of evoking different reactions, both positive and negative, from their viewers. Test designers should bear this in mind when selecting a presentational style for character-based simulations. Secondly, the question of whether candidates' reactions to animated characters could influence how they respond to test items has not been fully explored. Taking heed of the findings of MacDorman et al. (2010) it may be wise to err on the side of caution by ensuring that test items that use animated characters are adequately trialed to reduce the risk of inadvertently introducing adverse impact or otherwise distorting a candidate's responses. Research is warranted to further explore the influence of various media types on candidate preferences, particularly given the practical considerations that may influence organizations or testing companies to choose among them.

Further, the changing nature of the workplace and the increased reliance on technology has had a notable influence in the way business functions are conducted. For example, employers are finding ways to adapt their business and human resource models to meet the needs of the growing segment of Millennials entering the workforce. The increased reliance on technology for both work and social usage is a key differentiator between Millennials and previous generations (Pew Research Center 2010). This raises interesting questions regarding the use of *gamified* assessments and generational differences in response to the use of media-rich simulations.

Can we predict the way in which candidates react to serious games in employee selection? Laumer et al. (2012) found that candidates value fairness, ease-of-use, and usefulness more than they value enjoyment although the findings also support the notion that an enjoyable simulation is likely to be rated as easier to use. From a practical perspective, designers of gamified assessments can draw upon the experience of other serious games designers in measuring the reactions of candidates. For example, Moreno-Ger et al. (2012) proposed a tool, the Serious Game Usability Evaluator (SeGUE), which measures system and user related reactions by having users voice their feedback as they experience the simulation. Such a technique gives test designers useful insight into candidate perceptions and may be used to assess the potential influence of reactions on the validity of these gamified assessments. Clearly, a significant amount of research remains to be done in this area.

Exploring generational preferences for media format in simulations is also timely and novel. To our knowledge, the only study that has investigated this question is Tuzinski et al. (2012) who examined the level of engagement for various media types, including 2D, 3D, and video. Consistent with the technology preferences of Millennials, younger workers preferred the use of 3D media in simulations over other formats (Tuzinski et al. 2012). Here again, we find that the quickly advancing use of technology in the development of simulations requires us to think beyond what has been a traditional focus of candidate reaction research. Both practitioners and researchers need a clearer understanding of generational trends in the preference for and attitudes towards simulated technology.

Our review of this growing and trending area of candidate reaction has left us with more questions than answers. Changes in technology have impacted the relationships that candidates have with potential employers by creating social connections that may influence even consumer behavior. Consequently, the selection process has new implications for measuring other key organizational outcomes in relation to candidate reactions, such as the endorsement of companies through social media outlets. We urge researchers and practitioners alike to take a multidisciplinary approach to studying candidate reactions to technologically advanced simulations by incorporating findings from other areas, as we have done here with the uncanny valley, and to pursue the publication of research that can help advance theoretical frameworks.

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,* 121–136.

Arvey, R. D., & Sackett, P. R. (1993). Fairness in selection: Current developments and perspectives. In Schmitt, N., & Borman, W. (Eds.), *Personnel selection in organizations* (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.

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.

Bauer, T. N., Truxillo, D. M., Mack, K., & Costa, A. B. (2011). Applicant reactions to technology-based selection: What we know so far. In Tippins, N. T., Adler, S., & Kraut, A. I. (Eds.), *Technology enhanced assessment of talent* (pp. 190–223). San Francisco: Jossey Bass.

Beaty, J. C., Dawsom, 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,* 58–63.

Broderson, D. A., &, Murphy., C. (2010, April). *Applicant perceptions of online assessment and the companies using them.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Bryant, S. E., &, Malsey., S. (2012, April). *21st century assessment centers: Technology's increasing role and impact.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Cascio, W. F., & Phillips, N. F. (1979). Performance testing: A rose among thorns? *Personnel Psychology, 32,* 751–766.

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. (2004). An agenda for future research on applicant reactions to selection procedures: A construct-oriented approach. International Journal of Selection and Assessment, *12*, 9–23.

Drew, E. N., Lamer, J. J., Bruk-Lee, V., LeVine, P. J., & Wrenn, K. A. (2012a). Applicant reactions to multimedia simulations. Unpublished data.

Drew, E. N., Lamer, J. J., Bruk-Lee, V., LeVine, P. J., & Wrenn, K. A. (2012b, April). *Keeping up with the Joneses: Applicant reactions to multimedia SJTs*. Poster presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Drew, E. N., Bruk-Lee, V., Wrenn, K., & Levine., K. (2013, April). *Test taker dispositions in response to a multimedia SJT*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Houston, TX.

Eddy, E. R., Stone, D. L., & Stone-Romero, E. F. (1999). The effects of information management policies on reactions to human resource information systems: An integration of privacy and procedural justice perspectives. *Personnel Psychology, 52,* 335–358.

Flach, L. M., de Moura, R. H., Musse, S. R., Dill, V., Pinho, M. S., & Lykawka, C. (2012). Evaluation of the uncanny valley in CG characters. Proceedings of SBGames 2012. http://sbgames.org/sbgames2012/proceedings/papers/computacao/comp-full_14.pdf. Accessed 12 Feb 2013.

French, W. L. (1987). *The personnel management process*. Boston: Houghton Mifflin.

Gilliland, S. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18,* 694–734.

Giumetti, G. W., Wasko, L. E., & Sinar, E. F. (2010, April). *Mediated pathways linking internet testing features and applicant reactions*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Golubovich, J., & Ryan, A. M. (2012, April). *Demographic cues in video-based situational judgment tests*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Gutierrez, S. L. (2010, April). *Comparing examine reactions to multimedia and text-based simulation items*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Gutierrez, S. L. (2011a, February). *Moving beyond multiple-choice items: Examining the technological considerations and examinee reaction to a new point and click innovative item format*. Poster presented at the annual conference of the Association of Test Publishers.

Gutierrez, S. L. (2011b, April). Perceptions of fairness and opportunity to perform on CAT in personnel selection. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Hausknecht, J. P., Day, D. D., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57,* 639–683.

Hawkes, B. J. (2012a). *Multimedia situational judgment tests: Are animation and live action really equivalent?* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Hawkes, B. J. (2012b). *Test-takers' empathy for animated humans in SJTs*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Huffcut, A. (1990). Intelligence is not a panacea in personnel selection. *The Industrial Organizational Psychologist, 27,* 66–67.

Laumer, S., Eckhardt, A., & Weitzel, T. (2012). Online gaming to find a new job? Examining job seekers? Intention to use serious games as a self-assessment tool. *German Journal of Research in Human Resource Management*, *26,* 218–240.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies, 7,* 297–337.

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. (2009). Too real for comfort: Uncanny responses to computer generated faces. *Computers in Human Behavior, 25,* 695–710.

MacDorman, K. F., Coram, J. A., Ho, C.-C., & Patel, H. (2010). Gender differences in the impact of presentational factors in human character animation on decisions in ethical dilemmas. *Presence: Teleoperators and Virtual Environments, 19,* 213–229.

Moosa, M., & Ud-Dean, S. M. (2010). Danger avoidance: An evolutionary explanation of the uncanny valley. *Biological Theory, 5,* 12–14.

Moreno-Ger, P., Torrente, J., Hseih, Y. G., & Lester, W. T. (2012). Usability testing for serious games: Making informed design decisions with user data. *Advances in Human-Computer Interaction*, *2012,* 1–13.

Mori, M. (1970). Bukimi no tani [The Uncanny Valley]. *Energy, 7*(4)*,* 33–35.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). Alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

Murphy, K., & Davidshofer, C. (1998). Psychological testing: Principles and applications (4th ed.). Englewood Cliffs, New York: Prentice Hall.

Nakashima, R. (2011). Mars needs moms: Animation is too real, too creepy. http://www.smh.com.au/technology/technology-news/too-real-means-too-creepy-in-new-disney-animation-20110404-1cyt8.html. Accessed 16 Feb 2013.

Oostrom, J. K., Born, M. P., Serile, A. W., & van der Molen, H. T. (2010). Effects of individual differences on the perceived job relatedness of a cognitive ability test and a multimedia situational judgment test. *International Journal of Selection and Assessment, 18,* 394–406.

Parr, C. (2012). Unheimlich manoeuvres in the race to make CGI real. http://www.timeshighereducation.co.uk/story.asp?storycode=421455. Accessed 4 Feb 2013.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (2010). Innovative items for computerized testing. In can der Linden, W. J., & Glas, C. A. W. (Eds.), *Elements of adaptive testing* (pp. 215–230). New York: Springer.

Pavlus, J. (2012). Did the "Uncanny Valley" kill Disney's CGI company? http://www.fastcodesign.com/1663530/did-the-uncanny-valley-kill-disneys-cgi-company. Accessed 19 Feb 2013.

Pew Research Center (2010). Millenials: A portrait of generation next. http://www.pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf. Accessed 14 Feb 2013.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex sub group differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172.

Pommerich, M., & Burden., T. (2000, April). *From simulation to application: Examinees react to computerized testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Reynolds, D. H., Sinar, E. F., & McClough, A. C. (2000). *Evaluation of an internet-based selection procedure*. Paper presented at the annual meeting of the Society of Industrial Organizational Psychology, New Orleans, LA.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85,* 880–887.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. *Journal of Occupational Psychology, 55,* 171–183.

Rozin, P., & Fallon, A. E. (1987). A perspective on disgust. *Psychological Review, 94,* 23–41.

Ryan, A. M., & Ployhart, R. E. (2000). Applicant perception of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26,* 565–606.

Ryan, A. M., & Tippins, N. T. (2004). Attracting and selecting: What psychological research tells us. *Human Resource Management, 43,* 305–318.

Sanderson, K., Drew, E., Bruk-Lee, V., LeVine, P. J., & Wrenn, K. A. (2012, April). *For your eyes only? Reactions to internet based multimedia SJTs*. Poster presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Schleicher, D. J., Venkataramani, V., 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.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30,* 187–197.

Schuler, H. (1993). Social validity of selection situations: A concept and some empirical results. In Schuler, H., Fair, J. L., & Smith, M. (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 11–26). NJ: Erlbaum.

Sinar, E. F., Reynolds, D. H., & Paquet, S. L. (2003). Nothing but net? Corporate image and web-based testing. *International Journal of Selection and Assessment, 11,* 150–157.

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.

Tinwell, A. (2009). The Uncanny as usability obstacle. In A. A. Ozok, & P. Zaphiris (Eds.), *Proceedings of the HCI International 2009: Online Communities and Social Computing Workshop* (pp. 622–631). San Diego: Springer.

Tinwell, A., & Grimshaw, M. (2009). Bridging the uncanny: an impossible traverse? In O. Sotamaa, & A. Lugmayr, H. Franssila, P. Näränen, & J. Vanhala (Eds.), *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era* (pp. 66–73). Tampere: ACM.

Tinwell, A., Grimshaw, M., & Williams, A. (2011). The uncanny wall. *International Journal of Arts and Technology, 4,* 326–341.

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.
Tuzinski, K., Drew, E. N., Bruk-Lee, V., & Fetzer, M. (2012, April). *Applicant perceptions of multimedia situational judgment tests*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.
Viswesvaran, C. (2003). Introduction to special issue: Role of technology in shaping the future of staffing and assessment. *International Journal of Selection and Assessment, 11,* 107–111.
Wendt, A., Harmes, J. C., Wise, S. L., & Jones, A. T. (2008). *Development and evaluation of innovative test items for a computerized nursing licensure exam*. Paper presented at the annual meeting of American Educational Research Association, New York, NY.
Wiechmann, D., & Ryan, A. M. (2003). Reactions to computerized testing in selection contexts. *International Journal of Selection and Assessment, 11,* 215–229.
Wrenn, K., Drew, E., Buxo, N., Bruk-Lee, V., & Levine, P. J. (2011, April). *Applicant perceptions of multimedia situational judgment tests*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL

Chapter 4
Simulation Technologies

Ben Hawkes

4.1 Introduction

An exploration of the technologies employed in creating and deploying assessment simulations can only cover a subset of this expansive topic. Despite this, it is hoped that this chapter will enable the reader to understand some of the technologies involved in developing and deploying assessment simulations, the advantages and disadvantages of such technologies, and ultimately to make informed decisions about simulation development, regardless of the reader's role in that process.

The focus of this chapter is on assessment simulations that fulfill three criteria:

- Can be deployed to candidates in both *proctored and unproctored* settings. Deploying simulation assessments to candidates via the Internet imposes certain constraints on designers which are explored further on.
- *Capture candidates' responses*. Many simulations react to the user's input without capturing and storing that input. For assessment for selection, however, we need to be able to capture a candidate's response in order to score it or compare it to the responses of others.
- *Report those responses back* to a database such as a learning management system (LMS) or applicant tracking system (ATS). When used for online assessment, simulations are typically delivered from a platform that initiates the simulation, and then captures and stores the candidate's responses and scores. This may be a stand-alone assessment platform, or an ATS or an LMS.

Care has been taken to avoid technical jargon, or, where it is necessary for the understanding of assessment technologies, to explain it. For the purposes of this chapter, the starting point is the "script": the text of the assessment, perhaps created by an industrial–organizational (I–O) psychologist or other assessment professional, already trialed. The process of creating simulations has been segmented into three stages: content creation (creating the components of the simulation), authoring (tying

B. Hawkes (✉)
Kenexa, an IBM Company, London, UK
e-mail: ben.hawkes@uk.ibm.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_4,

Table 4.1 Types of simulation and their applications

Type of simulation	Example applications
Character-based simulation	Multimedia situational judgment test (SJT), presenting candidates with short scenes of customer interactions and asking them to select the best and worst response from a list of four or five options
Desktop simulation	Call-center simulation, presenting users with a facsimile of a call-center operator's screen. Electronic inbox or "e-tray" exercise, presenting candidates with documents and data to be analyzed and acted upon
Virtual environment (VE)-based simulation	Users interact with and manipulate representations of three-dimensional (3D) objects, from individual objects to complex machines all the way to full virtual environments

all of those components together), and deployment (putting the simulation online and connecting it to other online services such as an ATS). These three stages are not mutually exclusive: although some tools have a role in only one stage of the process, many of them span two or even all three stages.

Throughout the chapter, reference is made to various technologies, software, and websites. These are not intended as endorsements: the technologies available to simulation creators and users are varied, and those that are referenced in this chapter are provided as exemplars. Many of the more commonly seen technologies are listed in Sect. 4.5.

This chapter will focus on three common types of assessment simulation, outlined in Table 4.1.

4.2 Content Creation

In this context, "content" refers to everything that the candidate will see and hear during the course of the simulation, sometimes referred to as the collateral or assets. These assets could be text, images, animation, video, or audio.

To create images, simulation developers have several options. The creation of image content may be as simple as commissioning a series of photographs or buying suitable images from a stock library, or images may be created by a graphic designer in order to create the user interface. The remainder of this section focuses on the creation of other forms of multimedia: animation, video, and audio.

Fig. 4.1 Example of two-dimensional animation

4.2.1 Character-Based Simulations

To design character-based simulations, designers are faced early on with the choice of using "live action" (i.e., video of real people) versus animation. So what factors should designers consider in making this choice between these two media?

4.2.1.1 Animation

Originally, animation was a painstaking process: each second of footage would require 25 individually drawn images, to be shown one after another. Nowadays, through the assistance of animation software, such efforts are not necessary. At the very least the software streamlines the workflow, but in many instances it can create much of the animation automatically.

At the high-end of animation, of the quality shown in Disney Pixar movies, animators still use a painstaking approach. That approach comes at a high cost: the estimated budget for the film *Up* was $ 175 million (IMDB 2012). At 96 min long, that suggests a rate of $ 1.8 million per minute, or around $ 30,000 per second of animation!

However, most simulation designers are unlikely to have the budget of Disney Pixar. Instead, footage is commissioned from animators using software such as Autodesk's Maya and 3ds Max which have become almost "industry standard" for professional three-dimensional character animation, and Toon Boom, RETAS, and Toonz which have held a similar position amongst two-dimensional animation packages. Figures 4.1 and 4.2 show examples of two-dimensional and three-dimensional animation, respectively.

These programs provide a "blank canvas" with virtually no limitation to what can be created. To create a high-quality output from such software, however, requires extensive experience and knowledge, and this is reflected in the time and costs of creating animation. It is extremely difficult to give even a "ballpark" estimate on animation costs as it depends on the complexity of the scene, character movements,

Fig. 4.2 Example of three-dimensional animation

Fig. 4.3 Example simulation content produced using CodeBaby, an off-the-shelf animation tool

lip-synching, and so on. Hawkes (2012) supplied an identical brief to a selection of animation agencies in the USA, UK, and India. Their quoted costs ranged greatly, with the most expensive quote at ten times the cost of the least expensive.

Over the past 5 years, off-the-shelf character animation software designed for non-animators has become capable of producing high-quality outputs through a simple process of "drag and drop" editing. In the same way that presentations can be assembled by dragging and dropping shapes, text, and pictures, so animation can be created by dragging and dropping characters, backgrounds, and even movements. Figure 4.3 shows an example of the output of one tool, CodeBaby. In this example, off-the-shelf characters have been combined with a custom-built background environment.

However, this ease of use comes at a price. With these simpler-to-use tools, such as CodeBaby, Moviestorm, Xtranormal, and Muvizu, the designer's choices are restricted: characters, of a particular style, can only be customized to the extent permitted—or programmed—by the vendor.

However, a significant advantage of off-the-shelf animation tools is that the production time can be much shorter. One minute of animation might take just a few hours to create, whereas using professional animation software, one minute of animation might take days or even longer to create. There are various reasons for this. First, off-the-shelf tools often come stocked with a "library" of character movements and facial expressions that speed up the animation process. One of the more time-consuming stages in character animation is synchronizing mouth and lip movements to speech. Many of these off-the-shelf tools handle that automatically. Finally, such software packages often come with ready-to-use characters, environments, and objects, or at least offer the opportunity to download these for a cost from the vendor and third parties.

Between the consumer and professional packages lie other software such as Reallusion's iClone and DAZ Studio. Although not without their limitations, these offer much of the customization offered by professional software, but with greater ease-of-use.

Regardless of the style of animation that is selected, the characters must have a voice. Again, developers are faced with a choice: to use "real" people or to use computer generated speech, generally known as Text To Speech (TTS).

The quality of TTS has increased significantly over the past 10 years. Although it used to clearly sound machine-generated, now TTS vendors are able to create very clear human-like speech. However, developers of simulations are faced with a particular challenge when trying to embed emotional content into speech. While some TTS technologies allow text to be "marked up" to indicate where words and syllables should be stressed or delivered in a particular style, this still does not rival the expressive range of a human actor.

However, TTS still has a place for simulation developers. When prototyping a new simulation, it may not be cost-effective or practical to bring in voice actors, particularly if the spoken content of the simulation is likely to be changed frequently before its final incarnation. In this case, TTS can be used to create the spoken content rapidly and at low cost.

4.2.1.2 Video-Based Simulations

Video-based, live-action content offers several advantages over animated content. First, by capturing footage of real people, the simulation presents real behavior, including the subtle movements and expressions that we as observers rely on to understand the thoughts and feelings of others.

A second advantage comes with the potential speed of producing live-action footage. For certain types of material, the video production process can be much quicker than producing an equivalent animation. For example, a live-action version of a single two minute interaction taking place between two employees could be shot and edited in the space of several hours. To re-create this using animation, depending on the style of animation chosen, could take several days to produce. Animating the characters might only take a fraction of this time, with the remainder taken up by designing the characters and environment, and recording the voice audio for the animation to use.

However, live-action video does present some disadvantages. Firstly, although the time required to shoot simple interactions—such as the employee interaction mentioned—may be short, more complex interactions and situations can take significantly longer, when one factors in script learning, rehearsal, multiple camera angles, reshooting, and editing.

Secondly, animated sequences can more quickly and easily be updated than live-action footage, allowing rapid revision of existing test items to reflect changing psychometric (e.g., updated items), aesthetic (e.g., new branding), and localization needs (e.g., redubbing the animation into different languages). Making similar updates to live-action footage could well require extensive—and possibly expensive—reshoots.

A third disadvantage is that shooting video on location may present practical challenges. In the case of video being shot for a customer service simulation, if the video is being shot within a retail store then consideration must be given to planning the shoot to minimize the potential disruption that having a film crew may have on the day-to-day operations of that store. In other instances, it may not be practical to shoot a live-action version of the scenario. These instances might include the portrayal of unsafe working practices or catastrophic events: animation may be the only realistic option.

The fourth—and perhaps for many simulation developers, the most important—disadvantage of shooting live-action video is that the costs can become significantly higher than animation—perhaps to the extent that live action is not a feasible possibility. For example, a military simulation might require footage of many moving soldiers, ground vehicles, and aircraft. A live-action version of this might take months to plan and cost tens or even hundreds of thousands of dollars to shoot. The alternative of creating an animated version of this footage could be completed in shorter timescales and at a greatly reduced budget.

Of course, many simulations are less ambitious in their scope, and will require nothing like the budget and planning of a large-scale military simulation. Fox (2010) provides a list of 25 factors that impact the cost of corporate video production, including people (e.g., director, crew, actors, and editors), location, and equipment. Simulation developers can use a similar list to compare potential costs of animation versus live-action production.

4.2.2 Screen-Based and Desktop Simulations

Often, simulations need to re-create the working environment of a role that is either entirely PC-based, or are likely to use a PC much of the time—for example, call-center operators. In that case, the visual design of the simulation might simply be a re-creation of a screen interface.

In desktop simulations, the designer may choose to create a simple interface with a familiar graphical user interface (GUI) as one might see on a Windows or Mac "desktop." Figure 4.4 shows an example of this.

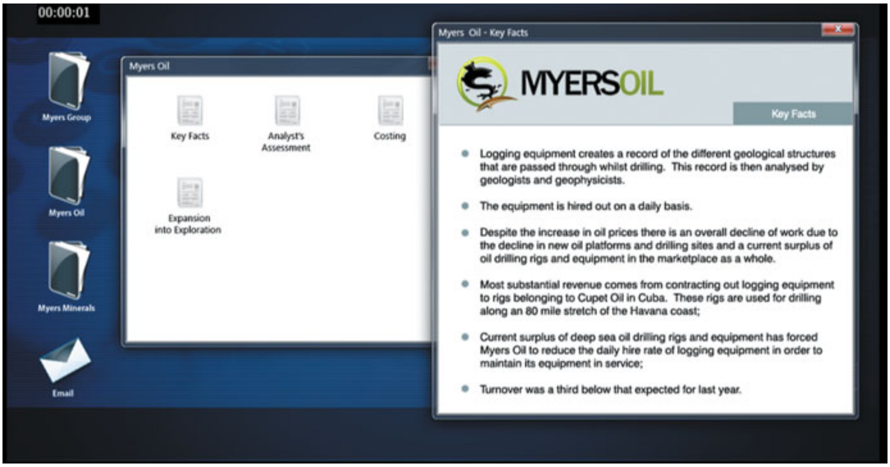
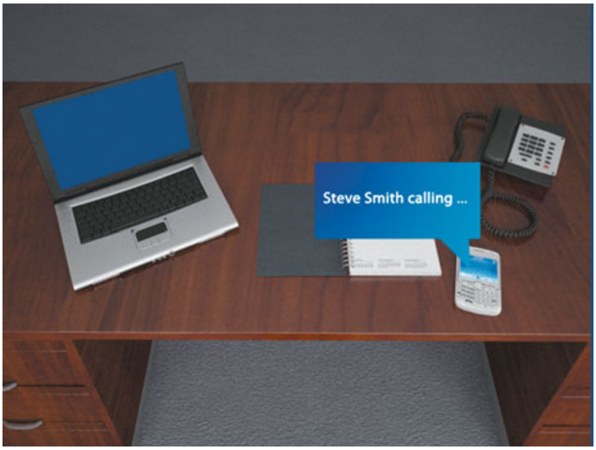


Fig. 4.4 Example of desktop simulation

Fig. 4.5 Example of a detailed desktop user interface



Alternatively, the desktop simulation may be more complex. Figure 4.5 shows an in-tray simulation taking place in a detailed three-dimensional office environment.

The choice of whether to use a fully rendered three-dimensional office, or simply to present a computer desktop may be as much to do with aesthetic choice as it is to do with budget. Regardless of the choice, care must be taken that the interface is easy to use. If users find the interface cumbersome or unfamiliar, then there may be a risk of introducing error into the assessment (Maxion and Reeder 2005).

4.2.3 Virtual Environment-Based Simulations

Virtual environment (VE)-based, or “virtual world” simulations present candidates with a re-creation of real-world objects, environments, and characters. They can

Fig. 4.6 Example of a virtual environment-based simulation assessment

often interact with these re-creations in realistic ways—for instance, disassembling a machine part or walking around a three-dimensional facsimile of a factory. Characters in the VE can be either entirely automated and computer-controlled, or they can be controlled by other people acting as virtual puppeteers (Fig. 4.6).

Although character-based simulation and desktop simulations may require modest animations and graphics, VE-based simulations often have far more complex requirements, for a three-dimensional virtual world needs to be populated with three-dimensional characters, environments, and props. It is the creation of these assets that often accounts for a large proportion of the development budget and time for VE-based simulations.

In many instances, however, these assets do not need to be created from scratch. Libraries of third-party characters, equipment, and other assets exist which can be used "off-the-shelf", or customized to meet the needs of the simulation. Typically, these assets are sold under a license, so care must be taken that the license permits the use for the asset that the simulation developer has in mind.

4.3 Authoring and Deployment

Authoring a simulation is the process by which the content (video, animation, audio, text, and images) is assembled into the finished product. It is similar in its objectives to the process of assembling video, animation, and other content into a presentation. The difference being that a simulation, as well as presenting information, also captures and stores candidates' responses.

Deployment of the simulation is the final stage of its development—the equivalent of letting it "into the wild." However, often this is the stage that simulation designers will consider first. This is done for a simple reason: the technology and methods for deploying the simulation will determine many of the choices that can be made during the content creation and authoring processes.

This section begins with an overview of some of the more commonly used tools to author and deploy assessment simulations before looking in detail at some of the challenges that are faced when we aim to deploy simulations in the context of online unproctored assessment.

4.3.1 Software for Authoring and Deployment

To maximize the reliability of any assessment—paper-based or computer-based—it is necessary that each candidate receives a standardized experience. That is not to say it should be identical: the principles of computer adaptive testing (CAT) are well established enough that we can be confident that candidates responding to different questions can still be considered to have taken the same test. Also, despite the multiple possible combinations of browser (Microsoft Internet Explorer, Google Chrome, Apple Safari, Mozilla Firefox, Opera, etc.) and operating system (Windows, Mac, Linux, Android, iOS), every year many millions of online assessments are delivered in a standardized and reliable form.

However, "traditional" assessments comprise text and graphics. Assessment simulations on the other hand may combine text, graphics, animation, video, and interactivity. So how do simulation creators ensure that the candidates' experiences are standardized?

Web pages are written using a language called Hypertext Markup Language (HTML). This instructs the browser software how to lay out the screen, what size and font to use for text, where to position images, and so on. However, the most common version of HTML in use—Version 4—does not itself provide the functionality that many simulations require. For example, it has limited capability to deliver video and audio, and it does not natively support animation.

To get around these limitations, multimedia and other rich Internet applications (RIA) use browser "plug-ins." These are small downloads that extend the capability of the browser and allow them to run software that has been created in other languages, not just HTML.

The plug-in that most Internet users will have encountered is Adobe Flash. Flash works within Microsoft Internet Explorer, Firefox, Safari, and other Internet browsers to deliver multimedia and interactive content. Over 99 % of PCs worldwide have Flash installed, although on mobile devices (phones, tablets) the percentage is substantially lower, around 50 % (Adobe Systems Incorporated 2012).

To understand the role of Flash and similar technologies, it might be useful to think back to a time when the World Wide Web was not supported by these technologies. As recently as 10 years ago, multiple conflicting standards for deploying animation, video, and even sound meant that the experience of using the Internet was far less seamless and transparent than today. Websites used a variety of standards and proprietary formats to deliver multimedia content. If the user's browser was unable to display the content, then they might be invited to download the appropriate plug-in.

As bandwidth increased and PCs became more powerful, a standard method of delivering rich, interactive content was needed, and Adobe (then Macromedia) Flash was launched (Macromedia 2002) and has since become the almost ubiquitous browser plug-in we see today.

Other plug-ins are commonly used to run simulations. Java, published by Oracle Corporation, and Silverlight, published by Microsoft, offer many capabilities to simulation developers that are similar to Flash. Lively discussions of the merits of Flash, Java, and related technologies can be found on the Internet (Brandt 2012; Draney 2012; Franco 2010; Le Grecs 2012), and simulation designers looking for guidance are unlikely to find a definitive answer. However, if Flash is ubiquitous and supported on so many platforms, why should simulation developers even consider other technologies for deployment?

One reason is that despite Flash being considered a “cross platform” language usable by different types of computers and operating systems not all mobile devices (phones and tablets) support Flash. In fact, no devices running Apple iOS—including iPad, iPhone, and iPod—run Flash within their browsers, nor are ever likely to (Jobs 2010). Support on other mobile devices—those running Android and Windows operating systems for example—is commonplace, but with Adobe’s 2011 announcement that Flash will no longer be developed for new mobile devices (Adobe Systems Incorporated 2011; Winokur 2011), it is likely that additional technologies will take the place of Flash as the most common method of delivering media-rich content, at least on mobile devices.

One of these technologies is the latest version of HTML, HTML5. Although the standard is not due to be finalized until 2014, most up-to-date PC, Mac and mobile-based browsers support it. The difference with this latest version is that animation, audio, and video are supported natively in the browser. In other words, the browser does not have to download a plug-in like Flash to deliver media-rich content. Fewer plug-ins mean fewer potential compatibility issues and, hopefully, fewer problems for the end-user.

However, compared to Flash, HTML5 has a limited set of features and capabilities. Graphics, audio, and video are well supported, but client-side applications—which are often needed to run simulations—receive inconsistent support from browsers (Deep Blue Sky 2012). Having said that, many simulations can be developed and run quite adequately in HTML5. However, more sophisticated simulations of the type often programmed in Flash may not be supported by HTML5.

There are ways of extending the functionality of HTML. One way is through the use of JavaScript. This is a programming language that runs within the browser and is capable of creating interactions from simple click and drag up to fully functioning, sophisticated word processors.

The choice of platform—Flash, Java, HTML, and so on—must be made very carefully. Designers should take into account the desired functionality and user experience for the simulation. In the interests of making the simulation as accessible as possible, designers should also consider whether users will have the necessary plug-in installed and whether that plug-in can even be installed at all. At the time of writing, in early 2013, simulation designers wishing to ensure that their assessments will run on iOS-based devices as well as older PC-based browsers may have to consider whether two versions of each simulation—one in Flash and one in HTML5—must be created to maximize the accessibility to the pool of candidates, and reduce the likelihood of candidates being unable to take the assessment. Over time, market penetration of devices and plug-ins change, so simulation designers should consult up-to-date statistics before selecting which technologies to use to create their assessments. Some sources for these statistics are listed at the end of this chapter.

4.3.1.1 Specialized Tools

Aside from the generic authoring tools such as Flash, there are many tools that are specifically designed for creating simulations for learning and development. Weiss (2012) counts over 140 of them. Many of those same tools can also create simulations suitable for assessment for selection.

These tools vary enormously in their capabilities, objectives, and look and feel. We have chosen a few examples of each style of simulation technology, but as iterated in the introduction to this chapter, their inclusion is in no way an endorsement of the product.

Many specialized simulation development tools offer the ability to export the finished simulation to Flash, Java, Silverlight, or many other formats discussed above. This enables the simulation to be deployed in the same way as an application that uses those formats. The advantage of using a specialized simulation authoring tool is that these technologies offer greater access to the non-programmer. Instead of writing lines of code, users can create simulations by dragging and dropping elements onto flowcharts. This presents two possible advantages: simulations can be built more quickly, and the development process is often more accessible to non-technical developers.

4.3.1.2 Linear Versus Branching Simulations

At their simplest, a simulation may be entirely linear: a video plays, question one appears, a second video plays, question two appears, and so on. Multimedia SJTs often use this format. This is a format that even simple simulation authoring software can produce.

Other simulations are nonlinear in that the candidate's responses are used by the simulation to determine which part they see next, almost like a "Choose Your Own Adventure" book. For example, in a sales simulation, the candidate could be given the choice of (a) asking more questions of the simulated customer, or (b) to present recommendations to the customer. The customer would then react realistically according to the candidate's choice, and the candidate is presented with further choices to make.

A simulation that offers branching after *every* stimulus offered to the candidate would rapidly become unwieldy. For example, if the candidate watches one video and is given three choices, each one leading to another video which in turn leads to three more choices, then for the fifth stage some 243 separate videos would be required. Instead, most branching simulations avoid this exponential growth by limiting the number of decision points or by limiting the number of options at each decision point.

Some authoring tools—for example Articulate Storyline and ClicFlic—provide a graphical interface with which to design such branching simulations. This interface often takes the form of a flowchart and the branching can be designed visually rather than using programming code. Using a graphical interface can make the simulation authoring process more accessible to nonprogrammers and reduce development time.

A branching simulation presents certain psychometric challenges: with its potential of thousands of different permutations of question order and candidate experience, the trialing and validation process needs to be carefully considered. This is not an issue that will be discussed here, suffice to say that any assessment, branching simulations included, should be trialed and validated to ensure that its use is appropriate and defensible.

4.3.2 Challenges Faced in Deploying Assessment Simulations

The use of simulations to deliver remote, unproctored assessments presents a series of challenges compared to the use of simulations in a proctored and controlled setting. In a controlled setting, administrators have control over the hardware, software, and testing environment. They also can offer direct and prompt IT support to candidates. Compare this to unproctored simulations which are being accessed by untrained users of unknown ability through a variety of browser software running on various operating systems and hardware, via Internet connections of varying quality, speed, and bandwidth.

Let us consider each of those constraints in turn, as each of them has an impact upon the design of assessment simulations.

4.3.2.1 Accessibility and Disability

We must be careful to consider the needs of candidates with disabilities. A visually rich simulation may well offer the sighted candidate an immersive, engaging experience, but unless suitable accommodation is made, there is a risk that candidates with visual impairments are disadvantaged. Equally, another simulation might make use of videos of conversations between characters, but without suitable accommodation, candidates with hearing impairments could be disadvantaged. So while simulations may well offer increased accessibility, creators must be careful to ensure that all candidates are assessed equally, regardless of computer literacy or disability.

In designing any assessment, simulations included, we must consider its accessibility. Commonly considered to refer to candidates with disabilities, accessibility also is relevant to other candidate groups, including older people and candidates with language impairment, or simply candidates who rarely, if ever, use a computer. Legislation often provides assessment publishers and users with a set of requirements that must be met. However, these requirements—and their very existence—are not universal, varying between countries and regions, and sometimes only providing the briefest of guidance. The Web Accessibility Initiative (WAI) of the World Wide Web Consortium (W3C) provides extensive guidelines and support materials to help designers understand and implement web accessibility (W3C 2012), and these are strongly recommended to designers. Simulation developers and users may also want to refer to the "Section 508" guidelines (United States Government 2011). Section 508 requires Federal agencies to take into the account the needs of all end-users when implementing electronic or other forms of information technology. Although targeted toward US Federal agencies, the guidelines may prove useful for any developer or user of simulations and assessments.

4.3.2.2 Processor Speed and Memory

The speed of the processor governs how fast the hardware can interpret and process instructions. A processor that is too slow for the software might run that software slowly, or even refuse to run it at all. The memory limits how much data can be stored. This is an issue that designers of internet-based simulations might come up against if, for example, the simulation required the computer to display full-screen, high-quality video. Bear in mind though that the candidate may have other software running on their computer when they access the simulation. Even though the simulation may only have modest requirements, it may be competing for resources with other software that might be running.

4.3.2.3 Screen Resolution

Screen resolution refers to the number of individual dots, or pixels, the screen contains. Typically, this is expressed by the number of pixels in each direction, e.g., 640×480 or 1024×768. Higher resolution means that higher detail can be displayed. As the numbers of pixels in both direction doubles, the amount of information contained on the screen quadruples: a screen of 1280×960 contains four times the data as a screen of 640×480.

Desktops and laptops typically have higher resolution than tablet devices, which have higher resolution than mobile phones. This is not always true, however: recent mobile phones are rivaling laptop displays for resolution.

4.3.2.4 Screen Size

Along with the resolution, simulation designers must take into account the physical screen size of the candidate's device. This is independent of screen resolution: a desktop display may be 22 in. diagonally, a tablet screen 12 in., and a mobile phone screen 3 in. Simulation designers must consider the sizes of the candidates' screens and design accordingly. That is not to say that all simulations must be designed to run on the smallest screen available, however: web pages can be programmed to detect the browser, operating system, and device upon which they are being displayed and, if the requirements are not met, to block its use and recommend to the candidate that he or she should use an alternative device.

4.3.2.5 Interface

The majority of desktop- and laptop-based software has to date been designed for keyboard and mouse (or touchpad). Users of these interfaces often become accustomed to the standard techniques of drag and drop, double-clicking, right-clicking, and so on. However, even with desktop computers there are differences. Mice used on Apple computers tend to have a single button. An instruction to a Mac-using candidate to "right click" in a simulation could well be met with confusion.

Many modern phones and tablet devices use touch screens. Early devices would allow one touch at a time. More recently, devices allow five or more simultaneous touches, giving the user the ability to literally physically manipulate on-screen information. This new type of user interface offers simulation designers new opportunities to engage and interact with users. On the other hand, many tablets and smartphones use "smart keyboards." These are on-screen keyboards that appear, often automatically, when the user is required to type information. As they are on-screen, simulation designers must be careful to ensure that the keyboard does not obscure important information. If they do, and if the user is forced to alternately minimize and restore the keyboard, they could well be at a disadvantage compared to real keyboard users who get to see the screen the entire time.

4.3.2.6 Sound

While most desktop and laptop computers, tablets and mobile phones have sound capability, this does not necessarily mean that they are available or accessible to the simulation. PC speakers may be unplugged, sounds may be muted, or sounds may be inaudible against loud background noise. If a simulation is dependent on audio content, candidates should be made aware of this at the beginning of the assessment, and given the opportunity to listen to sample audio to ensure that their device is playing the sound, and that it is audible in the environment. If users cannot hear the sound, they should be given the opportunity to use the simulation from another location or device.

4.3.2.7 Internet Connectivity

While broadband is increasingly commonplace, it is by no means universal. At the end of September 2012, 75.4 % of US households had fixed broadband (Mastrangelo 2012), with approximately 66 % of American adults having broadband connections at home (Zickuhr and Smith 2012). The number of adults who only have dial-up access at home is less than 3 %, and declining year after year (Zickuhr and Smith 2012).

So is it safe for simulation developers to design solely for broadband? Doing this could potentially limit the target audience's accessibility to the simulation. However, where a simulation is using audio, video, and animation, it may simply not be possible to create a version that will work via a dial-up modem.

Instead, simulation designers should design simulations for low-speed bandwidth use and, once created, they should be tested at those speeds. Many users of corporate and educational IT networks experience data speeds of 100 Mbps, 25–50 times faster than typical domestic broadband speeds. Testing the download speed of a simulation in a corporate environment will not emulate the experience of the typical home user.

4.3.2.8 Security

Whenever we use the Internet, some processes are carried out by the user's own computer (client-side), and some process are carried out by the host computer that the user has connected to via the Internet (server-side).

If we access a Flash-based simulation, then our computer downloads the Flash application and runs it. That application not only provides the user experience, but it also transfers the candidate's responses back to the server. But now that the simulation application has been downloaded client-side, it may be possible to disassemble that simulation, unpacking all of its contents and programming. This information could then be used by a candidate to cheat, or shared amongst other potential candidates. The simulation could even be "reverse engineered," or modified, to send back a series of correct responses to the server, enabling the candidate to receive a perfect score. This scenario does not just apply to Flash-based simulations, but any simulation that runs client-side code, including Java, Silverlight, and other technologies discussed.

Granted, there will only be a small proportion of candidates with the technical capability to do this. However, where a simulation might be administered to tens or even hundreds of thousands of candidates, there remains a possibility that its security could be breached.

There are steps that simulation developers can take to improve the security of simulations. The programming code can be "obfuscated," making it more difficult to disassemble and read. This may well deter some, but better security comes from minimizing the amount of data and information stored on the candidates' computers. This means putting as little data "client-side" as possible.

For example, a simple simulation might present a series of options to the candidate and they answer with a multiple-choice response. If that response is scored as correct or incorrect within the client-side Flash application itself—i.e., on the candidate's own computer—then the candidate essentially has the scoring key stored on their computer, albeit in a coded or obfuscated form. Potentially, that scoring key could be extracted, giving the candidate a list of correct answers. But instead, if the Flash application simply sends the raw responses back to the server for the server-side software to score, then the scoring key remains secure.

Through simple techniques such as this—as well as a number of far more sophisticated techniques—we can minimize the risk that simulated assessments become compromised.

4.3.2.9 Deploying Internationally

If the simulation is to be deployed to countries outside of the USA, then designers should investigate broadband speeds within those countries as they may not match those commonly seen in the USA. In some countries, an alternative, parallel assessment may need to be offered to accommodate candidates without access to broadband.

For those candidates without access to broadband, video or animation may not be a possibility: a short video of 30 s duration could take many minutes to download. An alternative, without resorting to a text-only assessment, might be to combine audio and still pictures, or even text and still pictures. The benefit of the latter approach is that applicants can then read and scroll through the text at their own pace, returning to any parts of the content to reread.

Simulation developers may run into technical challenges when deploying simulations into other countries. If the simulation is hosted—i.e., physically stored—in one country, say the USA, then users who are some distance away (e.g., in China) may experience a greater lag or delay than USA-based users would experience. Possibly, they may also experience a greater likelihood of their connection being lost. To overcome this issue, some simulation publishers host the simulation "in the cloud." This means that the simulation is hosted in multiple physical locations around the world. In a process that is completely transparent to the candidate, when they access the simulation they download it from the nearest host, instead of from a host located on the other side of the world.

4.3.3 Assessing in Virtual Environments

Some simulations can provide a richly detailed and immersive facsimile of the world in which users can engage, interact, and demonstrate their competencies, skills, and knowledge. Virtual environments are example of these types of simulations. The simulation designer has essentially a blank canvas to create an immersive environment in which users can see, hear, and experience situations which would not be practical to model in real life. Virtual environments are used extensively in military and medical training and assessment for this reason.

So if this is the case, why do we not see more VEs used for unproctored assessment? One of the reasons appears to be the learning curve. While VEs can offer a rich immersive environment, this can come at a cost to their usability: when people want to move around or manipulate objects in the real world, they simply do it—they walk and they pick things up. In a VE, however, these actions must be made through the interface between the user and the virtual world—typically the computer screen, mouse, and keyboard. Despite efforts of VE creators, this added layer decreases the usability of those environments compared to the real world.

One of the most extensively researched VEs is Second Life, a virtual world platform launched in 2002 by Linden Labs (Linden Research, Inc. 2012). Like other virtual worlds, it offers users an expansive world in which users take on the form of a virtual character or "avatar," allowing them to travel throughout the VE, interact with other users, and create or manipulate objects.

Some studies have compared real-world and virtual-world behavior in VEs such as Second Life, and found significant differences (Richardson et al. 2011; Satalich 1995). These findings suggest that the assessment of a candidate's performance in a VE is not necessarily an accurate measure of the performance that they might demonstrate in the real world. One of the reasons for this may be the steep learning curve of VE platforms (Mennecke et al. 2008). Assessing in a VE may introduce test error by favoring those candidates who have more experience of video gaming. Richardson et al. (2011) found that gaming experience was related to performance in desktop VEs. In addition, Ausburn (2012) found that performance was affected by gaming experience, gender, and age.

A source of error in VE simulations is the difficulty that some candidates have with learning the user interface and subsequently using that interface to navigate through the virtual three-dimensional world. This challenge appears to account for some of the differences in performance found by Richardson et al. (2011). One technique to eliminate or minimize the challenge of navigation is to remove that control from the candidate entirely. Instead of the candidate walking the character around the environment, he or she is taken around it as if they are "on rails." This has the result that the user still experiences some of the immersive quality of the VE, but with increased ease-of-use.

A second benefit of the "on-rails" approach is that the candidate experience can be controlled and standardized. In a VE in which a candidate is free to move around as they please, there is a risk that they may miss vital information. For example, if the assessment simulation takes place in a virtualized office environment and the candidate is free to move from one location to another, then they might rapidly navigate through the virtual office without noticing characters and information that could later prove to be instrumental in that candidate's approach to the task. Using an "on-rails" approach allows simulation designers to ensure that all candidates receive the same experience.

Figure 4.7 shows an example of an "on-rails" simulation. In this simulation, characters can interact with the patient and medical equipment through simple mouse clicks, increasing the ease-of-use for candidates.

Once the issue of usability is resolved, assessing using VE simulations offers the opportunity to move away from simple question/answer models of assessment and introduce many other ways of recording candidates' behavior. For example, a "traditional" assessment of spatial thinking may present candidates with a still image of a three-dimensional object, and five different images of unfolded versions of that object. They are then required to choose which one of the unfolded versions is the

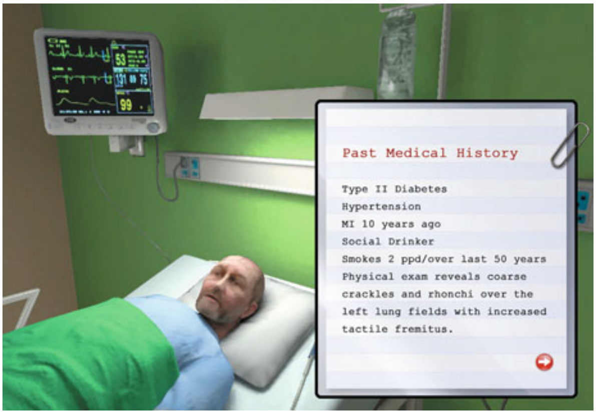


Fig. 4.7 An "on-rails" virtual environment-based simulation

same as the three-dimensional version. Once they have chosen, we have two data points that we can use for assessment: which answer option the candidate chose, and how long it took to respond.

However, if this same item is delivered in a VE-based simulation, we have the potential of measuring many more data points. If candidates are able to rotate the three-dimensional object, we can measure how many times the object was moved, what directions it was oriented in, or whether it was manipulated at all. We could also track mouse movement: did the candidate's mouse pointer hover over various answer options before clicking one, or was it moved directly to the chosen option? We could use eye-tracking technology to measure where on the screen candidates were looking: how long did they spend looking at the three-dimensional object? Did they look at each option equally, or did they glance at some options while paying more attention to others? Even a single VE-based spatial-thinking item like this could give rise to multiple data points that could be captured and used to calculate a candidate's score.

If we then move from item-based assessment to task-based assessment then the scope for capturing even more data points increases further. Take the example of an office-based simulation or inbox/e-tray. The candidate may have complete freedom to choose the order of e-mails they respond to in the simulation, how they plan their schedule, or which of the virtual characters they choose to interact with.

So as VE-based simulations impose fewer constraints upon the candidates' choices (compared to traditional multiple-choice tests), they can capture many more metrics about candidates' behavior.

With all these potential data points and variables, how then do simulation designers know which are relevant and which are spurious? After all, with the ability to capture so many variables in VE-based simulations, there is a likelihood that some of them will significantly correlate with future job performance, even though the correlation may be spurious and ultimately indefensible in a selection context.

There are methods that can help simulation designers focus on meaningful data arising from candidates' behavior in virtual worlds, and two examples of those methods are discussed here. The first is to develop a priori hypotheses about the relationship between candidates' VE behavior and real-world behavior. In the example of the spatial-thinking item mentioned previously, we might hypothesize that response time might be negatively correlated with spatial ability: candidates with high levels of spatial ability will answer the question more quickly. A second hypothesis might be that the extent to which the candidate rotated the three-dimensional object before choosing a response is also negatively correlated with spatial ability: candidates with high levels of spatial ability will need to rotate the object less before choosing a response. These hypotheses can then be tested by validating virtual-world performance against real-world performance measured by, for example, a "traditional" measure of spatial reasoning or by job performance. By developing a priori hypotheses in this way, simulation designers lessen the risk that candidates' suitability for a job will be measured by VE-based metrics that prove to be only spuriously correlated with real-world behavior.

A second approach is to treat the virtual world in the same way as physical assessment centers and use human assessors to observe and score candidates by using behaviorally anchored rating scales. For example, candidates participating in a VE-based simulation could be observed by assessors who are "present" in the VE, but invisible to candidates. Some VE platforms such as Second Life and SAIC's OLIVE allow for the simulation to be recorded and played back, not only as a movie but also as an immersive three-dimensional replay of the simulation, allowing assessors to repeatedly view the candidates' behavior, each time from multiple vantage points.

4.3.4 Knowing Your Audience

At the beginning of this section, deployment was described as one of the first factors that are taken into consideration when developing a simulation. Therefore, designers must very early on seek to understand the needs of their audience, the candidates. This information will inform not only the design of the simulation, but also the choice of technologies used to deliver that simulation.

Any information appearing in print, including this chapter, is soon surpassed by the rapidly evolving technology that drives the Internet. Therefore, simulation designers should seek the most current information available about their candidates and the candidates' technology—browser, plug-ins, operating systems, and so on. Two sources of information are StatOwl and W3Techs.

4.4 Conclusion

Deploying simulations for assessment over the Internet presents simulation creators with a series of challenges. However, as other chapters in this book demonstrate, there are compelling reasons for their use. The first step in designing a simulation is to understand the constraints imposed by this medium of assessment. Once these constraints are understood and accommodated, simulation designers can make extensive use of a range of tools and technologies to deliver engaging and valid assessment simulations.

4.5 Resources

This is not an exhaustive list, but is intended to help readers locate many of the resources mentioned in this chapter.

4.5.1 Authoring and Deployment

- Adobe Flash: http://www.adobe.com
- Articulate Storyline: http://www.articulate.com
- ClicFlic: http://www.clicflic.com
- HTML5: http://dev.w3.org/html5/spec/single-page.html
- Microsoft Silverlight: http://www.microsoft.com/silverlight
- Oracle Java: http://www.java.com

4.5.2 Character Animation

- 3ds Max: http://usa.autodesk.com/3ds-max/
- Blender: http://www.blender.org
- CodeBaby: http://www.codebaby.com
- Digital Video Toonz: http://www.toonz.com
- Moviestorm http://www.moviestorm.co.uk
- Muvizu: http://www.muvizu.com
- Reallusion Crazytalk: http://www.reallusion.com/crazytalk
- Reallusion iClone: http://www.reallusion.com/iclone
- RETAS: http://www.celsys.co.jp/en/products/retas/index.html
- Smith Micro Anime Studio: http://anime.smithmicro.com
- Toon Boom: http://www.toonboom.com
- Xtranormal: http://www.xtranormal.com

4.5.3 Text to Speech (TTS)

- Cereproc: http://www.cereproc.com
- Loquendo: http://www.loquendo.com

4.5.4 Accessibility

- Section 508 Guidance: http://www.section508.gov
- Web Accessibility Initiative: http://www.w3.org/WAI
- World Wide Web Consortium: http://w3.org

4.5.5 User Statistics

- StatOwl: http://www.statowl.com
- W3Techs: http://w3techs.com

References

Adobe Systems Incorporated. (2011). Flash to focus on PC browsing and mobile apps. Adobe featured blogs. http://blogs.adobe.com/conversations/2011/11/flash-focus.html. Accessed 7 Dec 2012.

Adobe Systems Incorporated. (2012). Statistics: Penetration by version. http://www.adobe.com/uk/products/flashplatformruntimes/statistics.displayTab3.edu.html. Accessed 7 Dec 2012.

Ausburn, L. J. (2012). Learner characteristics and performance in a first-person online desktop virtual environment. *International Journal of Online Pedagogy and Course Design, 2,* 11–24.

Brandt, A. (2012). Time to give Java the boot? PCWorld. http://www.pcworld.com/article/261843/time_to_give_java_the_boot_.html. Accessed 7 Dec 2012.

Deep Blue Sky. (2012). HTML5 and CSS3 support. FindmebyIP.com. http://fmbip.com/litmus/#html5-web-applications. Accessed 7 Dec 2012.

Draney, J. (2012). HTML5 and the future of online games—SitePoint. SitePoint. http://www.sitepoint.com/html5-and-the-future-of-online-games/. Accessed 7 Dec 2012.

Fox, J. (2010). What does a corporate web video cost? 25 Factors (with prices) that affect corporate video production costs (Web log post). http://onemarketmedia.com/blog/2010/03/what-does-a-web-video-cost-25-factors-with-prices-that-affect-video-production-costs/. Accessed 14 Dec 2012.

Franco, A. (2010). Mass confusion: The hysteria over Flash, Silverlight, HTML 5, Java FX, and Objective C. Anthony's Blog. http://anthonyfranco.wordpress.com/2010/02/01/mass-confusion-the-hysteria-over-flash-silverlight-html-5-java-fx-and-objective-c/. Accessed 7 Dec 2012.

Hawkes, B. (2012). Review of animation agencies rates. Unpublished raw data.

IMDb.com. (2012). Up (2009). http://www.imdb.com/title/tt1049413/. Accessed 7 Dec 2012.

Jobs, S. (2010). Thoughts on Flash. Apple.com. http://www.apple.com/hotnews/thoughts-on-flash/. Accessed 10 Dec 2012.

Le, Grecs. (2012). Java, Flash, and the choice of usability over security. Infosec Island. http://www.infosecisland.com/blogview/22381-Java-Flash-and-the-Choice-of-Usability-Over-Security.html. Accessed 7 Dec 2012.

Linden Research, Inc. (2012). History of Second Life. Second Life Wiki. http://wiki.secondlife.com/wiki/History_of_Second_Life. Accessed 7 Dec 2012.

Macromedia. (2002). Macromedia Flash MX—A next generation rich client. Macromedia. http://download.macromedia.com/pub/flash/whitepapers/richclient.pdf. Accessed 7 Dec 2012.

Mastrangelo, T. (2012). North American telcos struggle to gain broadband subscriber continues in 3Q12. The voice of broadband. http://broadbandtrends.com/blog1/2012/11/12/north-american-telcos-struggle-to-gain-broadband-subscriber-continues-in-3q12/. Accessed 7 Dec 2012.

Maxion, R. A., & Reeder, R. W. (2005). Improving user-interface dependability through mitigation of human error. *International Journal of Human-Computer Studies, 63,* 25–50.

Mennecke, B., Hassall, L. M., & Triplett, J. (2008). The mean business of second Life: Teaching entrepreneurship, technology and e-commerce in immersive environments. *Journal of Online Learning and Teaching*. http://jolt.merlot.org/vol4no3/hassall_0908.htm. Accessed 4 Dec 2012.

Richardson, A. E., Powers, M. E., & Bousquet, L. G. (2011). Video game experience predicts virtual, but not real navigation performance. *Computers in Human Behavior, 27,* 552–560.

Satalich, G. A. (1995). Navigation and way finding in virtual reality: Finding the proper tools and cues to enhance navigational awareness. Unpublished Master's thesis, University of Washington, Seattle, USA.

United States Government. (2011). Section 508 home. Section 508. http://www.section508.gov/. Accessed 7 Dec 2012.

W3C. (2012). Web Accessibility Initiative (WAI). World Wide Web Consortium (W3C). http://www.w3.org/WAI/. Accessed 7 Dec 2012.

Weiss, C. (2012). Bridge over troubled authoring tool waters. E-Learning 24/7 Blog. http://elearninfo247.com/2012/11/13/bridge-over-troubled-authoring-tool-waters/. Accessed 7 Dec 2012.

Winokur, D. (2011). Flash to focus on PC browsing and mobile apps; Adobe to more aggressively contribute to HTML5 (Web log post). http://blogs.adobe.com/conversations/2011/11/flash-focus.html. Accessed 9 Dec 2012.

Zickuhr, K., & Smith, A. (2012). Digital differences. Pew Internet & American Life Project website: http://pewinternet.org/~/media//Files/Reports/2012/PIP_Digital_differences_041312.pdf. Accessed 10 Dec 2012.

Chapter 5
Simulation Scoring

Eric Sydell, Jared Ferrell, Jacqueline Carpenter, Christopher Frost and Christie Cox Brodbeck

The history of psychological assessment stretches a century past, and until the advent of the Internet, it proceeded at a gradual pace. Now, as connected devices become ubiquitous, the methods we use to collect data are increasingly varied, and the amount of data our field captures is truly vast. For assessment developers, the game has changed—we are less focused on studying the properties of a response scale or particular item type and more concerned with the grand challenge of predicting human behavior. We are at a tipping point, at which our power to collect massive amounts of varied response data will allow us to reach the predictive potential of our field. Simulations are at the forefront of this future.

Although technology-based simulations represent an exciting and engaging future for the testing world, novel item types themselves will not revolutionize our ability to predict important outcomes. With current technology, there is no silver bullet that will significantly improve the size of our criterion-related correlation coefficients. This is not to say that there are no incremental gains that can be made—there certainly are—but we believe major predictive improvements will be made it two areas: (1) combining information across item types and assessment experiences, and (2) leveraging the power of increasingly large sample sizes.

Combinatorial Scoring Although a tremendous amount of research has been directed toward individual scales and item types, vastly less attention has been given

E. Sydell (✉)
Shaker Consulting Group, Cleveland, OH, USA
e-mail: Eric.Sydell@shakercg.com

J. Ferrell
e-mail: Jared.Ferrell@shakercg.com

J. Carpenter
e-mail: Jackie.Carpenter@shakercg.com

C. Frost
e-mail: chris.frost@shakercg.com

C. C. Brodbeck
e-mail: Christie.Brodbeck@shakercg.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_5,

to how diverse scales and item types interact to predict important outcomes. A high Extraversion score does not indicate that a subject will take every opportunity to speak. The complex human persona demands that if we are to achieve higher levels of predictability, we must take into account the effects of environment, mood, and more. The use of technology-based simulations as stand-alone assessments and the use of simulation exercises in combination with more traditional assessment item types hold promise for increasing predictive power through combinatorial scoring.

Big Data Our field's efforts at combinatorial scoring are drastically curtailed by lack of statistical power. Research is severely limited by small sample sizes. Many statistically significant findings have been reported that fail to hold up in cross-validation samples. However, the good news is that as connectivity and Internet delivery of assessments grow, organizations are increasingly able to provide large sample sizes for both validation projects and ongoing hiring needs. To be sure, many companies still do not collect numerical job performance data to the extent possible, but improvements are constantly underway. The big data movement has barely begun in the human resources arena; but, as it grows, we expect vastly greater ability to determine scoring methodologies that have ever-greater predictive power.

In addition, a fundamental shift is afoot in terms of how candidate data is collected. Whereas for nearly the entire history of our field, data have been collected using a question–response format, simulations are now allowing researchers to directly measure human behavior. In other words, we can now move from asking what a person *would do* in a certain situation, to observing how they *actually behave* in that, albeit virtual, situation. The online worlds being created are constantly becoming more lifelike, and as this occurs, we expect to see continual advances in levels of assessment realism.

How do you score a simulation? This question is unanswerable, as there are as many ways as there are simulations. You must first consider the purpose of the simulation—development, training, selection, etc. Moreover, if for selection, how will the scores be utilized? Will there be cut-off scores, subscores, broad or narrow scores? Once these issues are resolved, you can consider the type of simulation you will develop: Will it be a pure simulation measure or some combination of simulation and traditional item types? Will you seek to minimize adverse impact while also maximizing predictive validity against some type of criterion measure or measures? Will you have access to a validation sample? If so, how large will it be? All of these questions must be answered before determining the ideal scoring methodology.

The scoring of simulations is a dynamic and highly intricate topic. The macrolevel issues above will continue to influence the field for years; but, in this chapter, we also discuss a number of less nebulous topics. These include custom or local scoring models, broad versus narrow scores, automatic scoring of qualitative information, and branching logic.

5.1 Custom Scoring

The typical scoring philosophy among test vendors involves creating scoring routines based on analytical results culled from multiple samples of data. The intent is to guard against capitalizing on local variation by using scoring rules that have been shown to hold up in many different samples. We argue here that if the purpose of your assessment is to describe an individual in universal terms, as it is with many off-the-shelf assessments, this is the proper approach. However, if your purpose is to predict some outcome in a specific environment, then the situation becomes more complex. Too often, assessments that were designed to describe are inappropriately used to predict.

When the goal is to predict job performance outcomes, specificity matters. If the trait of Extraversion has a metaanalyzed validity coefficient of 0.3, does that mean it will predict for the role you are studying? One might surmise that it would be predictive of bank tellers' sales success, and yet at Shaker Consulting Group, our consultants have found a much less intuitive predictor to be a vastly greater and more stable indicator of sales success: computer skills.

If we compare two different sales positions, one a door-to-door sales position and the other an inbound call center position, we do not expect the trait of Extraversion to predict results identically. In the outside position, the primary relevant personality trait indicative of success may well be Extraversion, for it is incumbent upon the individual to assert him- or herself to unsuspecting and likely closed-minded potential customers. In the inbound role, Extraversion scores will likely not predict success at all, as the individual simply needs to answer the phone and ask scripted questions about whether a caller would like to purchase their products or services.

At Shaker Consulting Group, we create custom simulation modules for particular roles, but we typically hold constant the core measurement features of the simulation across clients. For example, a multitasking exercise might be customized to reflect the specifics of a role or organization with respect to the type of widget involved in some stacking exercise or the type of information shown in a call queue display. However, the core measure involving a numeric calculation coupled with a simultaneous task, such as clicking a button to take a new mock call, can well remain the same across different versions of the exercise. The validity evidence of the core measure is assessed and combined from many instances of administration, attesting to the stability of the measure.

One of the most valuable features of technology-facilitated simulations is the ability to include and deliver a wide variety of job-related assessment activities to candidates in a manner that more realistically depicts actual job tasks. The use of custom scoring with simulation-based assessment activities allows us to meet the goal of creating a high fidelity, realistic experience for candidates while also capitalizing on the rich data obtained from simulation exercises to predict on-the-job performance more precisely. Although the value of variety and realism of technology-facilitated simulations has led to an increased focus on creating novel, realistic exercises to assess different facets of performance, an important issue that we cannot afford to overlook involves decisions about the ways to summarize data to present candidate scores.

Researchers have recently examined whether meta-analysis, local study, or Bayesian analysis is the most accurate way of estimating local validity (Newman et al. 2007). Contrary to some conventional wisdom, local studies can actually provide more accurate estimates of predictor validity in some circumstances than meta-analytic methods. However, the best way to estimate the validity of a custom scoring procedure may be to combine the local estimate with Bayesian prior probabilities generated from meta-analytic results. Using this technique, selection scientists can weight a local validity result with the prior meta-analytic evidence to arrive at a more stable validity estimate (or Bayesian posterior).

In the scoring arena, one of the most common debates deals with the relative effectiveness of generating and reporting broad versus narrow competency scores. In order to assist test developers in deciding whether broad or narrow competency scores are most appropriate for a simulation-based assessment, the benefits and drawbacks of each as well as the best practices for implementation are discussed in Sect. 5.2.

5.2 Competency Scores: Broad Versus Narrow

5.2.1 Broad Performance Competency Scores

At a high level, simulations can be developed to yield specific scores around constructs such as multitasking ability, typing speed, cash transaction accuracy, and many other narrow scores. These may well be predictive of certain aspects of on-the-job performance. However, the present authors argue that simulations can be leveraged much more broadly to predict not only specific task performance factors, but also higher level competencies that represent a sizable portion of the job performance domain. In fact, more than other assessments, simulations offer the potential to predict overall job performance due to their ability to include diverse item types and provide a more realistic experience. Simulations offer the ability to measure a candidate in a more holistic fashion than an assessment developed around a specific item type or construct.

Broad competency scores combine multiple narrower facets and dimensions from either the same or different simulation exercises into more general, all-encompassing output reports. The goal of this combination effort is to utilize customized combinations of exercises to explain performance better than is possible through the utilization of narrower facets. These broad composites, highly popular in the world of customized assessments and simulations, are often created to align with an organization's own competency language, making them more easily interpretable by lay clients not trained in the technical aspects of personnel selection. For example, a broad competency score could be computed to predict overall performance based on a simulation, instead of reporting a score from each specific aspect of the simulation.

5.2.1.1 Benefits

A significant benefit of creating broader competency scores is that this practice allows key variables to be at the center of attention (Smith 2002). For example, a broad scoring composite could be created to measure expected overall job performance based on performance on a simulation. This has the benefit of putting performance at the forefront of attention when decision makers within the organization look at output reports.

Broad performance competencies also afford decision makers with increased efficiency not offered by narrower competency scores. Looking back at the example given above of a broad performance composite, this allows for an easy rank ordering of candidates on a key variable or variables of interest. As described by Hattrup (2012), no matter how many exercises are on a selection assessment, ultimately the decision boils down to a dichotomous choice between hiring and not hiring the candidate. Thus, this broad scoring methodology makes it much easier for decision makers to see the big picture instead of being caught up in an overabundance of narrow facet scores when drawing conclusions from assessments.

Multiple studies have shown that the use of composites can help to decrease the potential for an assessment to exhibit adverse impact (e.g., Bobko et al. 2007; Sackett and Ellingson 1997; Schmitt et al. 1997). This is generally due to the compensatory nature of broad competencies, in which case different scales all combine in a certain way to provide valid assessments with minimal risk for adverse impact. Other researchers have taken a more technical approach to finding optimal weighting schemes for maintaining high levels of validity while minimizing the risk for adverse impact.

De Corte and colleagues, in a series of studies (i.e. De Corte 1999; De Corte et al. 2007), took a technical approach to find the pareto-optimal tradeoff between validity and adverse impact. The pareto-optimal tradeoff attempts to solve the diversity–validity dilemma, wherein assessment providers are constantly in a tug of war between trying to provide the highest levels of validity while also minimizing the potential for an assessment to result in adverse impact—two tasks at odds with each other (De Corte 1999; Pyburn et al. 2008). The term "pareto-optimal," born from economics literature, refers to situations in which the increase of one factor is at odds with another factor. Applied to selection, increasing validity is at odds with decreasing the risk of adverse impact, and thus trying to optimize one inherently works at the expense of the other.

De Corte (1999) initially proposed a constrained nonlinear methodology for creating composites to minimize adverse impact concerns while working to maximize performance gains from a selection system. In De Corte's methodology, a constraint was placed into the weighting equation, setting the minimum acceptable adverse impact ratio. This value became a key consideration in the formation of the weights. A limitation of De Corte's methodology was that instead of optimizing both variables, the equation only optimized one (validity) while constraining the other (adverse impact potential). This issue was addressed by De Corte et al. (2007), in their article, which provided a new formula for the calculation of pareto-optimal tradeoffs between adverse impact and validity. The updated model presents numerous points that show differing levels of validity and adverse impact potential, based on differential weighting of the scales within the broad competency. One point included is the pareto-optimal level for validity and adverse impact in conjunction with each other.

The weighting equations described above provide evidence of examples where differential weighting of different subfacets within a broad composite allows for a compensatory system in which validity and adverse impact are optimally balanced. Scales can be weighted together in formations such that a single scale or exercise that is not only highly valid, but also at a higher risk for adverse impact (e.g., a cognitively loaded measure) can be combined with other predictors (e.g., personality) that are at a lower risk of violating adverse impact ratio cutoffs, creating a composite that is highly predictive of performance and also adheres to federal regulations on adverse impact.

5.2.1.2 Drawbacks

While there are numerous advantages to the utilization of broad scoring composites, there are also some distinct drawbacks to this practice, depending on the situation. In certain instances, broad competency scores may actually serve to disguise serious failings of candidates due to the inherent compensatory nature of the scoring system. For example, a candidate could score well on a broad composite without raising red flags that would be more likely to present themselves with narrow competency scores. Along the same lines, broad composites can potentially disguise the factors in a simulation that are the key drivers of performance making it hard to home in on key ways to increase organizational effectiveness.

Weighting issues, while possessing the potential to benefit a system in certain situations, also present potential drawbacks to the utilization of broad competencies in simulation scoring. As composites begin to integrate more factors, weighting of individual scales or exercises becomes a critical issue. The problem lies in the fact that weighting is not straightforward, thus subjectivity becomes injected in the scoring systems (Gatewood et al. 2010). For example, there are four main schools of thought on the weighting of predictors to form composites: (1) regression weighting, (2) reliability weighting, (3) a priori weighting, and (4) unit weighting (Hattrup 2012). The issue lies in the fact that there is much disagreement in the literature regarding which weighting scheme is optimal, leaving practitioners in a precarious position in terms of having to make and support a weighting decision. Doverspike et al. (1996) advise practitioners, no matter which scheme they utilize, to describe in detail the process implemented in the decision making regarding weights within a composite.

The final major drawback to broad performance composites strongly parallels any overly large organizational intervention. The issue is that when many parts are intertwined, a major change to one of the facets within a competency score could cause a host of weighting and predictability issues in the composite as a whole, leading to an entirely new set of headaches for practitioners. As such, practitioners need to be cognizant of this when choosing to utilize broad competency scores and ensure they understand the effects changing one part of the simulation can have on the properties of the competency score(s) as a whole.

5.2.2 Narrow Competency Scoring

Narrow competency scores are generally comprised of separate facets for different exercises and competencies and are presented as different metrics on the simulation output reports. For example, problem-solving skills would be separated from personality facets in the output report instead of potentially being combined to create an overall performance score as they might be in a system implementing broad competency scores. This approach to scoring has its own benefits and drawbacks, discussed in Sect. 5.2.2.1 and 5.2.2.2.

5.2.2.1 Benefits

The primary benefit of using narrow scales is their ability to present results for each facet or section in a more straightforward manner than their broad competency counterparts do. The results are much more transparent than many of the broader composite scores. This has numerous positive implications for practitioners, including ease of theoretically linking predictors with narrow criteria dimensions, increasing the ease of showing rationale behind inclusion of assessment aspects (Arthur et al. 2003; Christian et al. 2010).

The narrow scoring of facets also makes weighting less of an issue in most cases. This is because narrow competency scores generally do not require differential weighting of different simulation activities, instead, commonly requiring unit weighting of items into each narrow composite. This reduces the potential for subjectivity in the initial weighting of the competencies and in certain cases may make a system less vulnerable to legal action based on the scoring methodology.

Narrow composites offer the potential for more direct feedback than their broad counterparts do. The narrow composite approach is well suited to discovering and illuminating what specific facets actually drive performance while offering feedback that focuses on those specific facets. Moreover, they can be equally beneficial for the reverse situation, in which it may be important for decision makers to flag candidates for serious deficiencies on specific competencies or simulation exercises that have been shown to be critical to organizational success. The effects of this, on deciding which type of scoring competency to utilize, will be discussed in the best practices section (Sect. 5.2.4).

5.2.2.2 Drawbacks

While there are benefits to having more detailed output reports that include multiple facets, this method is not without its own drawbacks. The drawbacks here are often due to the interpretation of numerous narrow composites. Although broad composite scores are computed via an actuarial manner, wherein there are hard numbers to back up decisions, if care is not taken with training end-users on the meaning of various narrow facets, it is easy for hiring decisions to be based on softer interpretations, which may be more difficult to defend after the fact (Grove and Meehl 1996). This ties back to the tradeoff in which broad composites throw subjectivity into the creation of the composites themselves and narrow composites inject subjectivity into the potential for differential interpretations of the same set of scores, thus necessitating some sort of output report training to try and mitigate the potential for this to decrease the utility of the simulation as a whole. Moreover, the narrow facets, while interpretable to developers of assessment content, may not be interpretable to lay end-users. This issue ties into the concern about inconsistent interpretation of output reports across key decision makers. As an example, decision makers may not completely understand what a narrow facet, such as Extraversion, directly means, or how it would specifically relate to performance, causing different interpretations depending on who is reading the output report of a candidate.

Another concern with this approach is the potential to accumulate too many narrow facets or composites on a scoring output, leading to information overload, and thereby decreasing the administrative efficiency of making decisions from the simulations. Indeed, it is very common for simulations to have upward of 30 or more different narrow scales or composites, which can quickly become a nightmare for decision makers within an organization. As such, practitioners should be careful to focus on key composites that drive performance; in addition, they should even potentially consider removing more peripheral scales or composites that focus on predicting extremely narrow subsets of performance.

5.2.3 Psychometric Considerations

It is a fundamental element of psychometrics that you cannot have validity without reliability. This general dictum has been ingrained into graduate students' brains for ages (along with the idea that correlation does not imply causation). However, there is vast misunderstanding of the nature of the relationship between reliability and validity.

Theoretically, a measure must be reliable in order for it to be valid; however, in practice, it is extremely difficult to verify this relationship. The vast majority of scale development utilizes coefficient alpha as the reliability estimate of choice due to its simple computation. However, internal consistency is but one type of reliability estimate, and while internal consistency is important for scale interpretability, what if the scale is combined with other items or scales to yield a broader competency score? A heterogeneous scale may still be a reliable indicator of relevant characteristics. When the purpose of a simulation is shifted from description to prediction of real-world outcomes, interpretability is less important than high predictive power.

Many continue to focus on coefficient alpha for its ease of use. However, when creating broad competencies, we recommend following the newer approach of Linear Composite Reliability (Nunnally and Bernstein 1994). Although not discussed here, this approach provides an estimate of reliability that takes into account the reliabilities, relative weight, and variance of each component within the composite as well as the overall variance of the composite.

5.2.4 *Implementation of Broad or Narrow Competency Scores: Best Practices*

It is prudent to disentangle the benefits and drawbacks of broad versus narrow competency scores through explaining situations better suited to one methodology over the other. This section will begin with situations better suited to broad scoring composites, followed by an examination of situations in which a narrow scoring composite is the more appropriate choice.

5.2.4.1 Situations Best Suited to Broad Versus Narrow Scoring Composites

There are numerous situations in which either broad competency scores or narrow competency scores are better suited to achieving the goals of the simulation through which they are derived. Factors influencing the relative effectiveness of broad versus narrow competency scores are numerous, and thus an exhaustive list of situations is beyond the scope of this chapter. Nevertheless, there are certain general cues that can help practitioners decide whether to generate broad versus narrow competency scores based on a candidate's performance on a simulation. These clues can come via situational constraints, the purpose of the simulation itself, and the nature of the outcomes the simulation is designed to predict.

As discussed previously, situational constraints can determine the optimal composite construction methodology. One such organizational constraint deals with the time allotted to make decisions based on assessments relative to the number of people who complete an assessment. Based on a sheer lack of available time by key decision makers, the efficiency in decision making is often vital in organizational settings. For example, imagine a company that administers a simulation to thousands of candidates for a small number of job openings. It would be overwhelming for hiring managers to sift through report after report littered with narrow competency scores. Instead, a broad performance composite score would be ideal here, as the hiring managers could utilize applicant-tracking databases to sort candidates based on how well they are predicted to perform overall in this specific work environment. This would make deciding which candidates to advance to the next stage of the selection/promotion process much easier and more straightforward as opposed to trying to compare thousands of candidates on numerous narrow facets.

The purpose of the simulation can also be used as a deciding factor in whether to create broad or narrow competency scores. Although each situation will be different, there are some general situational factors that can affect whether broad or narrow competencies would be better suited. As discussed above, broad competency scores can increase the efficiency and uniformity of decisions, and therefore can often times be more practical for decisions being made in a hiring/promotion context, especially one with a high volume of candidates. On the contrary, the specificity of narrow competency scores allows for the ability to understand and alter specific behaviors and is well suited for developmental exercises in which it is valuable to be able to pinpoint specific opportunities for future skill enhancement, whereas broad competency scores would only be able to identify if there is a gap in performance at a much more general level, thus not being able to give specific suggestions for improvement.

The nature of the criteria may also affect the decision of whether it would be optimal to utilize broad or narrow scoring composites. Ideally, the goal is to match the criteria with the predictors, such that if the criterion is broad, a broad composite would be viewed as optimal, and if the criterion is narrower in nature, the scoring composite should be narrow to match as well. As an example, if the criterion of interest is organizational performance, it would be more advantageous to have a broad performance composite than to generate a plethora of narrow composites and expect decision makers to wade through the information and draw conclusions. Conversely, if the criterion of interest is communication skills, a narrow composite composed of exercises that tap this factor is going to be more beneficial than would an overarching performance composite.

While certain examples have been given in which broad or narrow competencies are preferred, it is also often the case that both are utilized in congruence with each other. For example, numerous simulations generate output reports that include broad competency scores and more narrow scores to help reap the benefits associated with the utilization of both. This methodology is beneficial because it allows decision makers to be able to employ a cursory screening of unqualified candidates as well as a more in-depth comparison of qualified candidates before proceeding to the next step in an organizational decision-making process.

Up to this point, we have discussed scoring considerations relevant to any assessment employing simulation exercises. As discussed in this chapter and others in this book, the array of simulation exercises used in assessment is vast and varied, thus it is beyond the scope of this chapter to cover scoring considerations specific to each type of simulation exercise used in assessments. However, we do want to highlight scoring considerations pertaining to some particular innovations in simulation exercises and test construction. The following section will discuss the use of automatic scoring in computer-based simulations and the techniques associated with the application of this methodology to simulation construction and scoring.

5.3 Automated Scoring of Qualitative Data from Simulation Exercises

An exciting innovation offered by technology-facilitated simulations is the opportunity to collect and automatically score open-ended responses from candidates. Compared with cumbersome, essay-style assessments of the past, simulations allow organizations to collect writing samples from candidates in novel ways and tailor the stimuli to job-specific situations. For example, a candidate may be presented with a hypothetical situation or problem and be required to generate multiple possible solutions or strategies to solve this problem. Embedding open-ended items within simulation exercises allows organizations to gain a more comprehensive understanding of the candidate's thought processes and complex problem-solving skills relative to a simple Likert-type scale item (Ackerman and Smith 1988; Birenbaum and Tatsuoka 1987). Furthermore, the range of possible responses is virtually unlimited, providing additional information that may be particularly useful when attempting to select candidates for higher level positions such as managerial or leadership positions (Zaccaro et al. 2000).

Until recent technological advances, the benefits of including qualitative elements in selection assessments were overshadowed by administrative impracticalities. Prior to the advent of automatic scoring methods, the time demands required to evaluate these responses were extensive, not to mention costly. Consider that each response must first be generated by the candidate before an appropriate scoring system is developed; responses must then be read in their entirety and finally scored. In addition to the time requirements, this scoring methodology is vulnerable to rater errors (Zaccaro et al. 2000).

The development of an automated scoring process promises to mitigate (if not eliminate) the disadvantages associated with qualitative scoring, while also maintaining the measurement benefits of this method. Reducing the time requirements in what is typically an extremely time-intensive process is the most obvious advantage of applying an automatic scoring methodology to qualitative data. Without an automated scoring process, multiple reviewers or raters would be required to read and score each piece of writing. An automated process eliminates this time consuming endeavor as all pieces of writing can be scored instantaneously. In addition to the reduced time requirements, automatic scoring processes introduce an increased level of objectivity to the scoring of qualitative data. A third advantage of applying an automated scoring process to written text is the comprehensive nature of the evaluation. Regardless of how efficient a human rater is or how closely they read a written sample, they will not be able to remember every word that they read and factor it into their final evaluation. Automated scoring methods by contrast, are capable of evaluating each piece of text in the response and using each word to develop a refined scoring process.

5.3.1 Automatic Scoring Methods

To date, automatic scoring methods have been applied more frequently in the education context (i.e., evaluating student essays, ACT, GRE) than within the selection domain (Attali 2004; Burstein and Chodorow 1999). Results from research in the education domain generally indicate that these automatic scoring methods can reliably reproduce human ratings, and in some cases, the automatic scoring methods actually appear to be more accurate than a human grader (Burstein and Chodorow 1999; Shermis 2012). Although there are numerous automatic scoring software programs available, the programs tend to be conceptually similar to each other[1] (Fielding and Lee 1998; Tesch 1990). However, these automatic scoring software programs vary somewhat in their methodology. Some programs utilize essays that were previously scored by human raters. These essays are divided into groups (e.g., high scores, average scores, and low scores), and the program is then trained to recognize the key differences between the essays in each group. In addition, many of these software programs evaluate writing samples based on grammatical properties such as subject-verb agreement, sentence completion, and punctuation.

Evaluating grammatical properties is not only popular in the educational domain, but it is also utilized in other contexts, including employee selection. The primary advantage of utilizing these types of techniques is that they are generalizable to nearly any context. However, there is perhaps a great deal more that can be uncovered by examining factors other than grammatical quality.

5.3.2 Measuring More Than Essay Quality

Utilizing automatic scoring to examine constructs beyond simple essay quality may be particularly valuable for organizations when assessing candidates. Consider that for many jobs composing a grammatically sound sample of writing may not be something an employee is required to perform. However, with technology-facilitated simulation exercises, qualitative item types can be used in novel forms to create a higher fidelity experience that taps numerous job-relevant constructs. There are some encouraging research results relating open-ended items to personality traits, leadership characteristics, and coping styles. For example, research on automatic scoring systems has shown their ability to predict personality traits, leadership characteristics, and individual coping styles based on qualitative response characteristics such as word choice (Fast and Funder 2008), simple word count (Hirsh and Peterson 2009; Lee and Cohn 2009), and idea complexity (Dudley and Cortina 2008). Assessing numerous constructs allows for job specific customizable scoring that achieves maximum validity. These benefits do not come easy, as there are numerous challenges associated with developing an automatic scoring system for qualitative items.

In general, when attempting to measure constructs beyond basic essay quality, scoring development is somewhat complex. For example, identifying words and phrases that indicate a high standing on a particular construct is not a simple process. Consider an assessment intended to measure a trait such as Conscientiousness. To develop the automatic scoring system, it first must be determined which words and phrases are more likely to be used by a person high on conscientiousness compared

---

[1] It is estimated that eight automatic scoring methods (AutoScore, LightSIDE, Bookeete, E-rater, Lexile, Project Essay Grade, Intelligent Essay Assessor, Crase, and IntelliMetric) represent approximately 97 % of all of the automatic methods used today to evaluate student essays (Shermis & Hamner 2012).

with a person low on conscientiousness. Although this task in itself is complex, it is made more so when considered in a selection context. As previously discussed, free form writing samples have generally been utilized to demonstrate the effectiveness of automated scoring methods to predict personality. When a candidate is composing a written sample for a potential employer to examine, they are not using a free form writing style. Candidates will generally try to put their best foot forward during the assessment process and therefore the range of responses is extremely restricted when compared with free form written samples. Therefore, the amount of words that will effectively differentiate among candidates on a particular construct is likely to be reduced in assessment contexts.

5.3.3 Automated Scoring of Qualitative Data: Future Directions and Suggestions for Practice

While identifying certain keywords and phrases to train an automated scoring system is not an easy task, there are software programs that can help to facilitate this process. One user-friendly program that can be utilized during this process is the Bayesian Essay Test Scoring System (BETSY; Rudner and Liang 2002). To use BETSY, one would divide writing samples into groups based on a particular criterion. For example, writing samples from highly extraverted individuals would be placed into one group while those from low extroverted individuals would be placed into another. The BETSY software would help to identify words that are used more frequently by members of the high Extraversion group. Users would be asked to identify words that differentiate between groups and are theoretically appealing. Eventually, these words and phrases can then be uploaded into software that identifies keywords, or BETSY could actually be used to score these written samples.

The BETSY software, along with many other software programs, utilizes a simple word count method. Initially, this word count function may appear overly simplistic. However, with respect to personality traits, there are numerous reasons why one would expect simple word choice to be related to candidate personality (Fast and Funder 2008). Furthermore, when utilizing a simulation that requires candidates to recall specific information from a particular passage, word count procedures can be very useful in scoring (Mumford et al. 2000).

There are some important issues that must be considered before implementing a word count procedure as part of the scoring for a simulation exercise. For example, consider an assessment that is designed to measure Achievement Orientation. It may be that words such as “motivated” and “driven” are identified as more likely to be used by a person that is high in Achievement Orientation. However, if a person uses a word such as “inspired” rather than “motivated” they would not receive an increased Achievement Orientation score despite the fact that they were attempting to convey the same message as another person who may have used the word “motivated.” Thus, careful consideration must be given to issues such as whether to score certain synonyms theoretically linked with the words identified by the word count software.

Automated scoring of qualitative item types is merely one example of innovative item types used in technology-facilitated simulations. Technological advances have also allowed for innovation in the construction of simulation-based assessments. For example, test developers now have the option to choose whether to use a nonlinear testing approach in place of a traditional linear testing approach, in order to provide a customized experience for candidates and achieve more precise measurement. Instead of presenting the same content to all candidates, nonlinear approaches such as branching logic or Computer Adaptive Testing (CAT) provide candidates with a customized experience based on their individual test performance. Applying nonlinear techniques to simulations for employee selection can enhance many of the benefits of simulation-based assessment.

5.4 Branching and Adaptive Testing

While still a relatively young approach to testing, nonlinear testing approaches are gaining popularity. The military introduced the first large scale CAT, the Armed Service Vocational Aptitude Battery (ASVAB) in 1982. CAT is a common type of nonlinear test design in which the test adapts to the candidate by successively presenting items representing a higher or lower level of the test construct based on the candidate's previous response (Drasgow and Olson-Buchanan 1999; Wainer 2000). For example, if a candidate incorrectly answers a mathematical ability item, he will be presented with an easier mathematical problem next. After multiple iterations of presenting items based on performance on the previous item, the test algorithm is able to pinpoint the candidate's level of the construct being measured. Since the introduction of the ASVAB, CAT has been increasingly used for professional and licensure examinations as well as academic entrance examinations such as the Graduate Record Examination (GRE). More recently, private sector companies have begun introducing nonlinear assessments to their employee selection processes. Beyond CAT, which may be primarily used when the goal is to shorten a test while maintaining or enhancing measurement precision, other nonlinear test approaches may be used to create branching or storyline experiences for candidates. The goal is to adapt the test to the candidate's performance in order to create a realistic simulation that enables more sophisticated modeling of potential performance. Both methods can be applied to simulation scoring and thus are discussed in the following sections.

5.4.1 Benefits of Nonlinear Testing

Nonlinear testing approaches can offer several enhancements to simulations for employee selection. Although typically associated with multiple-choice knowledge-based testing, nonlinear approaches can be leveraged across a variety of constructs and measurement formats. Techniques such as work samples, problem solving, and situational judgment tests can be implemented using various response formats. Beyond multiple-choice, candidates can interact with the content in ways that allow for more refined data collection such as clicking on a portion of a graph, dragging and dropping content, using a sliding bar, or manipulating content to build and create (Bejar 1991; Clauser et al. 1997). In addition, multimedia approaches—audio, video, animation, and a variety of imagery—can be easily incorporated. For example, a measure of mechanical ability can require candidates to view a video or image and then click on the part of a machine most likely to be causing a problem. Furthermore, because nonlinear testing allows for the customization of content based on previous responses, a candidate's problem-solving process can be explored in a way that mimics on-the-job scenarios. After selecting the machine part, candidates can be prompted to select which tools they would use to investigate the problem or to choose their next step, if the first hypothesis is incorrect. Storyline branching approaches such as these create an interactive and unique experience for candidates while also serving as a preview of the job. Importantly, they also give richer and more job-relevant data to potential employers. An added benefit for employers is that, unlike traditional approaches to evaluate branching or adaptive testing such as assessment centers, this complex information can be scored instantaneously.

The growing popularity of nonlinear testing can be attributed to several psychometric benefits touted over traditional linear tests, particularly CAT. By presenting targeted items based on a candidate's performance, fewer items are needed to zero in on the candidate's level of the measured construct. Candidates spend less time on items that are too easy or difficult, and more time on items that enable the fine tuning of their scores (McBride and Martin 1983). By reducing test length, test security may also be enhanced. Each candidate receives a different version of the test and is exposed to fewer items, making it less likely that test content will be shared or compromised. Even when compared with alternate forms of traditional tests, CATs help prevent cheating by giving a slightly different version to each candidate (Guo et al. 2009). This is especially beneficial for unproctored testing where content can be more easily compromised.

CATs have also been shown to enhance the reliability and validity of a test. Shortening a test through CAT also reduces fatigue associated with long tests (Tonidandel et al. 2002). Without fatigue effects clouding construct measurement, CAT can produce assessments that are more reliable and valid than linear tests despite their shorter lengths. Gains in validity also stem from the greater precision of adaptive tests. Because test content is successively targeted to candidates' individual performances, more information can be gathered about their specific ability levels. Although linear tests are often best for measuring medium levels of ability, CAT's customized content means that the test can quickly target test content for a wide range of ability levels (McBride and Martin 1983).

Importantly, adaptive testing is typically perceived positively by candidates as well. Using adaptive testing techniques to create simulations and storyline experience can feel more realistic and face valid for candidates (Hanson et al. 1999; Parshall et al. 2010). Proctor & Gamble also found in an extensive implementation of CAT that candidate perceptions of the appropriateness and fairness of the CAT were comparable to traditional tests (Gibby et al. 2008). In addition, the shorter length reduces test fatigue and the time needed to complete the test, making the test less taxing on the candidate. Research on the effects on test anxiety has been less consistent. Although anxiety is often reduced for those of medium and low abilities, adaptive testing may increase anxiety for those who do well early in the test as they will be quickly presented with difficult items (Tonidandel et al. 2002).

5.4.2 Complexities and Considerations

Although there are many benefits of nonlinear testing, this approach should be carefully considered before implementing. One important consideration is the technology comfort level of the intended test audience. Computer based testing in general may not be appropriate for audiences unfamiliar with the technology, and the expected computer proficiency of the target group should be considered when designing the interface of the test (Parshall et al. 2010). Creating an overly complex or customized test experience may limit the usability of a test.

In addition, the complexities and resource requirements inherent in this approach mean it is not always appropriate and must be carefully implemented. Implementing and maintaining an adaptive test is time and resource intensive. In the case of CAT, creating the test requires a large pool of items representing all levels of the targeted construct and a large sample to validate those items. Depending on the target group for the test implementation, this approach may or may not be suitable. For example, for small target groups, the expense and resources to develop the test may not be justified.

Complexities also arise in item development and the creation and implementation of complex scoring and item presentation algorithms. Mistakes in the development process can greatly impact test validity. Items must appropriately represent all levels of the construct, and algorithms must accurately choose and score responses based on previous responses. For example, a balance must be maintained between creating a short and statistically sound test and fully representing the construct domain. Too much attention to time and statistical considerations can result in narrow measurement of the construct domain for some candidates and create nonparallel forms of a test that do not allow candidates to be accurately compared (Huff and Sireci 2001). Implementing branching logic or adaptive techniques in simulation-based selection assessments requires rigorous item bank and test structure development as well as consideration of several scoring options.

5.4.3 Assessment Design and Scoring Considerations

The psychometric foundation guiding item calibration, item selection, and scoring for CAT is Item Response Theory (IRT). IRT advanced by Rasch (1960), Birnbaum (1968), and Lord (1970) describes the relationship between observed test performance and a test-taker's underlying ability on a particular trait. Typically, one or more item parameters (e.g., item difficulty, discrimination) are combined within a logistic function (Folk and Smith 2002; Lord 1980; Wainer et al. 2007). Although the simplest of models, the one-parameter logistic (1-PL) model considers just one item parameter, a three-parameter logistic (3-PL) IRT model is common among modern CAT designs.

The chosen model informs the calibration of test items, such that fitting an item response model to pretest data allows for the recording of each item's estimated parameters (Wainer and Kiely 1987). IRT also provides the basis for adaptively selecting test items based on an examinee's response to previous items (the item selection algorithm), and scoring the adaptive test as a whole.

Due to the heavy demands placed on test items in nonlinear testing environments, one of the costs of developing adaptive tests of any sort is the writing and calibration of items. Item pools for single-construct, IRT-based adaptive tests are necessarily large in order to cover the range of ability assessed and provide a sufficient amount of alternative items for the purposes of test security. In developing items for adaptive test formats, tests developers must consider the range of ability in the candidate pool, content coverage, and adhere to the assumptions of item parameter estimation models (Flaugher 1990; Hambleton 2002). Large pools of items also necessitate a substantial pool of trial items and large samples of pilot test-takers (Zickar et al. 1999).

Methods for adaptive item selection must be chosen with consideration of several test features, including psychometric characteristics, content specifications, and item exposure (Folk and Smith 2002). Popular item selection algorithms used in modern adaptive test designs are often based on the psychometric selection criteria of maximizing information about the examinee's current ability level (Folk and Smith 2002; Lord 1977; Rasch 1960). Alternatively, Bayesian item selection methods, such as the one described by Owen (1975), are also often used. Beyond these, other models for automated item selection have been developed with the goal of improving the balance between psychometric efficiency and content requirements (e.g., the weighted deviations model, Stocking and Swanson 1993; optimal constrained adaptive testing, van der Linden 1998). Van der Linden and Pashley (2010) provide detailed technical overviews of several modern alternatives to traditional adaptive item selection models. Item selection models often also include provisions for systematically monitoring item exposure for the purposes of item pool longevity and test security. For more details regarding item exposure, test security, and item pool maintenance, see Davey and Nering (2002) as well as Segall and Moreno (1999).

Test developers must consider multiple options for estimating final scores in order to achieve the optimal balance between precision, simplicity, and fairness. Administrators often confront difficulty in achieving a simplistic scoring model that can be explained to candidates with complex nonlinear tests. Scoring for adaptive tests depends upon test delivery methods and scoring is affected by how examinee responses are modeled. Therefore, developers must be particularly attentive to the manner in which the item parameter estimation and item selection algorithms impact final scores. For example, Owen's Bayesian ability estimate, tested for use in the CAT-ASVAB to update provisional scores after each item and provide a final score, has the undesirable property of providing final scores that depend on the order in which items were administered (Segall and Moreno 1999). That is, two candidates who answered the same items with the same responses may end up with different final Owen's ability estimates if they received the items in a different sequence.

Adaptive test design features, such as test length and response time modeling, are also important considerations relating to the generation and use of examinee scores. Adaptive tests can take the form of fixed or variable length. For variable length tests, two stopping rules may be used. A target standard error of measurement can be determined and additional questions are presented to the candidate until the target is met, or the tests can be stopped when a specified level of confidence in the pass/fail decision is met (Bergstrom and Lunz 1999; Kingsbury and Weiss 1983; Segall and Moreno 1999). Test developers must consider the incremental informative value of each additional item to determine whether fixed or variable length is appropriate for a particular assessment (Segall and Moreno 1999).

During the creation of the CAT-ASVAB, developers discovered a trend in candidate total time that was opposite of the anticipated response times from the pencil and paper version of the tests. That is, higher ability candidates were spending more time because they received more difficult questions requiring more time to answer (Segall and Moreno 1999). Related to the imposition of time limits, test developers and administrators must consider whether penalties should be imposed for incomplete tests. The necessity of such a penalty will depend on the particular scoring procedure applied to an assessment. For example, the Bayesian scoring procedure used in the CAT-ASVAB contained a bias such that a low-ability candidate could increase his or her score by answering the minimum number of items allowed, taking advantage of estimates that are too close to the population mean. Through a series of assessment simulations, the developers of the CAT-ASVAB settled on a penalty procedure that produces a final score that is "equivalent (in expectation) to the score obtained by guessing at random on the unfinished items" (Segall and Moreno 1999, p. 48). For more details on the penalty procedure implemented in this example, and time limit considerations for adaptive tests, see Segall and Moreno (1999) as well as Schnipke and Schrams (2002).

Building and adhering to the assumptions of item selection and scoring algorithms for adaptive tests can be a complex endeavor requiring extensive resources. Thus, it is important that the items and test format are designed and administered in accordance with measurement goals and the ultimate test purpose (Luecht and Clauser 2002). For simulation-based selection assessments, test developers must consider whether single item IRT delivery and scoring methods adequately fit the nature of the measurement experiences and responses captured through complex computer simulation examinations. As discussed in other chapters in this book, complex interactive exercises, such as those found in computerized simulation assessments, are a potential source for vast amounts of data. Luecht and Clauser (2002) discuss scoring methods for various complex computerized tasks (e.g., correcting embedded errors in an essay passage, producing mathematical expressions that represent a stimulus relationship, and managing patient information through data entry on order sheets) and the challenges of modeling the raw data for such complex tasks. Evident from this discussion and others (e.g., Wainer and Kieley 1987; Wainer et al. 2006) is that simulations containing complex computer-based tasks may be best suited for adaptive testing models that allow for modeling data representing multidimensional abilities and/or account for associations among subtasks.

Wainer and Kieley (1987) first introduced the concept of "testlets" as "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (p. 190). Wainer et al. (2007) proposed the use of testlets as the unit of construction and analysis in computer adaptive tests to alleviate such difficulties as context effects, item ordering, and content balancing which exist in most traditional algorithmic methods of test construction. Broadly, Wainer and Kieley (1987) propose that testlets can help in two ways: first, by allowing the test developer to recover more control over the test structure that is relinquished with automatic test construction algorithms; and second, by increasing fairness, such that scores for candidates of similar proficiency will be derived from tests of very similar content.

Testlet-based designs allow for the measurement of knowledge in several different content areas (Wainer and Kiely 1987; Wainer et al. 2007); they also allow interdependencies among sequential items referencing single stimuli (occurring, for example, in such situations that involve a large stimulus with several follow-up items; Wainer et al. 2007). Testlet-based designs are useful for allowing for sets of items with multiple response types within one test (Wainer et al. 2006), as is typical of many simulation-based selection assessments.

Developers of adaptive simulation-based assessments should also explore modeling options designed with the intent of allowing for measurement of multidimensional attributes and models that allow for generation of profiles of multiple components related to subtasks and sub processes. Several examples of such complex adaptive tests are provided in Williamson et al. (2006) as well as Segall (2010). In addition, Mulder and van der Linden (2010) discuss test-modeling options for multidimensional adaptive tests in detail.

5.4.4 Potential Future of Scoring Methodologies

Another area of interest to practitioners and academics alike deals with the vast potential for advanced and adaptive scoring methodologies for prehire simulations. Instead of simply combining and weighting the responses given in an assessment to yield a final score, more advanced methodologies may utilize detailed theoretical models that attempt to explain how a particular response to a simulation stimuli relates to other responses in the assessment as well as in other criteria.

Although much research has been conducted on this topic in other fields such as training evaluation, this is a vastly unexplored area to this point in scoring simulations for selection purposes. The ever-improving realism and complexity of prehire simulations, however, will likely push this research area to the forefront of the assessment field in the very near future. Indeed, one of the most exciting promises held by ever-improving simulations is the move from self-report, generic assessments to more performance-based, autonomous, and customized experiences for participants. As this move occurs, it is vital to be able to still gather information about higher level, complex, and even sometimes abstract abilities of candidates based upon observations of the lower level, concrete behaviors they perform within these high fidelity environments. Everything from the length of time a candidate spends doing something in a simulation to the amount of information they gather before acting in some manner within the simulation to their exploratory patterns can be of value in telling us more about a candidate, yet without improved scoring systems, much of this information may be wasted due to inadequacies in data capture technology. The end goal of complex scoring systems in simulations is to obtain similar, if not better, levels of scoring as could be obtained through having experts observe candidates as they complete the simulation, but through an automatic process that allows this to occur at a fraction of the cost. This section will continue with some examples of complex scoring of automatic tasks currently being utilized as well as an overview of the process utilized to develop these complex scoring systems. Exciting opportunities for future research in this area will also be examined.

Prehire assessment simulation developers would be well-suited to delve into the training field, where increasingly realistic, complex, and autonomous simulations have made it imperative to develop new assessment methods to ensure the training was successful. One example of this can be seen in the work of Koenig et al. (2010) who developed a theoretical framework for assessing performance in a simulation based on a naval ship and then followed up their theoretical framework with a computational design that incorporated elements learned from their initial attempts (Iseli et al. 2010; Koenig et al. 2010). The authors' two reports revolving around the naval simulation and its scoring are excellent resources providing details about the development of complex scoring systems for games in which examining a person's score falls short of adequately describing their performance. It is beyond the scope of this chapter to delve into all the intricacies described by these authors; however, we will provide a brief overview of vital steps in the assessment development process for a game-based simulation.

Koenig and colleagues used a preexisting 3D simulation of a naval ship to assess situational awareness as well as knowledge of how to deal with fires and floods (Iseli et al. 2010; Koenig et al. 2010). When utilizing a preexisting scenario, the steps that must be taken to create a valid scoring approach include: (1) the use of various specification editors to determine the domain represented by the game; in essence, this step defines what is being measured at different levels in the simulation; (2) the creation of an ontology development process, involving the definition of the domain and elements within the domain as well as the creation of element equivalence classes and the definition of relations, both within and between categories of objects defined earlier; this step involves specifying the theoretical model and showing relationships between variables; (3) the generation of a Bayesian Network to create a graphical relationship representing probabilistic relationships between variables; (4) the development of analysis tools based on the Bayesian Network; and (5) the choice of output generation tools and procedures to convey scores on pivotal aspects of the simulation to key stakeholders.

Although this type of modeling approach has most commonly been used to predict training outcomes in the past, it is hoped that as automated simulation design becomes more advanced our field may learn from these techniques. Ultimately, these techniques may help us move away from asking respondents to self-report their own psychological characteristics to enabling us to infer these characteristics directly through the decisions they make and behaviors they exhibit in the simulation. However, we do not necessarily expect modeling approaches like these to result in substantial enhancements to validity, though they may certainly add value. Rather, the primary benefit may be to better predict various latent constructs and develop our understanding of effective decision making.

5.5 Conclusion

Technological advancements can be plotted on an exponential growth curve, and in the early twenty-first century, civilization is at the knee of the curve (Kurzweil 2005). Vast amounts of information coupled with improving analytical tools will enable our predictive capabilities to asymptote. As a society, we are rapidly approaching a point where everything that can be known and understood will be. Simulations are also at the knee of their curve, and will soon peak as synergies between programming technology and the big data movement are realized.

There is substantial complexity in how simulations can be scored; and, since simulation design is evolving so rapidly, it is difficult to fully research any one technique. As our ability to simulate the real world in software evolves, we face the potential of being able to migrate from asking questions to observing behavior. This is a daunting possibility. Most of our psychometric history has been built on analyses of questions and responses and not on behavioral observations. There has simply been no automated way to observe true behavior until now. For the first time, realistic virtual scenarios are being used to simulate real life—giving job candidates the opportunity to demonstrate what they would *do* and not just what they *say* they would do.

The psychometric implications of this shift are vast. Self-report questions are conceptually simple: the researcher asks a question and the respondent answers it, resulting in a clean data point that may be linked directly to an underlying construct, which may be linked empirically and rationally to various criteria. Simulations allow us to ask job candidates to actually perform tasks—to brainstorm, multitask, respond to various stimuli, and much more. The responses they give are one type of data point; but we may also record and study their interactions with the virtual environment itself. For example, what can we learn from errant mouse clicks, repeated plays of the instructions, thorough viewing of different tabs of information in a problem-solving scenario, or response latencies on untimed exercises? In other words, we are increasingly able to capture indicators of style, and at least with our clients, we have found many of these indicators to be excellent predictors of real-world style and results. Conceptually, this shift is important: we are no longer looking at how a theoretical construct (e.g., Extraversion) predicts an emergent outcome (e.g., sales results); instead, we are concerned with what emergent prehire behavior can teach us about a new hire's emergent posthire behavior. Moving from self-report to emergent behavior measurement constitutes a fundamental advancement, and gets researchers one step closer to real world behavior.

However, as we have discussed, a potentially larger shift in predictive power will come from the convergence of big data and combinatorial scoring. Rather than relying on one particular item or scale construct, and expecting that lone data point to predict a meaningful outcome, modern simulations allow us to easily combine diverse item types and scales, while bigger data sets allow us to more easily examine interactions among these elements. We believe that these developments will create a golden age of assessment prediction power.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free response writing tests. *Applied Psychological Measurement, 12,* 117–128.

Arthur, W. R., Day, E., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*(1), 125–154.

Attali, Y. (2004) Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.

Bejar, I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76,* 522–532.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67–91). Mahwah: Lawrence Erlbaum Associates.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open ended versus multiple-choice response formats—It does make a different for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 385–395.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores a literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*(4), 689–709.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*(1), 83–117.

Clauser, B. E., Ross, L. P., Clyman, S. G., Rose, K. M., Margolis, M. J., Nungester, R. J., Piemme, T. E., Chang, L., El-Bayoumi, G., Malakoff, G. L., & Pincetl, P. S. (1997). Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment. *Applied Measurement in Education, 10,* 345–358.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 165–192). Mahwaj: Lawrence Erlbaum Associates.

De Corte, W. (1999). Weighing job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology, 84*(5), 695–702.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*(5), 1380–1393.

Doverspike, D., Winter, J. L., Healy, M. C., & Barrett, G. V. (1996). Simulations as a method of illustrating the impact of differential weights on personnel selection outcomes. *Human Performance, 9*(3), 259–273.

Drasgow, F., & Olson-Buchanan, J. B. (1999). *Innovations in computerized assessment*. Mahwah: Lawrence Erlbaum Associates.

Dudley, N. M., & Cortina, J. M. (2008). Knowledge and skills that facilitate the personal support dimension of citizenship. *Journal of Applied Psychology, 93*(6), 1249–1270.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94*, 334–346.

Fielding, N. G., & Lee, R. M. (1998). *Computer analysis and qualitative research*. Thousand Oaks: Sage.

Flaugher, R. (1990). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 41–64). Hillsdale: Lawrence Erlbaum Associates.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah: Lawrence Erlbaum Associates.

Gatewood, R., Field, H. S., & Barrick, M. (2010). *Human resource selection*. Mason: Thomson South-Western.

Gibby, R. E., Biga, A., Pratt, A., &, Irwin, J. (2008). Online and unsupervised adaptive cognitive testing: lessons learned. Paper presented at the conference of the Society for Industrial and Organizational Psychology, San Francisco.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance to test systems to small-scale cheating. *International Journal of Testing, 9,* 283–309.

Hambleton, R. K. (2002). New CBT technical issues: Developing items, pretesting, test security, and item exposure. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 193–203). Mahwah: Lawrence Erlbaum Associates.

Hanson, M. A., Borman, W. C., Mogilka, H. J., Manning, C., & Hedge, J. W. (1999). Computerized assessment of skill for a highly technical job. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 197–220). Mahwah: Lawrence Erlbaum Associates.

Hattrup, K. (2012). Using composite predictors in personnel selection. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 297–319). New York: Oxford University Press.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self- narratives. *Journal of Research in Personality, 43*, 524–527.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer testing. *Education Measurement: Issues and Practice,* 16–25.

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automatic assessment of complex task performance in games and simulations (CRESST Research Report No. 775)*. Los Angeles:

National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.
Kingsbury, G., & Weiss, D. J. (1983). A comparison of IRB-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 257–283). New York: Academic Press.
Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulations. (CRESST Research Report No. 771)*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for Studies in Education, UCLA.
Kurzweil, R. (2005). *The singularity is near*. New York: Viking.
Lee, H. S., & Cohn, L. D. (2009). Assessing Coping Strategies by Analyzing Expressive Writing Samples. *Stress and Health, 26*, 250–260.
Lord, F. M. (1970). *A theory of test scores. Psychometric monograph no. 7*. Princeton: Educational Testing Service.
Lord, F. M. (1977). A broad range tailored test of verbal ability. *Applied Psychological Measurement*, 95–100.
Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates.
Luecht, R. M., & Clauser, B. E. (2002). Test models for complex CBT. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 67–88). Mahwah: Lawrence Erlbaum Associates.
McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). New York: Academic Press.
Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback–Leibler information. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer Science+Business Media.
Mumford, M. D., Zaccaro, S. J., Harding, F. D., Jacobs, T. O., & Fleishman, E. A. (2000). Leadership skills for a changing world solving complex social problems. *The Leadership Quarterly, 11*, 11–35.
Newman, D. A., Jacobs, R. R., & Bartram, D. (2007). Choosing the best method for local validity estimation: Relative accuracy of meta-analysis versus a local study versus Bayes-analysis. *Journal of Applied Psychology, 92,* 1394–1413.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351–356.
Parshall, C. G., Harmes, C., Davey, T., & Pashley, P. J. (2010). Innovative items for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer Science+Business Media.
Pyburn, K. R., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61*(1), 143–151. doi:10.1111/j.1744-6570.2008.00108.x.
Rasch, G. (1960). *Probabilistic model for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*(2), 3–21.
Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*(3), 707–721. doi:10.1111/j.1744-6570.1997.tb00711.x.
Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*(5), 719.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 237–266). Mahwah: Lawrence Erlbaum Associates.

Segall, D. O. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer Science+Business Media.

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 35–65). Mahwah: Lawrence Erlbaum Associates.

Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In *Annual National Council on Measurement in Education Meeting* (pp. 14–16).

Smith, P. (2002). *Developing composite indicators for assessing health system efficiency. Measuring up: Improving the performance of health systems in OECD countries*. Paris: OECD.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Tesch, R. (1990). *Qualitative research: Analysis types and software tools*. London: Farmer.

Tonidandel, S., Quinones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87,* 320–332.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22,* 195–211.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W Glas (Eds.), *Elements of adaptive testing*. New York: Springer Science+Business Media.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wainer, H., Brown, L. M., Bradlow, E. T., Wang, X., Skoupski, W. P., Boulet, J., & Mislevy, R. J. (2006). An application of testlet response theory in the scoring of a complex certification exam. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 169–199). Mahwah: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–202.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah: Lawrence Erlbaum Associates.

Zaccaro, S. J., Mumford, M. D., Connelly, M. S., Marks, M. A., & Gilbert, J. A. (2000) Assessment of leader problem-solving capabilities. *The Leadership Quarterly, 11,* 37–52.

Zickar, M. J., Overton, R. C., Taylor, L. R., & Harms, H. J. (1999). The development of a computerized selection system for computer programmers in a financial services company. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 7–33). Mahwah: Lawrence Erlbaum Associates.

Part II
Simulations in Action

Chapter 6
Designing and Implementing Software and Computer Skills Simulations

Renee Barr and Christopher Coughlin

Most employers intuitively look to simulations to assess computer skills. Employers looking to assess 'computer skills' are typically looking for one of two types of assessments—general assessments of computer literacy or specific assessments focused on a particular piece of software. In addition, there is a wide range of skill level assessed, from basic (*Do you know how to save a file to a folder?*) to highly advanced (*How do you design secure networks for a global company?*). However, the bulk of the market for simulated computer assessments is in measuring basic computer literacy and knowledge of common software programs for high volume, relatively low skilled hiring. Accordingly, this chapter will examine how simulations assessing general computer literacy skills and specific skills using common software packages (e.g., Microsoft® Office) are developed, validated, and implemented (Fig. 6.1).

Simulations that focus on basic computer literacy or functioning largely involve assessing the candidate's ability to manage computer files and folders, access and use different software applications, navigate through an operating system, utilize the Internet, understand computer terminology, format and print documents, enter data, and perform other common work functions. Unlike simulations developed to assess a candidate's basic computer literacy, simulations for software programs are designed to assess a narrower skill set with a very specific program. Most of the common software programs used by the average worker come in a packaged group or office suite in which the different applications complement each other. In these office suites, there are usually a few programs that are widely used. These established programs include word processing software, database management systems, spreadsheet applications, email clients, and presentation programs. Other programs included in these productivity suites are not utilized as often by the average user, but still appeal significantly to certain workplace communities (e.g., graphics editors, collaborative software).

R. Barr (✉) · C. Coughlin
CEB, Alpharetta, GA, USA
e-mail: Renee.Barr@shl.com

C. Coughlin
e-mail: Chris.Coughlin@shl.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_6,

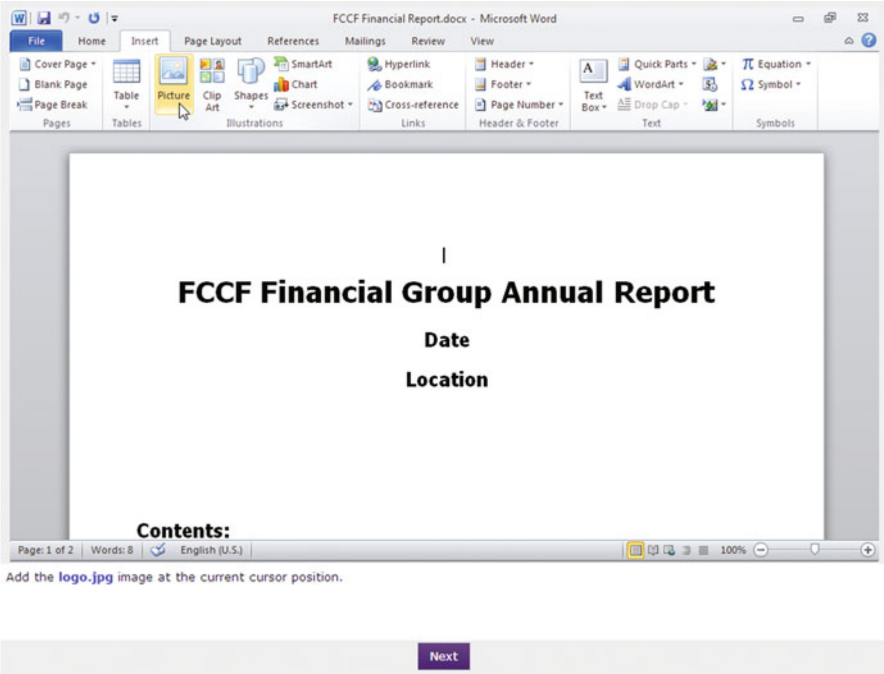


Fig. 6.1 Sample item from an MS Word® simulation

The most prevalent productivity suite, Microsoft® (MS) Office, has been on the market for almost 25 years. In the summer of 2012, Microsoft® announced that their Office suite had over a billion installations on computers across the globe (Sawers 2012). Also, the Office suite dominates the productivity software industry by controlling 90–95 % of the market share (Team 2013). According to Microsoft®, a consumer buys a new copy of the Office 2010 suite every second somewhere on earth (Arghire 2012). As indicated by its expansive reach, MS Office is commonly used by a variety of businesses because of its reasonable cost of volume licensing fees (i.e., five or more users), copious learning resources for the software itself, and the ease of working with other individuals and businesses who also utilize the Office software. Businesses use Office irrespective of their size or revenue. For example, a small "mom-and-pop" shop is just as likely to use Word, PowerPoint®, or Excel® as is a publicly traded conglomerate. Thus, it is not a surprise that simulation developers target the Office suite when building assessments for learning, training and development, and employee selection applications. In order to provide examples common to software simulations and readers, this chapter will utilize MS Office.

This chapter will review each step of the software simulation development and deployment process, taking care to note how the needs of each stakeholder (test provider, candidate, staffing clients, and corporate clients) influence the process. The development and deployment processes are broken down into four steps: design of the simulation interface, design of the simulation content, validation of the assessment, and implementation of the assessment.

6.1 Designing Software Simulation Interfaces

When designing software simulations, the importance of replicating the 'real' experience is paramount and potentially more critical than other types of simulations. Software simulations are job knowledge tests at their core, designed to measure how well a candidate knows a particular program (windows, Internet Explorer®, Excel®, etc). As a result, accurate measurement depends on the test, faithfully simulating the actual program. Further, test users have a strong preference for software tests that look and feel like the actual program. However, the test must not only look like the software, it must also behave like the software, allowing for all the same functionalities as the actual software. Any instance where the test does not mirror the appearance or functionality of the actual software should be addressed in the instructions and tutorials at the beginning of the test.

Replicating the look and feel of the software in the test environment is critical for multiple reasons. Because the testing medium and the tool used to demonstrate software knowledge is one and the same, there is an intuitive expectation of a high fidelity test on the part of test users. This correspondence between test medium and the knowledge being assessed also means there are measurement implications that result from using low fidelity tests. When the actual functioning of the software cannot be replicated in the test environment, the test instructions provide the best opportunity to inoculate test users' expectations and minimize measurement error.

When evaluating a MS Windows® simulation, both test users and candidates expect the test to look and function just like MS Windows® does. Users of most types of simulations would like them to be as realistic as possible, but in the end, strict high fidelity does not end up being the most important criteria. While users of a customer contact simulation may wish, it looked and functioned just like the proprietary software system they use on the job, they can usually be convinced that replicating their custom software would be cost prohibitive and that it is really most important to measure a candidate's ability to use that *type* of software effectively. A new hire in a customer contact position will most likely receive weeks of training on the proprietary software once hired, making the ability to learn a software program more important to successful selection. However, software simulations are actually intended to measure a candidate's knowledge of a particular program. Lack of fidelity to the actual software intuitively feels inadequate to test users and candidates, and there are real measurement concerns that support this impression.

Using a software program is a very visual experience. The location and appearance of menus, icons, and data entry fields are fundamental to how these programs are used. When a user hovers over a menu, clicks on an icon, or enters information into a field, the program reacts the same way every time. A simulation that does not accurately replicate these characteristics leads to increased measurement error and negative candidate reactions. During the test, they are preoccupied by worries regarding these differences and may alter how they ‘use’ the system to answer questions as a result. After the test, they do not feel like they had a fair opportunity to demonstrate their knowledge of the program, and they are correct.

While maximizing face validity is an important goal, it is not feasible to develop a simulation that corresponds exactly to each candidate’s experience with the software. In recent versions of office suites, individual users are allowed and often encouraged to customize the program with elements that will better facilitate their individual work goals. For example, an MS Word 2010 user is able to add buttons (e.g., E-mail, Quick Print) or remove the default buttons (e.g., Save, Undo) on the Quick Access Toolbar so they can easily locate and perform their most frequent activities. While this customization results in higher satisfaction with the application and allows for quicker completion of particular software tasks, it presents a problem for simulation developers trying to create an interface that mimics exactly what each candidate views when using their version of the program at home or work. This issue can be mitigated by establishing a simulation model that replicates the default program made by the software proprietors. While the customized interface features cannot always be replicated for every candidate, the default settings are usually established as such by the software developers to maximize the benefits to the general software user.

Elements of the display size (i.e., screen resolution, aspect ratio) are another set of factors impacting developers’ ability to replicate each candidate’s unique experience with the software. In their own experiences with the software, candidates have a number of different display settings they might use to view the programs in accordance with their personal preferences. If a user has their monitor set at a display resolution of 1024×768 (in pixels) and views Word 2010 with this monitor setting, then the software program adjusts the interface to that particular resolution. So, a user who views Word 2010 with a screen resolution set at 800×600 sees a different Word interface than a user who works in Word with a screen resolution of 1920×1080. For example as seen in Figs. 6.2 and 6.3, the number and configuration of the options available on the menu bars change depending on the display size. This creates a limitation when designing one common interface for the simulation. To address this issue, the simulation developer might construct the interface at or a size below the most commonly used resolution, as indicated by data on the current usage share of different screen resolutions.

Simulation developers not only need to consider candidate expectations regarding display size, they also must take into account how the simulation will be displayed when administered to candidates. The software simulations usually use another platform to create the simulated interface and item-specific content. These software applications, designed to deliver rich internet applications or animation, are used to mimic the component (e.g., Word 2007) precisely so the candidate sees it as the software application being measured. Whichever software (e.g., Adobe® Flash®, Microsoft® Silverlight®, HTML5) is used to create the simulated interface, the simulation developer must ensure that it is supported by all web browsers (e.g., Internet Explorer®, Mozilla Firefox) and, if the software isn’t supported fully, then finding out if the limitations will impact the simulation is essential in assuring a quality simulation.

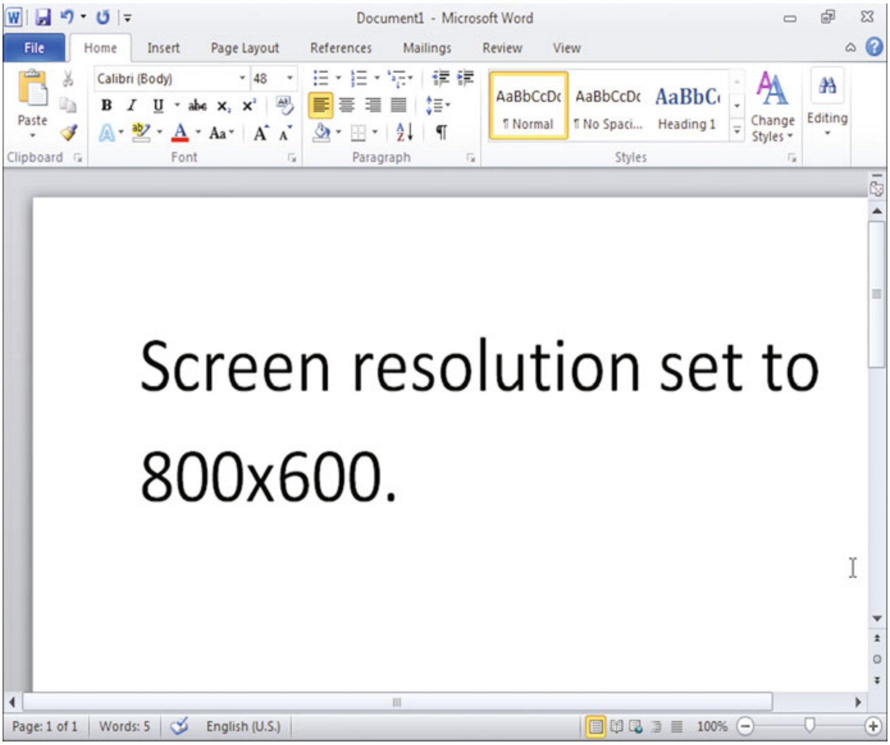


Fig. 6.2 View of MS® Word when screen resolution is set to 800 × 600. Note how many menu options are minimized or collapsed

Test users prefer software simulations to multiple-choice software tests because most people do not memorize the layout and functioning of a program by rote. A notable exception is syntax-based programs where writing code is the primary means of interacting with the software program. Most solid intermediate level users of MS Excel® do not have the menu bars and pop-up windows memorized. This makes writing multiple-choice items that accurately reflect the candidate's ability to use the software challenging. The test ends up being a better measure of who has memorized a book on how to use the program rather than who actually knows how to use the program itself. You may know how to conditionally format cells in MS Excel®, but could you write step-by-step instructions without looking at the program? Most test administrators are not looking to identify software experts or trainers using assessments, they are looking to identify functional users. Simulations that do not quite replicate the real look of a program encounter the same issues as a multiple-choice test: candidates answer some items incorrectly, not because they do not know how to perform those tasks in the program, but because the test is confusing and does not reflect how the program is used on the job.

Most of today's popular software programs offer multiple ways to perform an action (e.g., dropdown menus, shortcut keys, entering formula or code). To accurately

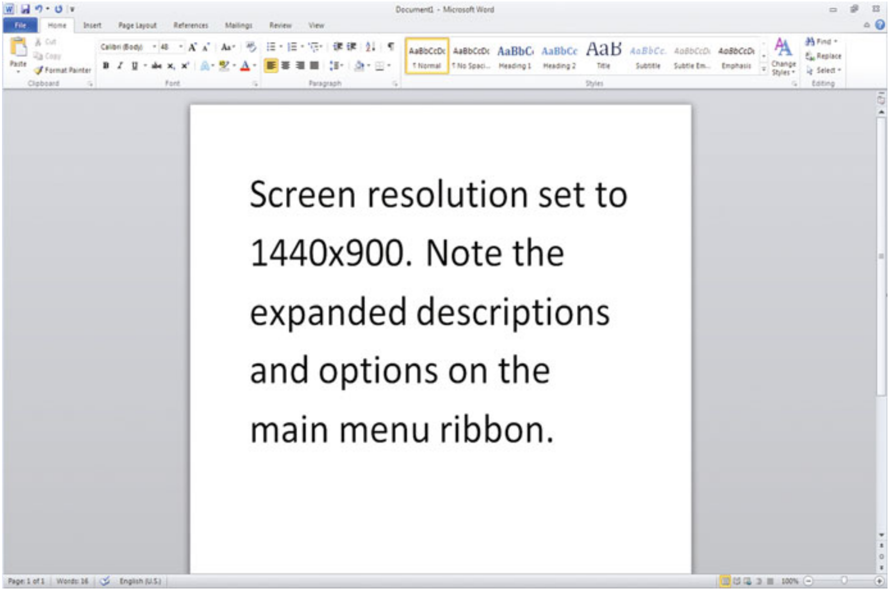


Fig. 6.3 View of MS® Word when screen resolution is set to 1440 × 900. Note the additional visible menu options

capture the look and feel of a software program in a simulated test, the simulation would ideally allow for items to be correctly answered in multiple ways. There is an added layer of complexity as each item essentially has multiple correct answers, but unless there is a job related reason for restricting the mode of correct answer, then allowing for these multiple correct responses is critical to accurate measurement of the ability to use the software on the job. For example, in MS Word 2010 there are multiple methods to bold a particular piece of text, as shown in Fig. 6.4. The three most common methods are probably to select the text and (1) use the shortcut keys 'CTL + B', (2) go to the home menu tab and click on the 'B' in the 'Font' section, or (3) right click and then left click on the 'B' in the mini toolbar above the shortcut menu. This is not an exhaustive list of options, and it would be difficult to assert than any one method is fundamentally better or a better indicator of knowledge of the program. The shortcut key is faster, but not so much as to impact the job performance of the average user. Given that most software simulations are not custom designed for a particular job or client, it would be challenging to determine the one best answer or most widely used method across test users. Finally, if a simulation does not allow for multiple methods of answering an item, then this would have to be explained in the test instructions or item stem.

Test tutorials and instructions do not make the simulation more realistic; instead, they can serve to minimize the negative impact of whatever aspects of the simulation that are not realistic. If there are limitations to how the simulation emulates the actual software, it is best to inform candidates of these discrepancies in the test instructions. This will prevent candidates from having to guess at the discrepancies and may

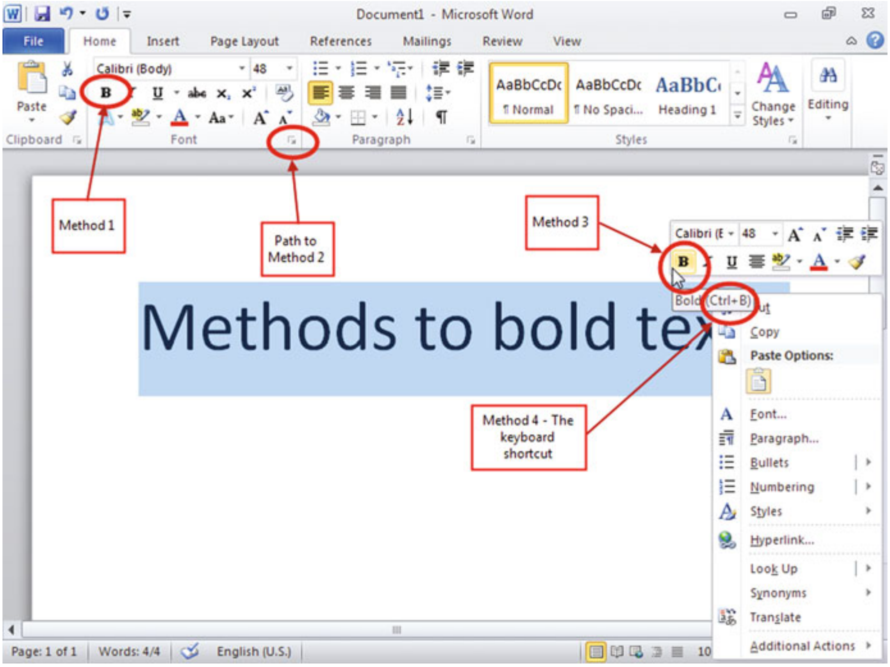


Fig. 6.4 Illustration of the multiple correct methods to bold text in MS® Word

even prevent any negative candidate reactions. Ideally, the forward instructions for a software simulation would include an interactive tutorial, allowing candidates to get a feel for the interface before being evaluated. If needed, the item stems themselves can include instructions that account for incidents where the simulation does not fully replicate the software's capabilities. For example, if the simulation cannot accommodate using shortcut keys to copy an object in PowerPoint®, the item stem could read 'Copy the smart art object on slide 2 and paste it on slide 3 without using shortcut keys.' Here the instructions start to bleed into the simulation's item content which is inexorably interrelated to the overall simulation interface. However, if an item stem requires numerous caveats in order to operate within the interface's limitations, then it would probably be best to find another item within the content domain that does not require caveats. Numerous caveats could make the instructions and/or item stems too confusing or verbose. Ideally, the interface and item stems complement each other to create an uncomplicated candidate experience.

6.2 Designing Software Simulation Content

When determining the content of a software simulation assessment, the ultimate goal is to identify and measure the knowledge that is most indicative of successfully using the software on the job. Achieving this aim requires simulation developers to balance multiple considerations. How is the software typically used on the job? Are test administrators interested in identifying multiple skill levels? What limitations does the software interface place upon the items that can be created? What influence do past or future versions of the software have on the content of the assessment currently under construction? Managing these sometimes competing issues will play a large role in determining how successful the assessment will be. Do test administrators like what they see? Do candidates feel they had a fair chance to demonstrate their skill? Is the test valid for its intended use? Does the test ultimately predict who will perform well on the job?

The first step in designing the simulation content is identifying the software knowledge needed on-the-job and organizing it into meaningful knowledge domains. Assuming the software simulation is being designed for general use across a wide variety of users rather than being custom-developed for a particular user or job, a wide net needs to be cast in order to understand how the software is used on the job. Focus groups or interviews with software users representing the scope of jobs and industries that will be using the assessment allow the simulation developer to broadly define the universe of test content that will be needed to be created. In addition to focus groups and interviews, the simulation developer could acquire information from software trainers, master users, teachers, and application proprietors to better understand the most common and important tasks. Many software programs have a larger range of capabilities than are commonly used on the job, and understanding which capabilities are actually relevant to job performance is critical. Often understanding the relevant capabilities corresponds to understanding the range of skill levels that are of interest to test administrators. Many test administrators are much more concerned with identifying individuals with basic or intermediate skills than they are with identifying expert users. Test administrators are much more likely to want to know if a candidate can print a specific area of cells from a MS Excel® worksheet, than if the candidate can write a macro. The simulation developer should come away from the focus group, interview, and information review process with a confident understanding of the most common and important areas of knowledge for the software in question.

This broad understanding of the important knowledge areas is then organized into knowledge domains with multiple skill levels. The knowledge domains can be thought of as test scales even if they are not actually scored and reported as individual scales. They are a framework for organizing the test content that can play an important role in customizing the test, validating the test, designing the test report, and using the test to make hiring decisions. The knowledge domains organize similar software skills together and often contain skills that are related, but with different expertise levels (i.e., basic, intermediate, and advanced). For example, a simulation for MS Excel® might have knowledge domains for entering data, formatting, and using formulas. The entering data domain might contain multiple skill levels: basic tasks, such as typing text directly into a cell; intermediate tasks, such as automatically displaying data from another cell on the same worksheet; and advanced tasks, such as automatically displaying data from a cell on another worksheet or workbook. When compiling this information, there are a number of resources that a simulation developer can locate and use to help organize information into knowledge domains and, subsequently, into subsets of expertise. Most software developers provide comprehensive training manuals, user forums, help guides, materials for certification testing, etc. that can direct this domain organization process. Also, if this is a new version of an already established software program, utilizing the content or topics of any previous software simulations could prove useful. Once the knowledge domains and skill ranges are created, they directly inform item development.

One of the most attractive benefits of using a simulated software assessment is the ability to ask a candidate to perform an action in the program. Simulated software assessment items become open-ended requests to demonstrate knowledge rather than closed multiple choice questions about the knowledge. From a measurement perspective, multiple choice questions have inherent guessing effects which are minimized in the simulation assessment format. Because of contextual software clues and the opportunity to attempt an item more than once, guessing effects are not eliminated in simulation-based assessments, but they are less problematic than an assessment where candidates can randomly select an answer from a group of four or five choices. Thus, from the candidate's perspective there are only item stems while the interface itself is one large distractor.

Writing good software simulation items is grounded in the same item writing fundamentals as any other type of item writing activity. Item stems should be clearly worded. They should be directly related to the knowledge domain they are supposed to measure and the targeted skill level. They should not be double-barreled unless you are intentionally asking the candidate to perform a multistep action. The primary constraint on software simulation item writing is the simulation interface and the technology behind it. The item must be answerable within the interface and the simulation must be able to correctly score the item response. If, in the interface design process, it was deemed too cost prohibitive to build certain aspects of the actual software program, then these areas should not be available to item writers.

The most important difference between traditional item writing and item writing for software simulation assessments is identifying the correct answer to an item. In traditional item writing, the item writer has the benefit of writing close-ended questions that define the universe of possible correct answers in the responses' options given to the candidate. The item writer should not have this kind of freedom when creating content for a robust software simulation. If the ideal simulation replicates the actual program's capabilities and the simulation developer wants to evaluate the candidate's ability to use the program in a natural manner, then an ideal item is open-ended with multiple correct responses. The item writer must identify every correct method of answering the item so that they can all be programmed as a correct response. If an item asks the candidate to cut and paste a piece of text, it is not enough for the item writer to identify a single method, or even the three most common methods, for performing this task. It is critical that the item writer identify each and every method for performing the task. Each method must then be programmed as a correct response. If the simulation interface cannot accommodate a correct method, then its robustness is lessened and the item stem must be written in a manner that tells the candidate not to take that action.

Identifying and implementing every plausible method ensures that the candidate can use atypical methods and still correctly complete the task given in the item stem. Most software developers will provide their preferred methods to complete a particular task, but, with a little research, the complete list can be identified for each task. Most productivity suites, when successful, will attempt to improve and expand on each version of the suite and create a bigger and better release in the future. Since these software applications have so many versions, a simulation developer must remember to incorporate legacy methods for certain tasks. Legacy methods, in this context, can be expressed as methods that worked in an earlier version of a software application, but have been abandoned by the proprietor to focus on the preferred, newer method to complete this task. However, these legacy methods usually work in every future version of a program. For example, a method that was emphasized for Microsoft® Word 97 might not be mentioned as a viable method to complete the same task in Word 2010, but, it will still usually work in the latter version. Discovering these legacy methods might be difficult for the simulation developer, but these are necessary for software users who developed the skills to complete the task years ago and allowed these skills to generalize to subsequent versions of the software.

After determining all the correct methods for each question, the simulation developers are also tasked with addressing which elements of the assessment should be scored as incorrect. Usually any individual action that is not advancing the simulation towards completing the task in the question stem is scored as incorrect. However, there are some actions in the software that will elicit a neutral response. These neutral responses (colloquially known as "dead elements") to certain candidate actions are established because some methods are not necessarily directly completing the task assigned in the item stem, but can be viewed as innocuous, ambiguous, or indirect to the task at hand. If these method uncertainties can not be resolved in the assessment instructions or item stem, then it is best to have nothing happen when the candidate attempts to perform this neutral action. For example, imagine that a candidate is asked to change the font of some selected text in Word 2010 and they opened up the Font dialog box to complete this task. However, they do not like the positioning of the dialog box within the simulation and they attempt to move it via a left click and drag of the title bar. Since this is not necessarily an incorrect action, the simulation developer can make a left click of the title bar of the dialog box a dead element when programming the item and nothing will appear to happen within the simulation if and when the candidate attempts to left click this title bar. Thus, candidates are not penalized for committing a neutral action.

Another item writing consideration is ensuring that the item content does not stray away from the targeted software application. It is increasingly common for productivity suites to interconnect its component programs, allowing users to seamlessly move between programs to complete tasks. For example, MS Word allows a document to be sent via email as an attachment. When completing this task, the action is initiated in Word, which delivers the user to an opened email in the user's email client (e.g., MS Outlook) with the document attached. From here, the user must know how to use the email client—an entirely different program—to complete sending the document. This task becomes problematic as the basis for an MS Word item because it crosses the content domains of two separate programs. While MS Word and MS Office skills are likely to be positively correlated, this type of item makes for messy measurement and should be avoided. If this type of dual program task is essential to include in the assessment, then the item should only score the candidates' response where it resides inside the targeted program. Using the example of attaching a Word document to an email, only the candidate's action inside Word would be scored.

Item writing for the most popular software titles must often also take into consideration previous versions of the software. Programs such as MS Windows® and MS Office have had numerous versions since their inception in 1981 and 1989, respectively. Assessments have been built around those versions and organizations have conducted job analyses and validation work around those versions. It is advantageous to those organizations if the simulation built to assess the newest version of a program corresponds where appropriate to previous versions of the assessment. In practice, this typically means that much, but not all, of the new items directly or indirectly correspond to items in the simulation for the previous version. Updated versions of software programs are often produced to incorporate user feedback on common problems. Enhancements are carried forward in each subsequent version. When a software program has been updated many times, users' most common tasks were addressed long ago and the new enhancements rarely impact common business-oriented tasks. If the current version of the software program has completely new functionalities or no longer allows other functionalities, then the item content will need to reflect this. However, much of the time the same functionalities exist and are executed in a similar fashion as in previous version of a software program. Saving a file, opening a file, formatting, printing, and the like are important skills in each version of MS Office. It would not be unusual to have the same item stem across numerous versions. For example, the item stem "Save the open file in the 'Project XYZ' folder using the name 'Timeline'" would be applicable to any MS Office program released in the last decade. The interface, menu, and icons that the candidate would use to achieve this action might look a bit different across versions, but the item stem does not need to change. This correspondence of item content across versions helps to streamline the job analysis and validation work an organization must conduct when moving from older versions of the software simulation to a newer one.

Like any other assessment, once the items are written, a method of quality assurance (QA) should be applied to the written items and their corresponding answers or methods. The simulation developer can utilize software subject matter experts, super users, industrial and organizational (I-O) psychologists, etc. to certify that the items are appropriate for the simulation and fulfill the requirements set forth by the item writer (e.g., clearly worded, directly related to the knowledge domain). Confirming the items follow these standards and guidelines described earlier will allow the simulation developer to progress to the next steps in the development of the simulation, which is usually the tangible creation and programming of the simulation itself.

Unlike other assessments, once the simulation has been created, additional QA measures need to be applied to assure that the simulation works properly. Identifying that the response data for every item are captured accurately, the "right" methods and interface are programmed correctly, and the "wrong" or incorrect options and dead elements are coded accurately, this is essential to guarantee that the simulation adequately mimics the real software application and works properly. If the candidate is taking the simulation through an Internet browser, the simulation developer must ensure the software (e.g., Adobe® Flash®, MS Silverlight®) used to mimic the software being measured is not causing any interactive problems with the browser. Different browsers can interact with the simulation software and absorb some of the simulated functions of particular keyboard shortcuts. For example, if HTML5 and JavaScript® are used to create a simulation for MS PowerPoint® 2010, then the simulation developer must certify that all key press combinations for every item work properly and do not interact or become absorbed by all supported browsers or operating systems. Given the ever expanding number of commonly used web browsers, this QA testing can be lengthy in terms of time and content covered. These additional QA measures should be comprehensive and allow the simulation developer to identify and resolve any simulation defects before the assessment is released to the market.

6.3 Validating Software Simulations

Validating software simulations is not fundamentally different than validating any other type of knowledge or skill assessment. Test administrators must demonstrate the job relatedness of the test content through job analysis and validation. Like many knowledge and skill assessments, software simulations are more likely to be content-validated than criterion-validated. As such, job analysis is often the critical step in implementing a software simulation assessment that will have utility for the test user and minimize legal risk.

Job analysis plays the primary role in the implementation of software simulation assessments because it is the step that ensures the correct knowledge domains, skill levels, and versions of the software program are being targeted. These are the basic characteristics that must be evaluated to demonstrate the job relatedness of software simulations. Organizations routinely seek to assess candidates' skills on common software programs across a wide range of positions. Even when the organization is interested in a single job title (e.g., Administrative Assistants), the individuals with this title can sit in a range of locations and/or business units. The level of MS Excel® and MS PowerPoint® skills required for an Administrative Assistant may vary considerably depending on, if they sit in the Finance, Marketing, or Human Resources (HR) departments. It would not be uncommon to find Administrative Assistants in a location that opened in the last couple years using the most recent version of MS Office, while the Administrative Assistants sitting in an older location associated with a regional company acquired by the multinational last year are using an older version of MS Office. Job analysis indicates if it is appropriate for the same MS Office simulation assessment to be used across these positions.

Determining which version of a software simulation assessment is appropriate for use in an organization's selection process can be surprisingly complex. Ideally, an organization is using the most recent version of a software program throughout their organization, making the determination of which version of the assessment to use is straightforward and simple. Unfortunately, this is not the case for most test administrators. It is much more common for an organization to find that there are multiple versions of the software in use or that they are using an arguably out-of-date version of a software program. In the middle of 2012, many organizations were still using MS Office 2007 (Kelly 2012).

Both large and small time gaps between the most recent version of a software program and the version in use in an organization can make it difficult to determine which version of the assessment is most appropriate. When the gap is large, young or recently trained candidates may be particularly disadvantaged by continued use of the older version of the software assessment. These individuals are more likely to be experienced in the most recent version of the software program and less likely to have been exposed to the old version of the software program. A recent high-school graduate may score poorly on a MS Excel® 2007 assessment, not because they are inexperienced in using MS Excel® or unable to learn MS Excel®, but because they have only been exposed to the menu layout of MS Excel® 2010. However, recent software releases can disadvantage experienced candidates. For both monetary and logistical reasons, organizations are often slow to adopt new software releases. An experienced candidate may be currently employed at an organization using MS Excel® 2007. This candidate uses MS Excel® on a daily basis at an above average skill level. However, the position the candidate is testing for is at an organization that uses MS Excel® 2010 and this is the version of the simulated assessment the organization uses during the selection process. This experienced candidate is likely to score lower on the 2010 version of the assessment than she would have on the 2007 version. This is unfortunate for the hiring organization because, while this candidate might take a few weeks or months to adjust to the new version of MS Excel®, it is highly probable that she would again attain her previous skill level which is more indicative of her long term success in the position. Thus, when deciding which version of a software simulation assessment to implement, test administrators must consider the time gap between the most recently released version of the software program and the version currently in use, in addition to the job analysis results.

Once a version of the software simulation has been selected, it will ideally be validated. In practice, it is quite common for software simulations to be implemented with little or no job analysis and validation work. Perhaps because these assessments have tremendous face validity when they are well constructed, test users perceive there to be both minimal legal risk and minimal risk of implementing a test that has little or no utility. When software simulation assessments are validated, a content validation approach is almost always employed. Test administrators have very little appetite for criterion validation approaches to establishing the job relatedness of software simulation assessments. This likely arises from the perception that they are low risk assessments, but there is also a practical reason that criterion validation approaches are avoided. Concurrent criterion validation of software simulation assessments would be challenging in many organizations because the incumbent test scores would likely suffer from severe range restriction. When organizations seek to use a software simulation assessment in the selection process for a position, it is typically because that program is used on a daily basis in the position. Few incumbents of any meaningful tenure are retained in these positions if they are not performing adequately with the software. Organizations are also generally uninterested in predictive criterion validation strategies for these positions. When the perceived risk of implementing without criterion validation is so low, why take the very real risk of hiring a candidate who scores poorly on the assessment? When the perceived risk is so low, why expend any additional resources or time on criterion validation at all?

Because software simulations are designed to assess mastery of a content domain, content validation is arguably the most appropriate approach to take. Content validating software simulation is a very straightforward process, especially if the test content is already organized into knowledge domains of varying skill levels. Subject matter experts (SMEs) can be asked to rate the job relatedness (i.e., importance, frequency, etc.) of each knowledge domain and the skill levels within each. Actual items can be used as concrete examples of each domain and skill level to help SMEs make accurate ratings. For an even more thorough content validation effort, SMEs can rate the job relatedness of each individual item. Angoff, or a similar type of expectancy rating, can be collected to help inform the process of setting a qualification standard for the assessment. Between the job analysis and content validation efforts, test users should have all the information necessary to make sound decisions regarding how to implement the software simulation in their selection processes.

6.4 Implementing Software Simulations

With their high face validity and familiar subject matter, software simulation implementations can appear very straightforward at the outset of a project. However, the ubiquity of many software programs throughout an organization can present some unusual implementation issues. Deciding which test to use, how to administer it, and where to set a passing standard while maintaining job relatedness and consistency is often a surprisingly complex assignment from an HR process perspective.

6.4.1 Choosing a Test

Once an organization decides to implement software simulations, how does it decide which tests to use for a particular opening? Even if the organization has conducted a job analysis and content validation effort around the testing, this can be a challenging decision. For example, XYZ Corp. needs to decide how to implement their MS Office testing for administrative positions. Should XYZ Corp. require all candidates to take PowerPoint®, Excel®, and Word simulations? Doing so would simplify the process and ensure consistency in the assessments' use; however, it would likely result in some candidates taking tests not actually related to the specific administrative position for which they are applying. Administrative employees' use of PowerPoint®, Excel®, and Word likely varies by the department they sit in and the person they support. A similar issue can arise over which version of a test to use. Many organizations update their software in a piecemeal fashion, resulting in multiple versions being in use at the same time, which complicates ensuring both consistency and job relatedness when choosing a test. As a result, the person tasked with deciding which test to use for a particular opening can have several competing demands influencing what is the most job-related and consistent choice.

Determining who should be making the decision regarding which assessment to use is a key point of consideration. HR can make a centralized decision, facilitating consistency of process but potentially decreasing job relatedness. Alternatively, authority to decide which test to use can be given to hiring managers who are closer to the job. Allowing hiring managers to make the decision may increase job relatedness, but will almost certainly reduce consistency in the testing process. Finally, recruiters can be tasked with the decision in the hopes that they are in a position to better balance consistency and job relatedness. If given training and tools on how to make a testing decision, recruiters and hiring managers can be an attractive solution. However, decentralizing the decision in this manner creates a more complex process to manage and oversee.

The decision of which test to use and who gets to make that decision is not a "one size fits all" decision. In general, the more confidence an organization has in the judgment of those in the field and the less legal risk an organization faces, the more likely it is that these decisions will be decentralized and somewhat customizable to the requisition being filled. If these decisions will be decentralized, it is imperative that decision aids are created to help ensure their job relatedness and consistency. Some simple decision tree type documents outlining the factors to consider and summarizing the content and skill level of the available test options can make a large impact on how well these processes are executed upon implementation.

6.4.2 Administration Considerations

Once an organization has decided which tests to use, it then has to consider a variety of administration factors. How many chances does a candidate have to answer each question? Can the candidate take the test again? Will the test be timed? Will the test be administered remotely? Because the test in question asks candidates to demonstrate skills, the answers to these questions can be somewhat different than expected. While most selection assessments do not allow candidates multiple opportunities to answer a question, it is not uncommon for software skills assessment to do so, and it is not as great a threat to test measurement as it might seem. The bulk of the market for software simulation assessments consists of MS Office and basic computer literacy. In addition, most of the jobs in question do not require expert skill. Instead, most of the jobs in question require good working knowledge of the program. At this skill level, it is not uncommon for someone to click through a couple of menus or taskbars to remember how to perform a task of low frequency or intermediate skill. Thus, it is not unreasonable to allow candidates two attempts to perform a task. When one considers that most software simulations consider a response to be incorrect if a candidate makes one click outside of a correct path, it is arguably more reflective of on-the-job characteristics to allow two attempts, than to allow only one. In addition, it is quite easy for candidates to make inadvertent clicks during a test session. Scoring these inadvertent clicks as incorrect responses would only add error into the skill measurement.

Test-retest policy decisions for software simulations most closely reflect best practices around general job skill and knowledge assessment. Organizations should avoid applying test-retest standards from cognitive and personality assessment to software simulations. Because a candidate can improve his or her MS Office and basic computer literacy skills with a short training course or self-study with an online tutorial or book, it is reasonable to think that candidates could improve their skill level in a short amount of time. The feasibility of quick skill improvement would support a relatively short retesting window, perhaps only 4–6 weeks. If the software simulation is adaptive and candidates do not see the same items each time they take the test, then it is easier for organizations to support a short retesting policy. However, if the software simulation is static and the same items are presented at each test session, then concern over candidates becoming too familiar with the test content becomes a legitimate issue for organizations.

The advisability of making a software simulation timed is determined by the nature of the timer and the evidence for implementing it. If job analysis indicates that a software simulation needs to assess for expert knowledge or the ability to complete tasks in a high volume or time sensitive manner, then implementing a speeded test may make sense. In this situation, the time limit on the test puts pressure on the candidate and not all candidates are expected to complete the assessment. The test instructions should tell the candidate that there is a time limit and that many candidates do not complete the assessment. More commonly, organizations want to implement a timer on a software simulation to prevent candidates from spending far too much time making their way through the test. The intent is not to rush the candidate, but rather to ensure the process is completed in a reasonable amount of time. In this situation, the time limit should be set such that the vast majority of candidates have more than enough time to complete the assessment. Organizations or test providers can examine historical data on test time to identify the time point where all but the outliers have completed the assessment. The test instructions should indicate that the test is timed and that most candidates are able to comfortably complete the assessment within the allotted time.

In today's market, most software simulation assessments for employee selection are administered remotely. Software simulations have right and wrong answers that candidates could look up in a remote setting, traditionally an argument for proctored testing. However, software simulations are most typically administered for high volume entry level positions where the cost of high touch proctored assessment is not justified by the relatively low risk of cheating. In addition to the aforementioned benefits of having a test timer, a reasonably set time limit can deter a candidate from cheating. A reasonably set timer creates enough concern about having enough time that it discourages candidates from spending their finite testing time trying to research each item via the Internet. If an organization was highly concerned about cheating on these types of assessments, re-administering the assessment in its original or shortened form to candidates at the time they are interviewed in person would probably be the most economical course of action.

6.4.3 Setting Qualification Standards

Finally, organizations must decide how to score software simulations in the assessment process. Should the organization set a single pass/fail point? Should the pass/fail point depend on the amount of skill required by the job in question? Some of the complexity in this decision will be influenced by how an organization has decided to go about determining which test is administered. Is that decision centrally made by HR, or is it delegated to the recruiters or hiring managers? If the test is determined centrally, it is more likely that a uniform pass/fail qualification standard can be set and will be adhered to throughout the organization. However, there is some argument for not setting a single pass/fail qualification standard, and it follows much of the same reasoning reviewed earlier in the context of deciding which test to use. If the type and level of software skill needed varies greatly from job to job in the organization, then a single qualification standard may not be useful. It may be more effective to provide decision-making tools for the recruiter and/or hiring manager to help them determine what skill level the role demands and then apply that standard to candidates for the job in question. Test reports that indicate skill level in different areas may provide more useful information than a pass/fail report.

6.5 Summary and Conclusion

Basic computer literacy and familiarity with common office software is necessary for a wide range of jobs, and employers want to assess candidates' skills in those areas prior to hiring them. Software simulations are a natural solution to this business need. Like any skills assessment, a strong content validation approach is important during both development and implementation of these assessments. Because these assessments measure computer skill and are administered via computer, there are many technology considerations that must be accounted for in both development and implementation. Successful implementation of software simulations as part of a selection process requires careful consideration of which skills to measure, at which skill level, and in which version of the software. In addition, setting uniform qualification standards for these assessments is often more difficult when compared to personality or cognitive assessments. However, software simulations have substantial face validity and are generally well received by both candidates and employers.

References

Arghire, I. (2012, July 10). Microsoft's Office has over one billion users. Softpedia. http://www.news.softpedia.com. Accessed 8 Dec 2012.

Kelly, H. (2012, July 16). Can a new Office suite keep Microsoft on top? CNN. http://www.cnn.com. Accessed 2 Dec 2012.

Sawers, P. (2012, July 9). Microsoft: 1.3 billion Windows users, Office is on 1 billion desktops (and more stats). The Next Web. http://thenextweb.com. Accessed 12 Dec 2012.

Team, T. (2013, January 9). An overview why Microsoft's worth $ 42. Forbes. http://www.forbes.com. Accessed 12 Dec 2012.

Chapter 7
How to Measure Contact Center Skills Using Multimedia Simulations

Brent Holland and Dawn Lambert

7.1 Introduction

Large companies frequently rely on contact centers to address consumer demand for quick and efficient access to information, problem resolution, and to communicate the value of new products or services. Contact centers are fast-paced, technologically sophisticated businesses that route inbound customer inquiries and outgoing solicitations to employees anywhere in the world. Most contact center jobs require employees to interact with customers, whose demeanor ranges from friendly to hostile, while navigating a complex array of databases that provide access to customer and product- and/or service-related information. Contact center employees usually perform these activities under time pressure and with sophisticated systems monitoring their communication style, reliability, and performance. The combination of job complexity, speed, oversight, and an endless stream of customers causes significant psychological strain that overwhelms ill-suited and maladroit employees, leading to burnout, absenteeism, and attrition.

The contact center industry has turned to realistic multimedia simulations, which allow job candidates to play the part of a fictitious contact center representative to evaluate a candidate's contact center skills (e.g., computer, multitasking, and data entry). Leading simulations have evolved from quasi-interactive situational judgment tests (SJTs) to microcosms of modern centers, complete with training, interactive dashboards, timers, and branching that allows candidates to escalate or deescalate a customer's emotional response based on the skill with which they manage the interaction. Asking a candidate to play the part of a contact center representative creates an engaging experience, presents a realistic preview of the job, and provides

B. Holland (✉) · D. Lambert
FurstPerson, Inc., Chicago, IL, USA
e-mail: Brent.Holland@furstperson.com

D. Lambert
e-mail: Dawn.Lambert@furstperson.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_7, © Springer Science+Business Media New York 2013

a company's talent acquisition team greater visibility into the candidate's likelihood to perform well in training and on the job.

This chapter discusses the design, development, deployment, and business impact of interactive, multimedia contact center simulations. Before addressing the role of simulations, however, it is important to understand the contact center industry and the complexity of its jobs and environment.

7.2 A Brief History of Contact Centers

Contact centers are an efficient and cost-effective alternative to in-person interaction with consumers. A contact center is a facility or home-based office from which a company's representatives interact with consumers via telephony and other technologies (e.g., e-mail and instant messenger) while accessing databases to help answer questions, solve problems, solicit new business, and myriad other activities (Holman 2003; Merchant 1998). The terms call center and contact center are often used interchangeably, but they are not identical. A call center refers to an environment in which representatives use telephony-based technology to interact with customers whereas a contact center, a term we will use throughout the remainder of this chapter, describes modern environments that use multiple channels to facilitate interaction with consumers. A contact center is typically invisible to the consumer.

Many contact center environments and jobs reflect the application of scientific management principles to white collar work (Russell 2008). As described by Bagnara (2000), the first contact center appears to have opened in the late 1960s in response to a Federal court order requiring Ford Motor Company to create a toll-free phone number to streamline a vehicle recall. In these early centers, jobs required limited knowledge, skills, and autonomy (Callaghan and Thompson 2001), but management also maintained constant oversight of representatives' performance. This organizational structure caused stress (Taylor et al. 2003) and disengagement (cf. Holman 2002).

The 1970s and 1980s brought changes in consumer attitudes toward service and advances in technology that facilitated more efficient call distribution, but contact centers continued to be plagued by rigid management and poor work design (Hauptfleisch and Uys 2006). More modern, less intuitive products (e.g., electronics) created greater demand for quality customer support. Bagnara (2000) suggested that changes in consumers' attitudes affected the nature of incoming calls. The representative's role evolved from resolving claims to supporting products and the customer's lifecycle. Not equipped to handle the shift to product and consumer support technologically or in terms of the knowledge, skills, abilities and other qualities (KSAOs) of front-line representatives, many centers experienced long wait times, plummeting performance, and dissatisfied callers. The emergence of computers, advanced telephony, and automatic call distribution systems in the late 1980s and early 1990s helped improve contact centers' performance. Nevertheless, management practices and job design failed to adapt to the changes, prompting Wickham and Collins (2004) to describe contact centers as "white collar factories."

The contact center industry has grown at an astonishing rate since the mid-1990s, both domestically and abroad, as companies have attempted to connect directly with customers more quickly and inexpensively (Burgess and Connell 2004; Gilson and Khandelwal 2005; Holman 2003; Taylor and Bain 2005). Holman et al. (2007) pointed out that the contact center sector emerged in most countries between 1997 and 2002. A number of factors enabled this growth, such as competition (Ellis and Taylor 2006), improvements in technology and infrastructure (Norling 2001; Van Gass 2003), and reduced costs of conducting business both near- and off-shore.

Management practices and job design remained relatively unchanged even as the contact center industry has advanced technologically and expanded globally. Research suggests that flexible, autonomous service models outperform rigid, predefined processes (Knights and McCabe 1998). Batt (2002) demonstrated that high-involvement work practices produced lower attrition and higher sales than standardized processes. Contact centers that allow for more independence and collaboration benefit from a more engaged and satisfied workforce (Holman 2002; Loveman 1998). In a comprehensive survey of US contact centers, Batt et al. (2004) reported that the industry averaged 33 % annual turnover with a 6 % daily absenteeism rate; outsourced centers experienced 51 % annual turnover and a 10 % daily absenteeism rate. The results also indicated that centers using high-involvement practices experienced significantly lower turnover (25 % annually) and absenteeism (5 % daily). Despite research illustrating the value of creating an environment that fosters discussion and autonomy, the management structure and job design in many centers—especially outsourced centers—continues to reflect a contemporary form of Taylorism (Russell 2008).

The contact center industry has grown from a single center providing claims support in the late 1960s to a $300+ billion global industry. Contact centers reflect companies' attempts to cater to consumer demand for instantaneous high-quality service. Increased competition and pressure to reduce operating costs have forced many companies to move operations offshore and/or outsource operations to companies that specialize in operating contact centers (i.e., business process outsourcing or BPO), though only 14 % of contact centers in the USA are outsourced (Batt et al. 2004). The offshore investment has become an economic boon for many developing countries (Holman et al. 2007). Despite the industry's expanding footprint, sophisticated technology, and potential to differentiate a company's customer service, many centers continue to be plagued by poor job design and managerial practices that demotivate and disengage employees.

7.2.1 Nature of Contact Center Jobs

Contact center jobs tend to be complex, change frequently, and operate at a frenetic pace. Employees perform multiple tasks with frequent interruptions, engage in repetitive movements, and process complex information in noisy environments, often under time pressure, while their performance, communication, and efficiency are continually monitored (Bakker et al. 2003). These jobs are often highly scripted, routine, and do not give employees control (Bakker et al. 2003; Zapf et al. 2003). A contact center employee sits for extended periods in front of a computer and may be required to wear a hands-free headset (Zapf et al. 2003). The more customers with whom a representative interacts during a shift—either via voice or other medium—the more routine and potentially boring each interaction becomes (Zapf et al. 2003). The work often places emotional as well as mental demands on employees, who must be ready to respond to any number of issues, resolve them quickly, and do so while maintaining a friendly demeanor. Contact center jobs are often associated with high levels of psychological and physiological strain, which lead to burnout, absenteeism, turnover, and other withdrawal behaviors (Bakker et al. 2003; Holman et al. 2002; Sprigg and Jackson 2006; Sprigg et al. 2007; Zapf et al. 1999; Zapf et al. 2003).

The industry's growth has forced contact centers to support a variety of jobs that can be performed in different environments (e.g., physical centers and home offices) and across multiple channels (e.g., voice, email, and chat or instant messenger). Unfortunately, little detailed information is publicly available about most contact center jobs' competencies or KSAOs, so Holland and Lambert (2008) created a survey to collect subject matter expert (SME) ratings on 53 competencies. SMEs completing the survey make four ratings per competency (i.e., necessity at entry, practicality of finding it in the labor pool, its potential to distinguish successful from average performance, and likelihood of trouble if ignored), which are combined to form an importance score that allows the 53 competencies to be sorted from most to least important. The survey database includes ratings on 101 contact center jobs by 2,928 SMEs (e.g., supervisors, incumbents, trainers, and managers) from 16 countries and five continents.

The most important competencies for six common contact center jobs and the channels through which those jobs are performed are summarized in Table 7.1. The diversity and sophistication of contact center jobs is illustrated by the number of channels, environments, and competencies rated as important (i.e., 35) for one or more jobs. The first 15 rows of the table highlight universal competencies important for the six jobs, regardless of whether it is performed in a physical or home office. These 15 competencies reflect personal responsibility, effective communication, emotional control, and comfort with change, technology, and simultaneous work activities. The remaining competencies tend to be more specific to a job (e.g., persuasiveness for outbound sales) or environment (e.g., autonomy for home-office jobs). Job and environment-specific competencies play a central role in understanding the unique and/or nuanced differences between jobs that are vital when determining how to assess candidates, whether to hire a candidate, and then into which job to place a new hire that will maximize the likelihood of success.

7.2.2 Challenges Hiring Contact Center Representatives

Contact center jobs place significant demands on the representative, so successful performance reflects an assortment of individual differences variables. The key to hiring the right contact center representative is to identify the KSAOs that underlie effective performance and measure them accurately. In the remainder of this section, we discuss two of the primary challenges associated with hiring and keeping successful contact center representatives: recruiting and screening.

Recruiting an adequate supply of well qualified candidates is a precursor to hiring the right people. Implementing strategically driven processes to sustain a steady supply of contact center talent is akin to taking out an insurance policy on a center's future. Many centers, unfortunately, seem to favor a just-in-time philosophy—the same approach used to manage physical inventory—in which a pool of "replacements" is identified quickly by posting an advertisement on a job board, rewarding employees for referrals or by calling a staffing office. The problem is that low wages, weak benefits, undesirable shifts, and inflexible schedules force many qualified candidates to seek alternative employment. The final candidate pool often lacks the qualities and skills essential for success in a contact center, but the need to meet hiring goals can outweigh the desire to hire the highest caliber candidates. The result is that hiring managers often take chances on marginal or fringe candidates.

The second obstacle is accurately assessing candidates' likelihood of success. Well-designed screening processes are built on job analysis and empirical validation, create an engaging experience, and are monitored using closed-loop analytics (i.e., continuous statistical analyses linking test or assessment scores to key performance outcomes) to remain aligned with performance outcomes. In our experience, many centers do a reasonable job assessing candidates' knowledge (e.g., computer software and hardware), ability, and personality characteristics. However, when it comes to measuring essential job-relevant skills, far too many centers miss the mark.

The most common jobs in an inbound contact center require a core set of competencies (e.g., first 15 competencies mentioned in Table 7.1), though three of those skills are particularly important. First, in most centers, candidates must possess basic computer navigation skills prior to being hired. The reason is twofold: (1) employment training is generally conducted in a setting that requires the new hire to interact with a computer, and (2) the contact center job almost always requires the representative to toggle between multiple screens, use the Internet, search databases, etc. Candidates lacking basic computer skills struggle with training and frequently fail. For those who successfully complete training, the on-the-job experience can be especially frustrating and often leads to early attrition. Second to computer navigation skills is keyboarding skill. Keyboarding is different from typing skill because it focuses on speed and accuracy of data entry rather than typing formal, punctuated sentences. During calls, representatives input identifying information to access databases necessary to address a customer's questions. After the call has ended, representatives enter notes that summarize the purpose of the call and its resolution. Candidates who lack keyboarding skills often fail to meet productivity goals. Finally, one of the most essential contact center skills is also one of the most difficult to measure—multi-tasking. A fact of life in an inbound, phone-based contact center is that a representative will have to talk and listen to customers while entering information into a database, reading a summary of the customer's account, etc. The skill to "talk and type" or "talk and read" is frequently hailed by industry insiders as an

Table 7.1 Common contact center jobs' core competencies and channels

	Common contact center jobs											
	Customer care		Inbound sales		Technical support		Customer retention		Collections		Outbound sales	
	B&M	W@H	B&M	W@H	B&M	W@H	B&M	W@H	B&M	W@H	B&M	W@H
Competencies												
Compliance	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Listening	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Composure	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Oral communication	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Dependability	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Integrity	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Tact	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Professionalism	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Learning aptitude	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Accountability	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Stress tolerance	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Adaptability	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Multitasking	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Computer skills	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Keyboarding skill	▲	▲	▲	▲	▲	▲	▲	N/A	▲	▲	▲	▲
Problem-solving	▲	▲			▲		▲	N/A	▲			
Openness to feedback	▲	▲	▲					N/A	▲		▲	
Decision-making	▲	▲					▲	N/A	▲		▲	
Detail orientation	▲	▲	▲	▲				N/A				
Probing	▲	▲	▲	▲	▲	▲		N/A				
Computer knowledge		▲		▲	▲	▲		N/A		▲		▲
Autonomy		▲		▲		▲		N/A		▲		▲
Perseverance		▲						N/A		▲	▲	▲
Problem recognition					▲			N/A				
Empathy						▲	▲	N/A				
Conflict resolution							▲	N/A	▲			

Table 7.1 (continued)

Positive attitude			▲				▲	N/A	▲	▲	▲	▲
Resilience								N/A				▲
Sales aptitude			▲	▲				N/A			▲	▲
Self-confidence			▲	▲				N/A			▲	▲
Self improvement				▲				N/A				▲
Relationship building								N/A			▲	
Persuasiveness								N/A				▲
Ambition								N/A				▲
Flexibility							▲	N/A				▲
Channel												
Voice	●	●	●	●	●	●	●	●	●	●	●	●
Chat/IM	●	●	●	●	●	●	●	●				
E-mail	●	●					●	●	●	●		

Sample sizes range from $N = 10$ for W@H collections to $N = 1{,}385$ for B&M customer care

B&M Brick-and-Mortar, *W@H* Work-at-Home

Competency definitions from job analysis survey are as follows: *Compliance* follows rules, regulations, and guidelines, *Listening* pays attention to what people say without interrupting, *Composure* controls emotions when dealing with hostile people; *Oral communication* communicates in spoken form in a clear and grammatically correct manner, *Dependability* follows through on commitments; can be counted on, *Integrity* adheres to ethical principles, *Tact* recognizes what to do or say to avoid offending others, *Professionalism* displays proper business etiquette, *Learning aptitude* learns and applies new information quickly and accurately, *Accountability* takes responsibility for actions and consequences, *Stress tolerance* manages the physical and emotional effects of pressure, *Adaptability* adjusts oneself to changing conditions or situations (e.g., policies), *Multitasking* works on several tasks at the same time without making errors, *Computer skills* navigates and uses a computer effectively to accomplish tasks, *Keyboarding skill* uses a keyboard to enter information quickly and accurately, *Problem-solving* uses logic and reason to identify solutions to problems, *Openness to feedback* accepts and applies others' feedback to enhance own performance, *Decision-making* makes choices that serve the business and its customers well, *Detail orientation* produces work that is error free and high in quality, *Probing* examines issues closely by asking questions, *Computer knowledge* displays working knowledge of computer hardware and software, *Autonomy* works well with little or no supervision, *Perseverance* remains focused on completing the job and is not easily distracted, *Problem recognition* identifies patterns or trends that may signal potential problems, *Empathy* identifies with and understands another person's situation or feelings, *Conflict resolution* solves disagreements between others diplomatically, *Positive attitude* upbeat person who sees the good in people and situations, *Resilience* rebounds quickly from difficult or stressful situations, *Sales aptitude* positions offers and convinces others to purchase, *Self-confidence* believes in oneself and one's skills and abilities, *Relationship building* establishes connections with others based on respect and trust, *Persuasiveness* convinces others to take a desired action, *Ambition* shows strong desire to achieve results, *Flexibility* changes one's approach to accommodate unique customers' needs essential factor for an employee to be successful. Employees who struggle to multitask tend to become overwhelmed by the job, fail to meet performance standards, provide poor service, and usually leave the job early in their tenure.

Finding, hiring, and retaining employees with the characteristics necessary for long-term success can provide an organization with a significant competitive advantage (McCulloch and Turban 2007; Sawyerr et al. 2009). The US Department of Labor (2000) recommends a "whole-person approach" to assessment to help ensure adequate measurement of the predictor domain. In many cases, using a variety of assessments is a good strategy (Dunnette 1966) for increasing the defensibility of the pre-hire selection system as a whole (Pulakos and Schmitt 1996), and is supported by research on the incremental validity of various selection methods (Schmidt and Hunter 1998). However, in instances where a particular domain has not been measured well, such as with contact center skills, until the emergence of multimedia simulations in the last decade, adding more assessments to a battery will lengthen the screening process without adding incrementally to prediction.

Once we know what to measure, we must ensure that we measure it well. According to the *Principles for the Validation and Use of Personnel Selection Procedures* (hereafter *Principles*; Society for Industrial and Organizational Psychology 2003), "validity is the most important consideration in developing and evaluating selection procedures" (p. 4). This is because the validity of a pre-hire tool provides evidence of its job relevance, which is essential for defensibility (Equal Employment Opportunity Commission 1978). Thus, in order to know that we have measured what we intended to measure, and that we have measured it well, we must gather validity evidence. In the next section, we briefly explore the history of multimedia simulations.

7.3 Using Simulations to Select Better Employees

Organizations have relied on decision-making processes to identify the best job candidates for thousands of years. Plato described physical and cognitive ability assessments used to determine selection for state service in ancient Greece (Jeanneret and Silzer 1998). Since that time, organizations have experimented with a dizzying array of tools and methods to improve accuracy, such as interviews, graphology, biographical data, personality inventories, general mental ability tests (GMAs), SJTs, and work samples and job simulations. The underlying goals are to design pre-hire screening processes that:

- predict important work outcomes (e.g., job performance and turnover) accurately,
- create a positive candidate experience (e.g., face valid, job related, engaging),
- produce a fair outcome that is unaffected by demographic characteristics,
- reflect cost effective methods, and
- administer easily, particularly when hiring volume is high.

Many assessments fail to meet one or more of the above goals (see Table 7.2). Work sample tests and job simulations represent attractive options based on the first three

Table 7.2 Alternative selection procedures

Reference	Selection procedure	Published validity estimates[a]	Candidate reactions	Potential for adverse impact	Cost to develop	Feasibility for high-volume hiring
McDaniel et al. (1994)	*Interview*	Vary depending on interview type—.18 for unstructured; .24 for structured	Vary depending on nature of questions	Moderate to high (also potential disparate treatment concerns for unstructured)	Moderate to high (costly and time consuming to train interviewers)	Low (one-on-one administration required)
Rothstein et al. (1990)	*Biographical data*	.26	Potential negative (unfair, invasive, or not job related)	Low	Low	Moderate to high
Pearlman et al. (1980)	*General mental ability*	Vary depending on measure—.21 for general mental ability; higher for many specific measures	Vary depending on face validity of test	High	Low	High
Barrick et al. (2001)	*Personality*	Vary depending on characteristic—.21 for conscientiousness	Vary depending on nature of questions	Low	Low	High
McDaniel et al. (2001)	*Situational judgment*	.26	Generally positive (fair, job related)	Low	Moderate to high (time/cost may be better spent on higher fidelity simulations)	High
Roth et al. (2005)	*Work samples*	.26 (Regarded by many researchers as the most valid predictors of job performance)	Positive (fair, job related, engaging)	Low	Moderate to high	High

[a]Values represent sample-weighted correlations

criteria (validity, candidate perceptions, and low adverse impact), but struggle to meet the last two criteria because they can be costly, time consuming, and difficult to develop and/or administer.

Concerns about the accuracy and defensibility of some pre-hire screening procedures prompted challenges to conventional thinking about validation and testing (Guion 1967). Wernimont and Campbell (1968), for example, argued for the adoption of a *behavioral consistency model* over the *classic validity model*, which assumed that assessment scores are only an indicator of future behavior. The behavioral consistency model suggested that by sampling candidates' actual behavior using high-fidelity assessments, rather than relying on indirect signs, researchers could predict job performance more accurately (cf. Guion 1998; Roth et al. 2005). Behavioral sampling reached its height in the 1960s partly due to the growing popularity of managerial assessment centers (Bray and Grant 1966).

Published articles often classify simulations as a type of a work sample without discussing the rationale or criteria on which the decision was based. A commonly accepted definition of work sample comes from Ployhart et al. (2006), who state that "a work sample test is a test in which the applicant performs a selected set of actual tasks that are physically and/or psychologically similar to those performed on the job" (p. 538). However, one important difference between these two approaches is that a candidate completing a work sample will perform a subset of a job's tasks, often in the actual environment and using the requisite tools and equipment, whereas a candidate completing a simulation will perform job-related activities in a fabricated environment. Nevertheless, disentangling research on work samples from research on job simulations is nearly impossible because much of the published literature subsumes simulations under the "work sample" label.

There are at least four benefits of using job simulations or work samples in employee selection (Callinan and Robertson 2000; Robertson and Kandola 1982). First, work samples are among the most valid predictors of job performance (Hunter and Hunter 1984; Reilly and Warech 1993; Roth et al. 2005), outperforming GMA in at least one large-scale meta-analysis (Schmidt and Hunter 1998). Earlier meta-analyses included a smaller-scale review (Hunter 1983) of work sample tests in relation to supervisor ratings of performance in non-military samples and a review of work sample studies published from 1964–1982 (Schmitt et al. 1984). The Hunter (1983) and Schmitt et al. (1984) studies reported mean corrected correlations of .42 and .38, respectively. These results align with Roth et al.'s (2005) more recent meta-analysis that reported a mean corrected coefficient of .39. Finally, a frequently cited meta-analysis examining the validity of assessment centers (Gaugler et al. 1987), which consist primarily of work samples and simulation exercises, reported mean corrected coefficients of .53 (ratings of employee potential) and .36 (actual job performance).

Second, simulations appear to cause less adverse impact than some other types of assessments, such as GMA. Schmitt et al. (1996) reported that GMA scores routinely reflect a difference of 1.00 standard deviation (SD) favoring White over minority candidates whereas work samples show a difference of only .38 SD between White and African–American candidates, and virtually no difference between White and Hispanic–American candidates. More recent research (Bobko et al. 2005; Roth et al. 2008) suggests that subgroup differences may be higher than Schmitt et al. estimated (e.g., Roth et al. reported a .73 SD difference between African–American and White candidates), though still lower than the difference produced by GMA. The relatively modest group differences probably contributed to Ployhart and Holtz's (2008) suggestion that simulations represent one way to avoid or decrease adverse impact in the selection process. Furthermore, an analysis of Federal court cases (Terpstra et al. 1999) that found work sample tests had been defended successfully in six of the seven cases on record.

Third, job simulations and work samples show high face validity (Callinan and Robertson 2000; Cascio and Phillips 1979) and are viewed favorably by candidates (Hattrup and Schmitt 1990; Hausknecht et al. 2004; Vance et al. 1989). According to Hausknecht et al. (2004), positive candidate reactions are influenced not only by face validity and job relatedness, but also by the perceived validity of the selection procedure. A recent meta-analysis of candidate reactions (Anderson et al. 2010) confirmed earlier findings that candidates preferred work samples over GMA, biographical data, and personality inventories.

Finally, simulations provide candidates with a realistic job preview (Callinan and Robertson 2000) because they often mirror important aspects of the job on which they are based. As early as 1973, O'Leary touted the use of job simulations, citing two primary benefits: (1) they help the organization learn important information about a job candidate's suitability based on demonstrated behaviors and skills and (2) they "[enable] the applicant to learn something important about the job's suitability ... (p. 148)."

There are three major challenges associated with the design and deployment of multimedia simulations. The first challenge is that it is difficult to create a simulation capable of immersing a job candidate into a series of high-fidelity scenarios, particularly in complex jobs (Callinan and Robertson 2000; Lievens and De Soete 2012). Creating an artificial setting that replicates a work environment, complete with realistic situations, encounters, interactions and problems, is a painstaking process that requires the design team to draw on the collective expertise of many SMEs and complete numerous iterations to perfect the product. The time and effort required to build a realistic simulation that accurately measures a candidate's job related skills is one reason behind the rise of low-fidelity simulations or SJTs (Motowidlo et al. 1990), which use noninteractive, job-relevant vignettes to assess what candidates would do in different scenarios.

The second challenge concerns hurdles associated with administering a multimedia simulation across a diverse technological landscape. The technology that enables a simulation to work properly usually requires testing computers to meet or exceed certain technical specifications (e.g., hardware, software, and Internet access). Adhering to the minimum technical requirements is simplest in company-owned or proctored testing centers, though Internet firewall and other security-related monitors periodically interfere with a candidate's ability to access and complete a simulation. As more companies have adopted remote-testing policies (i.e., allowing candidates to test from home), the number of candidates affected by technical glitches has skyrocketed due to incompatible equipment, missing or an incorrect version of software, poor or slow Internet connection speed, and user error (Barak and English 2002; Garland 2012; Tippins et al. 2006). According to some estimates (Fluck et al. 2009; FurstPerson 2012), as many as 20–40 % of candidates testing from home experience some type of technical issue while completing a simulation. Although it appears rare for a technical glitch to affect a candidate's score on a simulation, these nuisances may negatively impact the candidate's experience and perception of the prospective employer.

The last major challenge associated with simulations is the cost to design and develop them. It takes a team hundreds of hours to understand a job, identify its essential functions and skills, and parlay that knowledge into a storyboard and production-ready script. The implication is that companies often spend tens of thousands of dollars before beginning to produce the simulation (Roth et al. 2010). After hiring actors, graphic designers, instructional designers, creating beta versions of the tool and, eventually, a final production-ready product, the investment in a simulation is often between $100,000 and $500,000 or more, depending on its length and the complexity.

The challenges associated with multimedia simulations undermined their popularity in the past (Callinan and Robertson 2000; O'Leary 1973; Schmitt and Mills 2001), but improved technologies (e.g., broadband, Adobe® Flash®, and HTML5) minimized two of the major barriers to entry (i.e., administration and cost), thus paving the way for a new generation of high fidelity assessments.

During the 1990s, job simulations required companies to install sophisticated software on every computer being used to administer the assessment. The installation requirements limited the locations from which a candidate could complete the simulation, and a company's internal security settings frequently complicated the process. Adding to the problem was that the multimedia files used by many simulations required significant random-access memory (RAM) and hard drive space to operate properly, and many testing computers failed to meet the minimum requirements. In addition, releasing an update or launching a new version of a simulation required the software to be reinstalled on every computer, creating new technical challenges and complex version-control issues with every release.

Improvements in the speed and reliability of Internet access during the late 1990s and early 2000s, along with advances in software technology, enabled web-based delivery of job simulations. Delivering simulations via the Internet alleviated some of the hardware and version-control problems, but not many computers used standard internet communication ports at the time. Therefore, companies selling pre-hire screening simulations during these formative years had to work closely with a customer's IT team, sometimes for weeks at a time, to configure Internet access. Widespread availability of broadband Internet access (68 % of US households reported having broadband access in 2010; US Department of Commerce 2011) and the use of standard internet ports helped address one of the constraints surrounding web-based delivery of simulations. A second development, software supporting web-based multimedia applications without requiring a simulation to be installed on a local computer (e.g., Flash® and Silverlight®), helped overcome most of the remaining administrative hurdles. In fact, today, the most common technical problems encountered by candidates are easily corrected with a free software update.

Animation authoring software has reduced the costs of building multimedia simulations. Building a simulation in the 1990s was typically a custom project that required a programmer to spend countless hours writing code. New technologies, such as Flash®, simplified and streamlined the authoring process. The efficiency gained from more advanced authoring software has reduced the investment required to build elegant multimedia simulations. For a more extensive treatment of the technologies available for simulation developers, we refer the interested reader to Chap. 4 (Hawkes, this book).

7.4 Contact Center Simulations

The nature of contact center work makes it challenging to identify candidates with the greatest likelihood of success. The work is complex and varied, and qualified candidates must possess a multitude of skills, abilities, and personality characteristics. As such, contact centers provide the ideal context for examining the value of accurate, high-quality pre-hire assessments that simulate the complex nature of the work to evaluate candidates' job relevant skills (Frisch 1998).

Contact centers have used simulations to evaluate job candidates' skills for more than a decade. Much of this work has been obscured from public view because organizations in the test publishing and contact center industries gained competitive advantages from designing and using simulations. The following review of published literature on contact center simulations, therefore, summarizes only a fraction of the work on these tools. Highlighting the disconnect between research and practice is that there appear to be fewer than a dozen peer-reviewed empirical articles on contact center simulations, but a Google® search on the term "contact center simulations" produced over 300,000 results.

Some of the earliest published research on contact center simulations stems from the work of Sidney Gael, Donald Grant, Douglas Bray, and their colleagues at AT&T in the late 1960s and early 1970s. The AT&T team created an interactive job simulation to study customer service representatives (Gael and Grant 1972) and telephone operators (Gael et al. 1975). Participants received detailed training on how to use job aides, equipment, and perform the job properly before participating in the study. The simulation began with a call from an administrator, playing the role of a fictitious customer, to a participant. Participants interacted with the caller and used job aides and equipment to manage the call. A second group of administrators observed participants' behavior and evaluated their performance. Perhaps reflecting the time and cost associated with conducting these simulations, the AT&T team used scores from the simulation as a criterion measure in their studies. In the service representative study, the predictor battery included "a role playing interview modeled after tasks performance by [Service Representatives] in telephone contacts with customers"

(p. 136), which significantly correlated with training proficiency and job performance measured by the simulation.

In one of the only other peer-reviewed articles examining the validity of contact center simulations in pre-hire screening, Schmitt and Mills (2001) studied service representatives using a "high-fidelity computerized job simulation" to measure job relevant skills. Similar to the simulations used by Gael and colleagues three decades years earlier, trained assessors evaluated candidates while completing a realistic telephone-based role-play initiated by other assessors acting as customers. During the role-play, candidates used computers to navigate databases while handling the fictitious customer's call, which corresponds to the work contact center employees perform today. Schmitt and Mills found that scores on the job simulation significantly predicted on-the-job performance ($r = .32$).

More recent contact center simulation studies adopted methods similar to those used by Gael and Grant (1972), Gael et al. (1975), and by Schmitt and Mills (2001). However, these studies focused on emotional labor (Rupp and Spencer 2006), self-regulatory behaviors (Zyphur et al. 2007), and emotional regulation (Chi et al. 2011; Goldberg and Grandey 2007) rather than on the validity of simulations in pre-hire settings.

Although most recent studies do not improve the field's understanding of simulations as predictors of post-hire performance, they provide a glimpse into the evolution and ever-increasing complexity of these tools. Murthy et al. (2008), for example, used interactive, multimedia simulation software, complete with branching technology (i.e., programming that moves the customer engagement closer toward or farther from resolution based on the quality of the participant's decisions), to train employees how to interact with different types of customers in a realistic, yet low-stakes setting. The confluence of three streams—technological advances, pressure to more closely mirror the complexity of person-to-person interactions, and creating an immersive and interesting candidate experience—forced the field to blend test construction and film production into a unified development model.

7.4.1 Designing a Modern Contact Center Simulation

The value of a contact center simulation as a pre-hire screening tool is predicated, at least in part, on its ability to immerse a candidate into a particular job. Achieving such a high level of fidelity requires multidisciplinary collaboration to ensure the simulation conveys a job's essential duties and measures a candidate's performance accurately. The steps involved in developing a contact center simulation are shown in Fig. 7.1.

Designing a modern contact center simulation represents a significant undertaking that generally proceeds sequentially through the ten steps shown in the figure. The time that it takes to complete the steps varies based on the type of job, availability of resources, access to data, and the complexity of scoring and reporting, but a moderately complex three-call version of a contact center simulation can be designed, developed, validated, and released in 8–12 months at a cost of $150,000 to $300,000.

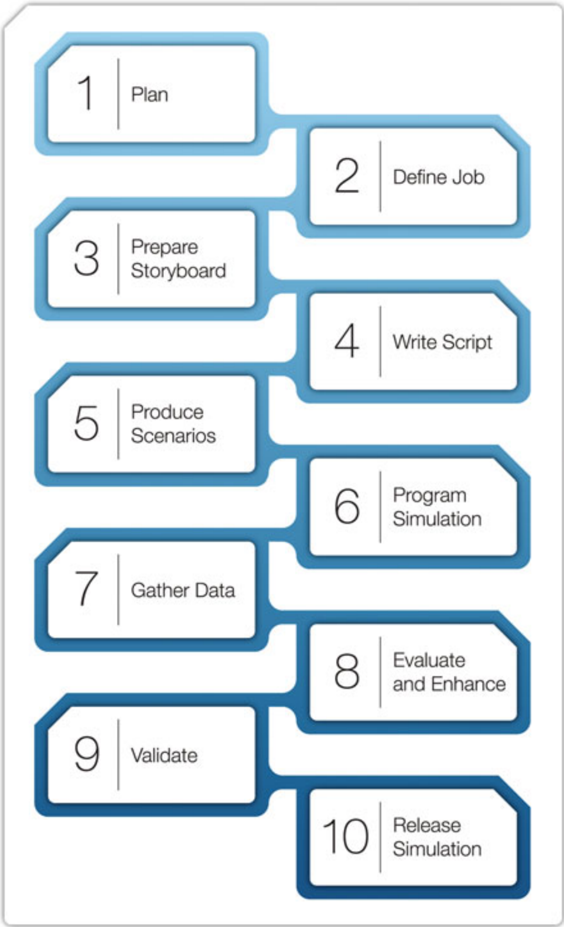


Fig. 7.1 Contact center simulation development steps

7.4.2 *Profile of a Modern Contact Center Simulation: CC Audition®*

Two factors inspired development of CC Audition® (FurstPerson 2006; Holland and Lambert 2011). First, providing staffing services to the contact center industry allowed FurstPerson, Inc. to experience firsthand the challenges of identifying candidates capable of performing well and remaining in a contact center job for more than a couple of months. The company's leadership team believed that the difficulty of making successful temporary placements represented an opportunity to differentiate its practices by using pre-hire screening tools that would improve the performance and tenure of temporary staff. After reviewing the assessment landscape and best practices at the time, the company attempted to adopt a holistic measurement strategy that focused on identifying the unique blend of personality characteristics, biographical factors, skills, and GMA essential for successful performance and retention. However, we recognized almost immediately that the lack of well-designed, high-fidelity contact center skills assessments in the market was a gap that needed to be closed.

Second, we saw potential value in combining elements from the video game industry with modern test design to create an engaging, fun candidate experience that portrays the contact center job realistically while measuring essential skills accurately. The notion of creating a "fun" candidate experience received enthusiastic endorsement from contact center recruiting teams, which frequently struggled to move large volumes of candidates through the entirety of the hiring process. Ultimately, the need for a high-fidelity contact center skills assessment that provided candidates with an informative and entertaining experience laid the groundwork for CC Audition®.

The simulation's development mirrored the steps outlined in Fig. 7.1. The process began by collecting data to clarify (a) contact center jobs' essential tasks and duties, (b) the types of inquiries fielded by representatives, and (c) the competencies required to perform the job(s) well. The industry- and job-related intelligence helped the design team prepare eight unique storyboards (e.g., interactive voice scripts, contact center dashboards, and knowledge bank). A customer advisory board reviewed the storyboards for accuracy, realism, and relevance to contact center jobs. After incorporating the advisory panel's feedback into the scenarios, the team drafted a module to teach job candidates, regardless of previous contact center experience, how to perform the representative's job in the simulated contact center. The design team then partnered with actors to record audio files and programmers to develop a beta version of the simulation. Finally, field testing permitted the design team to finalize scoring and evaluate the simulation's psychometric properties and validity.

The simulation is a multimedia assessment that predicts contact center representatives' job performance. An interactive dashboard places candidates in a fictitious contact center, whether in a physical contact center (Fig. 7.2) or from a home office (Fig. 7.3), in which they play the part of a contact center representative. The format of the simulation contains a training module (Fig. 7.4), sample customer call to practice the job in a low-stakes setting, and then between one and three scored customer engagements (Fig. 7.5). The simulation measures a candidate's critical contact center skills (see Table 7.3).

The simulation was designed to measure contact center skills consistently and accurately. Its success delivering on these objectives depended on its internal psychometrics and external validity. Reliability summarizes the consistency of the simulation's measurement. A selection tool that fails to measure its intended constructs reliably is probably inappropriate to use in high-stakes decision making processes, such as hiring. Table 7.4 presents two estimates of the simulation's reliability: (a) Cronbach's (1951) alpha, a well-accepted statistic for estimating consistency of measurement across test items and (b) test-retest reliability, which reflects the stability with which a test measures a construct across repeated administrations to the same candidate. Cronbach's alpha and test-retest reliability coefficients reflect acceptable reliability at .70 or higher, good reliability at .80, and excellent reliability at .90 or greater (Nunnally 1978). Both Nunnally and the US Department

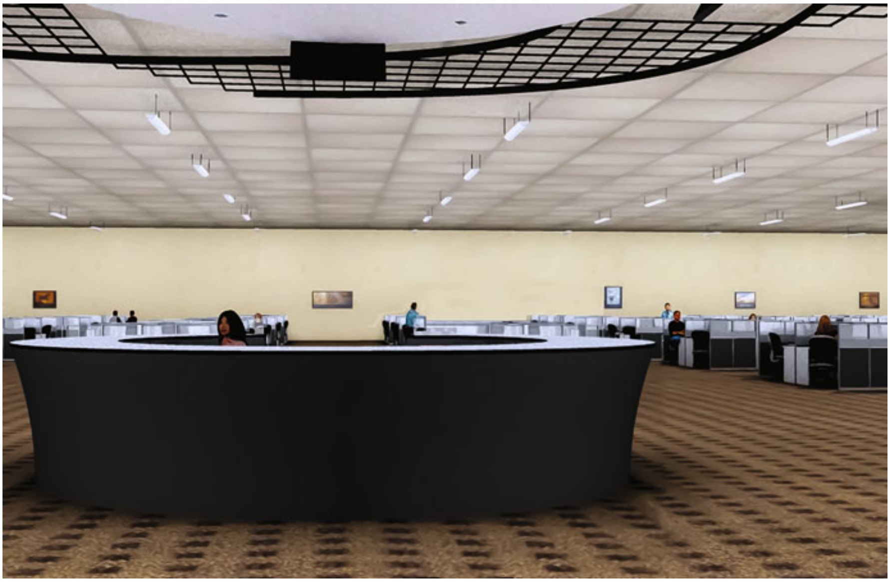


Fig. 7.2 Brick-and-mortar contact center simulation

Fig. 7.3 At-home contact center simulation

Fig. 7.4 Simulated contact center training

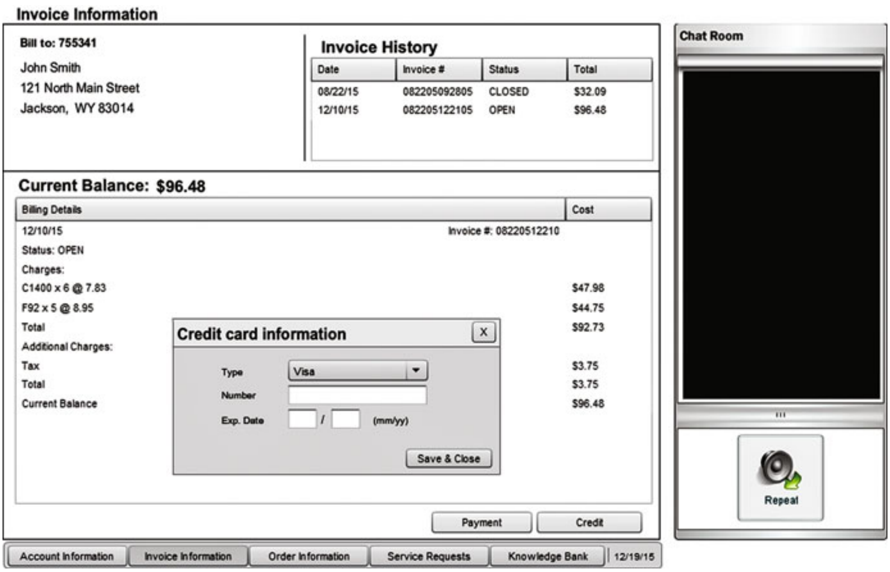


Fig. 7.5 Simulated customer account interface

of Labor (2000) argued that the type of reliability, type of measure, and the way in which the measure is being used are essential considerations when evaluating whether an instrument possesses a satisfactory level of reliability. According to the US Department of Labor's guidelines on interpreting reliability coefficients, an acceptable test-retest reliability coefficient may be lower than .70 for constructs that

Table 7.3 Competencies measured by CC Audition®

Competency	How it is measured
Computer skills	Computer skills are assessed by requiring candidates to complete actions such as accessing customer payment and order information, navigating between multiple screens, and searching for information using a knowledge bank
Accuracy	Accuracy skills are assessed by evaluating a candidate's attention to detail (e.g., following procedures) and entering information and data correctly
Multitasking	Multitasking skills are assessed using proprietary timers that track the speed at which a candidate performs multiple activities correctly
Customer service potential	Customer service potential is measured by the quality and efficiency of the customer engagement
Sales potential	Sales potential is measured based on the candidate's success identifying sales opportunities, positioning the proper offer, and influencing the customer's decision by overcoming resistance and suggesting alternatives

Table 7.4 CC Audition® reliability estimates

Study sample	*N*	Type of reliability	Reliability estimate
Contact center representatives from insurance, financial services, and outsourcing organizations	168	Cronbach's alpha (α)	.94
Candidates applying to a variety of contact center organizations	457	Cronbach's alpha (α)	.95
Candidates who applied more than once to a contact center organization	376	Test-retest (r_{tr})	.75

Table 7.5 CC Audition® meta-analytic validity results

Criterion	k	*N*	Avg *N*	r_{obs}	SD_r	ρ_v	ρ	SDρ	% VE	90 % CV
AHT	34	6,471	190	.19	0.09	.32	.37	0.16	46	.22
CSAT	19	3,153	158	.14	0.07	.30	.34	0.17	53	.17
QA	24	5,937	247	.18	0.10	.36	.41	0.25	66	.30
Sales	8	924	116	.19	0.08	.32	.37	0.16	64	.21

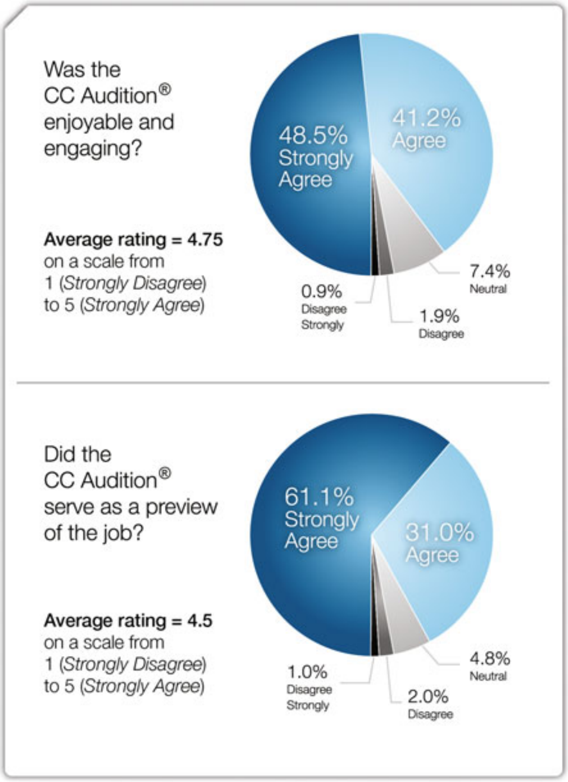
k number of studies, *N* number of participants across *k* studies, *avg N* average number of participants within each study, r_{obs} mean observed validity, SD_r *SD* of observed correlations, ρ_v operational validity (corrected for range restriction and criterion unreliability only), ρ true validity at scale level (corrected for range restriction and predictor-criterion unreliability), SD_ρ *SD* of true validity, *% VE* percentage of variance explained, *90 % CV* credibility value, *AHT* average handle time per call; *CSAT* customer satisfaction, *QA* call quality, *Sales* sales performance

are expected to vary over time (e.g., skills that are likely to improve with practice and experience). In contrast, for Cronbach's alpha, Nunnally cautions that "in those applied settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated" (p. 246). The estimated alpha and test-retest coefficients suggest that the simulation is measuring constructs consistently, both internally and across repeated measures.

Demonstrating consistent measurement, though a necessary condition, does not ensure the simulation measures the intended construct(s). Accumulating criterion-related validation results provides evidence that candidates' simulation scores relate to important on-the-job outcomes, such as work productivity, work quality, or employee retention. According to the *Principles* (Society for Industrial and Organizational Psychology 2003), "validity is the most important consideration in developing and evaluating selection procedures" (p. 4) because it supports its job relevance, utility, and legal defensibility (Equal Employment Opportunity Commission 1978). A summary of meta-analytic validity coefficients for the simulation, across contact center jobs and four common performance metrics, is presented in Table 7.5. The table illustrates the simulation's ability to predict a broad array of criteria, which range from QA (quality assurance; $\rho = .41$), an index summarizing a representative's adherence to formal protocols, to CSAT (customer satisfaction; $\rho = .34$), based on post-call attitudinal survey data gathered from customers. The results correspond to those reported by Schmitt and Mills (2001), though a noteworthy difference is that this simulation does not incorporate trained assessors but the simulations within the Schmitt and Mills analysis did.

The simulation's operational validity, an estimate of an assessment's validity in practice, places it among some of the most valid predictors in employee selection today. Although the simulation's operational validity tended to be slightly lower than GMA ($\rho_v = .37$; Schmidt et al. 2006), it exceeded the estimates for work samples ($\rho_v = .26$; Roth et al. 2005) and Big Five openness ($\rho_v = .20$), conscientiousness ($\rho_v = .19$), and emotional stability ($\rho_v = .17$) factors (van der Linden et al. 2010). These results, albeit preliminary, are encouraging and will hopefully stimulate more research into the potential value of multimedia simulations in contact center employee selection.

Fig. 7.6 Contact Center Candidate Survey Feedback



In addition to designing a reliable, valid assessment, the team also attempted to develop a simulation that would deliver an engaging candidate experience. Job seekers spend hours conducting job searches and completing pre-hire applications and assessments. An organization that is able to create a fun and interesting pre-hire process may be better positioned to hire its most coveted recruits. Figure 7.6 highlights survey results from more than 5,200 contact center job candidates. On a scale from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*), respondents rated the simulation as engaging and enjoyable ($M = 4.74$) and a realistic preview of the job ($M = 4.50$). The feedback reinforces the potential for multimedia contact center simulations to enhance the candidate experience while providing a lifelike window into the job, thus potentially encouraging ill-suited candidates to opt out of the process.

7.5 Conclusion

Multimedia simulations are valuable indicators of a candidate's likelihood to succeed in a contact center's frenetic and intense environment. Contact centers are often characterized by low pay, poor management practices, and rigid, punitive policies. The culture of dysfunction that tends to permeate many centers creates innumerable challenges for talent acquisition teams that must balance the need to "fill seats" with the desire to hire qualified candidates. The need to more accurately select candidates capable of performing well in contact center jobs helped inspire the development of multimedia simulations. Modern simulations allow candidates to experience life as a contact center representative while auditioning for a job from anywhere in the world, and give an organization's talent acquisition team direct visibility into a candidate's potential to achieve on-the-job success. Amid a constellation of problems that characterize many contact centers, multimedia simulations represent a promising tool to help centers identify the most skilled job candidates who will succeed in a complex and challenging role.

7.6 Case Study

7.6.1 Using a Multimedia Simulation to Improve Customer Engagement

Acquiring and retaining customers in the ultra-competitive telecommunications' industry depends on a company's pricing, product, and customer service strategies. Consumer behavior appears to be more influenced by rational and practical considerations, such as saving money, acquiring the latest devices, and/or escaping poor service, than brand loyalty. A company's success distinguishing its brand will ultimately hinge on building a passionate customer following. This review describes a multinational telecommunications company's (referred to as Telecom Z) attempt to begin enhancing customer engagement by using a multimedia simulation to improve the quality of its front-line representatives.

Telecom Z sits at the intersection of old and new technologies. On one hand, its intricate network of wires, cables, fiber, and wireless technologies gives consumers uninterrupted, on-demand access to programming, information, and other people virtually anywhere in the world. On the other hand, front-line representatives often deal with a difficult job, inflexible processes, unfriendly policies, and weak leadership. These behind-the-scenes challenges led to longer customer calls, more repeat calls because a problem was not resolved correctly, and lower customer satisfaction.

An inconsistent, fragmented front-line representative hiring process contributed to the company's customer service problems. With tens of thousands of representatives dispersed geographically, Telecom Z's pre-hire approach shifted from one uniform process to several independent processes over time, usually reflecting the preferences

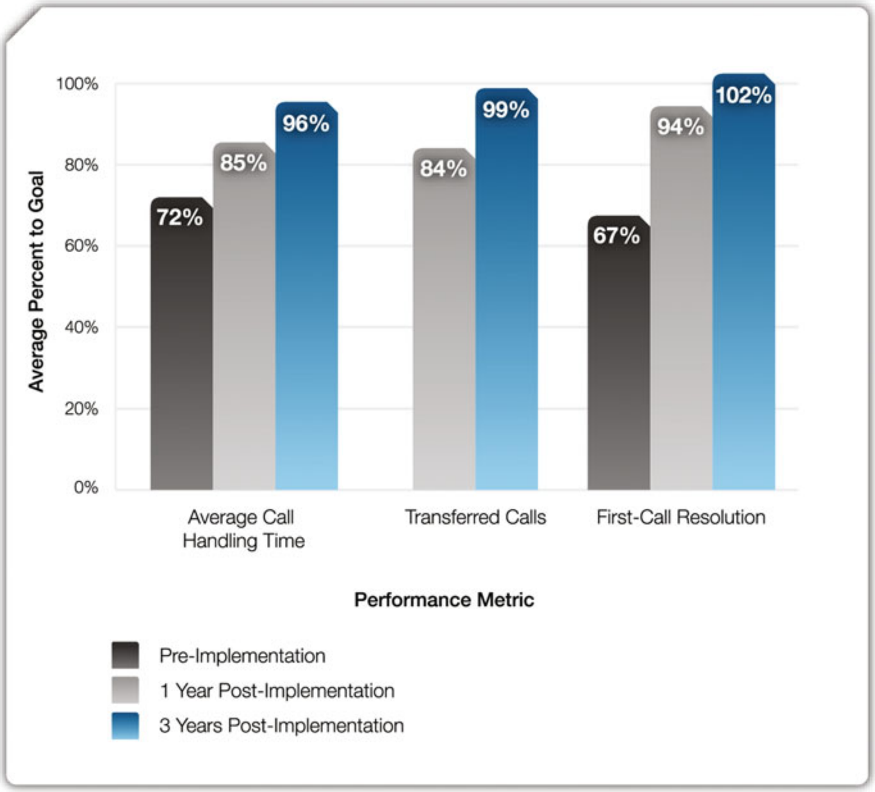


Fig. 7.7 Impact of CC Audition® on Telecom Z's performance over time

of a business unit's leadership team. The inconsistency prevented performance-based comparisons of the various pre-hire approaches.

One component of Telecom Z's customer engagement strategy involved standardizing its front-line representative hiring process. A comprehensive job analysis, encompassing hundreds of participants, identified the core KSAOs required by the representatives' jobs, some of which included computer navigation skill, multitasking skill, ability to learn and apply information, and customer service orientation. A review of alternative pre-hire screening tools capable of measuring the representative jobs' essential KSAOs suggested that Telecom Z consider incorporating a multimedia contact center simulation.

Telecom Z conducted a large-scale concurrent validation study of the simulation to gauge its potential to improve call handling, problem resolution, and customer engagement. The results, based on more than 1,000 representatives, supported the validity and utility of the simulation. Based on the strength of the findings, Telecom Z incorporated the simulation into its pre-hire process in all of its contact centers.

Telecom Z has achieved significant performance improvements since implementing the simulation (Fig. 7.7). The simulation's results are summarized in terms of

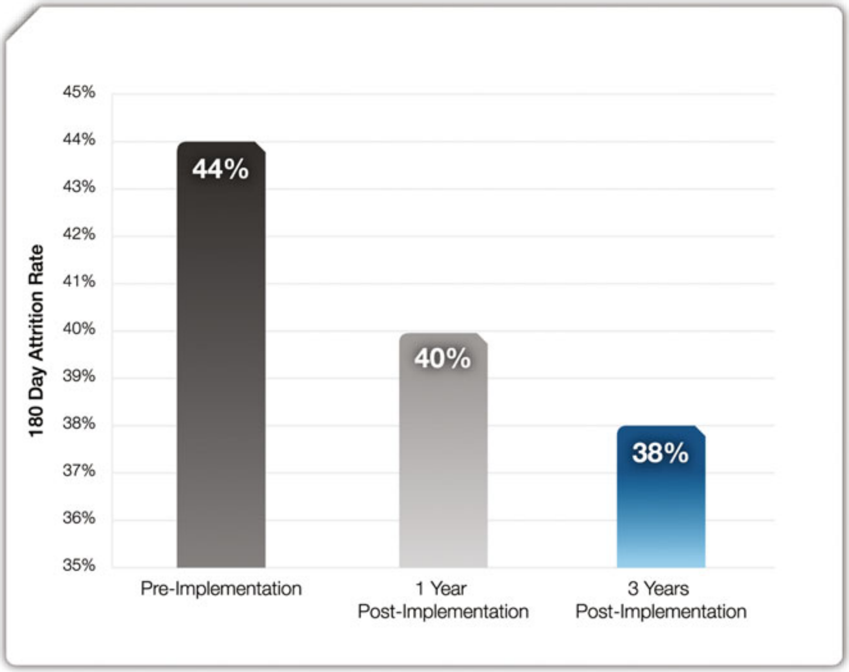


Fig. 7.8 Impact of CC Audition® on Telecom Z's attrition over time

the average percent to goal (i.e., average representative performance relative to the goal set by Telecom Z—100 % or higher means the representatives have achieved or surpassed the minimum expectation on that metric). The simulation has been instrumental in helping Telecom Z improve AHT (33 %) and Repeat Calls (52 %) while simultaneously driving down 180-day attrition (14 %; see Fig. 7.8). These performance gains translate into 8,588 fewer hours of handle time and 16,200 fewer repeat calls per month. When the performance gains are combined with the reductions in attrition, Telecom Z is receiving a return on its investment of approximately 3,300 %.

Although Telecom Z's results provide a solid, data-based case for using multimedia simulations, no pre-hire tool can deliver results in vacuum. Telecom Z's results reflect a collaborative partnership in which data are continually analyzed and reviewed to ensure the simulation is delivering maximum value to the business.

References

Anderson, N., Salgado, J. F., & Hulsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment, 18,* 291–304.

Bagnara, S. (2000). Towards telework in call centers. http://www.telework-mirti.org/bagnara.htm. Accessed 1 Feb 2013.

Bakker, A. B., Demerouti, E., & Schaufeli, W. B. (2003). Dual processes at work in a call centre: An application of the job demands—resources model. *European Journal of Work and Organizational Psychology, 12*, 393–417.

Barak, A., & English, N. (2002). Prospects and limitations of psychological testing on the internet. *Journal of Technology in Human Services, 19*, 65–89.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1/2), 9–29.

Batt, R. (2002). Managing customer services: Human resources practices, quit rates, and sales growth. *Academy of Management Journal, 3*, 579–597.

Batt, R., Doellgast, V., & Kwon, H. (2004). The U.S. call center industry 2004: National benchmarking report. Ithaca, NY: Global Call Center Network.

Bobko, P., Roth, P. L., & Buster, M. A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment, 13,* 1–10.

Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs: General and Applied, 80*, 1–27.

Burgess, J., & Connell, J. (2004). Emerging developments in call center research. *Labor and Industry, 14*(3),1–14.

Callaghan, G., & Thompson, P. (2001). Edwards revisited: Technical control and call centers. *Economic and Industrial Democracy, 22*, 13–37.

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8,* 248–260.

Cascio, W. F., & Phillips, N. F. (1979). Performance testing: A rose among thorns? *Personnel Psychology, 32*, 751–766.

Chi, N., Grandey, A. A., Diamond, J. A., & Krimmel, K. R. (2011). Want a tip? Service performance as a function of emotion regulation and extraversion. *Journal of Applied Psychology, 96*, 1337–1346.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Dunnette, M. D. (1966). *Personnel Selection and Placement*. Oxford: Wadsworth.

Ellis, V., & Taylor, P. (2006). You don't know what you've got till it's gone: Recontextualising the origins, development and impact of the call centre. *New Technology, Work and Employment, 21*, 107–122.

Equal Employment Opportunity Commission (1978). Uniform guidelines on employee selection procedures. Federal Register, 43, 38, 290–38, 315.

Fluck, A., Pullen, D., & Harper, C. (2009). Case study of a computer based examination system. *Australasian Journal of Educational Technology, 25*, 509–523.

Frisch, M. H. (1998). Designing the individual assessment process. In R. Jeanneret, & R. Silzer (Eds.), *Individual psychological assessment* (pp. 135–177). San Francisco, CA: Jossey-Bass.

FurstPerson. (2006). *CC Audition technical manual*. Chicago: Author.

FurstPerson. (2012). *Pre-hire assessment completion rate report*. Chicago: Author.

Gael, S., & Grant, D. L. (1972). Employment test validation for minority and nonminority telephone company service representatives. *Journal of Applied Psychology, 56*, 135–139.

Gael, S., Grant, D. L., & Ritchie, R. J. (1975). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology, 60*, 411–419.

Garland, S. (2012). New online tests hold promise, perils. The Hechinger Report. http://hechingerreport.org/content/new-online-tests-hold-promise-perils_8885/. Accessed 1 Feb 2013.

Gaugler, B. B., Rosenthal, D. B., Thornton III, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493–511.

Gilson, K. A., & Khandelwal, D. K. (2005). Getting more from call centers. *The McKinsey Quarterly. Nov 22,* 1–8.

Goldberg, L. S., & Grandey, A. A. (2007). Display rules versus display autonomy: Emotion regulation, emotional exhaustion, and task performance in a call center simulation. *Journal of Occupational Health Psychology, 12*, 301–318.

Guion, R. M. (1967). Personnel selection. In P. R. Farnsworth, O. McNemar, & Q. McNemar (Eds.), *Annual review of psychology* (pp. 191–216). Palo Alto, CA: Annual Reviews.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hattrup, K., & Schmitt, N. (1990). Prediction of trades apprentices' performance on job sample criteria. *Personnel Psychology, 43*(3), 454–466.

Hauptfleisch, S., & Uys, J. (2006). The experience of work in a call centre environment. *SA Tydskrif vir Bedryfsielkunde, 32*(2), 23–30.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639–683.

Hawkes, B. (2013). Simulation Technologies. In M. S. Fetzer, & K. A. Tuzinski (Eds.), *Simulations for employee personnel selection*. New York: Springer.

Holland, B., & Lambert, D. (2008). Contact Center Job Analysis Survey. Chicago, IL: FurstPerson, Inc.

Holland, B., & Lambert, D. (2011). CC Audition technical manual (2nd ed.). Chicago, IL: FurstPerson, Inc.

Holman, D. J. (2002). Employee well-being in call centers. *Human Resources Management Journal, 12*, 35–50.

Holman, D. J. (2003). Call centers. In D. J. Holman, T. D. Wall, C. W. Glegg, P. Sparrow, & A. Howard (Eds.), *The new workplace: A guide to the human impact of modern working practices*. Chichester, UK: Wiley.

Holman, D. J., Chissick, C., & Totterdell, P. (2002). The effects of performance monitoring on emotional labor and well-being in call centers. *Motivation and Emotion, 26*, 57–81.

Holman, D., Batt, R., & Holtgrewe, U. (2007). *The global call center report: International perspectives on management and employment. Sheffield*. UK: Global Call Center Network.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, and job performance, and supervisory ratings. In R. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257–266). NJ: Erlbaum.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Jeanneret, R., & Silzer, R. (1998). An overview of individual psychological assessment. In Jeanneret, R., & Silzer, R. (Eds.), *Individual psychological assessment* (pp. 3–26). San Francisco, CA: Jossey-Bass.

Knights, D., & McCabe, D. (1998). When 'Life is but a Dream': Obliterating politics through business process reengineering. *Human Relations, 51*, 761–798.

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 383–410). New York: Oxford UP.

Loveman, G. W. (1998). Employee satisfaction, customer loyalty, and financial performance: An empirical examination of the service profit chain in retail banking. *Journal of Service Research, 1*(1), 18–31.

McCulloch, M. C., & Turban, D. B. (2007). Using person-organization fit to select employees for high-turnover jobs. *International Journal of Selection and Assessment, 15,* 63–71.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

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.

Merchant, L. (1998). The 1998 international call centre benchmarking report. Author.
Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640–647.
Murthy, N. N., Challagalla, G. N., Vincent, L. H., & Shervani, T. A. (2008). The impact of simulation training on call center agent performance: A field-based investigation. *Management Science, 54*, 384–399.
Norling, P. (2001). Call centre companies and news patterns of organization. *Economic and Industrial Democracy, 22*, 155–168.
Nunnally, J. C. (1978). *Psychometric theory (2nd ed.)*. New York: McGraw-Hill.
O'Leary, L. R. (1973). Fair employment, sound psychometric practice, and reality: A dilemma and partial solution. *American Psychologist, 28*, 147–150.
Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*(4), 373–406.
Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex sub-group differences and adverse impact in selection. *Personnel Psychology, 61*, 153–172.
Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and research*. Mahwah, NJ: Lawrence Erlbaum Associates.
Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9*(3), 241–258.
Reilly, R. R., & Warech, M. A. (1993). The validity and fairness of alternatives to cognitive ability tests. In L. Wing, & B. Gifford (Eds.), *Policy issues in employment testing*. Boston: Kluwer.
Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and application reaction. *Journal of Occupational Psychology, 55*, 171–183.
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.
Roth, P. L., Bobko, P., McFarland, L. A., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. *Personnel Psychology, 61*, 637–661.
Roth, P. L., Buster, M. A., & Barnes-Farrell, J. (2010). Work sample exams and gender adverse impact potential: The influence of self-concept, social skills, and written skills. *International Journal of Selection and Assessment, 18,* 117–130.
Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75*(2), 175–184.
Rupp, D. E., & Spencer, S. (2006). When customers lash out: The effects of customer interactional justice on emotional labor and the mediating role of discrete emotions. *Journal of Applied Psychology, 91*, 971–978.
Russell, B. (2008). Call centers: A decade of research. *International Journal of Management Reviews, 10*(3), 195–219.
Sawyerr, O. O., Srinivas, S., & Wang, S. (2009). Call center employee personality factors and service performance. *Journals of Services Marketing, 23*, 301–317.
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., Oh, I., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59*, 281–305.
Schmitt, N., & Mills, A. E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86*(3), 451–458.
Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. In C. L. Cooper, & I. T. Robertson (Eds.), *International review of industrial and organizational psychology, Vol. 11*. Chichester, UK: Wiley.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422.

Society for Industrial and Organizational Psychology. (2003). Principles for the Validation and Use of Personnel Selection Procedures (4th ed.). Bowling Green, OH: Author.

Sprigg, C. A., & Jackson, P. R. (2006). Call centers as lean service environments: Job-related strain and the mediating role of work design. *Journal of Occupational Health Psychology, 11*, 197–212.

Sprigg, C. A., Stride, C. B., Wall, T. D., Holman, D. J., & Smith, P. R. (2007). Work characteristics, musculoskeletal disorders, and the mediating role of psychological strain: A study of call center employees. *Journal of Applied Psychology, 92*, 1456–1466.

Taylor, P. D., & Bain, P. (2005). India calling to the far away towns: The call centre labor process and globalization. *Work, Employment and Society, 19*, 261–282.

Taylor, P. D., Baldry, C., Bain, P., & Ellis, V. (2003). 'A unique working environment': Health, sickness and absence management in UK call centers. *Work, Employment and Society, 17*, 435–458.

Terpstra, D. E., Mohamed, A., & Kethley, R. B. (1999). An analysis of federal court cases involving nine selection devices. *International Journal of Selection and Assessment, 7,* 26–34.

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.

U.S. Department of Commerce. (2011). *Exploring the digital nation: Computer and internet use at home*. Washington, DC: Author.

U.S. Department of Labor, Employment and Training Administration. (2000). *Testing and assessment: An employer's guide to good practices*. Washington, DC: Author.

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*, 315–327.

Van Gass, C. (2003, May 7). Call centre industry finds SA's pulse. Africa News Service (online). http://allafrica.com/stories/200305070615.html. Accessed on 1 Feb 2013.

Vance, R. J., Coovert, M. D., MacCallum, R. C., & Hedge, J. W. (1989). Construct models of task performance. *Journal of Applied Psychology, 74*(3), 447–455.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372–376.

Wickham, M., & Collins, G. (2004). The call centre: A nursery for new forms of work organization? *The Services Industry Journal, 24*(1), 1–18.

Zapf, D., Isic, A., Bechtoldt, M., & Blau, P. (2003). What is typical for call centre jobs? Job characteristics and service interactions in different call centers. *European Journal of Work and Organizational Psychology, 12*, 311–340.

Zapf, D., Vogt, C., Seifert, C., Mertini, H., & Isic, A. (1999). Emotion work as a source of stress: The concept and development of an instrument. *European Journal of Work and Organizational Psychology, 8*, 370–400.

Zyphur, M. J., Warren, C. R., Landis, R. W., & Thoresen, C. J. (2007). Self-regulation and performance in high-fidelity simulations: An extension of ego-depletion research. *Human Performance, 20*, 103–118.

Chapter 8
Show Me You Can Do It: The Use of Interactive Simulations in Manufacturing Settings

Matthew O'Connell, Amie Lawrence and Theodore Kinney

Manufacturing jobs vary widely. Some involve relatively straightforward, light industrial tasks such as sorting and packing lightweight materials. Others are much more physically demanding and involve heavy lifting, climbing, or working in awkward positions. And some others are more cognitively demanding because they create highly complex outputs and require workers to monitor equipment and fine-tune machinery. In many manufacturing jobs, exposure to dangerous materials and environmental hazards is common, and the chance for injury is inherent to the job.

As technology advances, manufacturing organizations are using more robust, predictive, job-related, and engaging simulations in their personnel selection processes. Leveraging these technology advancements and using high-fidelity, multimedia simulations can help to prevent these organizations from employing workers that will turnover, make costly mistakes, or significantly increase the risk of injury to themselves or others. Effective performance in many manufacturing jobs requires behaviors that are ideally suited to simulation-based assessments. Competencies such as attention to detail, safety orientation, multitasking, processing speed, and work pace are routinely found to be critical to success in high-performing manufacturing employees (O'Connell and Reeder 2008). These types of behavioral competencies are ideally suited for multimedia, simulation-based measurement methods.

Work sample testing is a form of assessment involving the use of hands-on performance measures, whereby a candidate or incumbent performs a given task or set of tasks under conditions comparable to those found in the position in question (Callinan and Robertson 2000). The primary philosophy behind this approach to assessment lies in the theoretical foundation set forth in the seminal works of both Wernimont

M. O'Connell (✉) · A. Lawrence · T. Kinney
Select International, Inc., Pittsburgh, PA, USA
e-mail: moconnell@selectintl.com

A. Lawrence
e-mail: alawrence@selectintl.com

T. Kinney
e-mail: tkinney@selectintl.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_8, © Springer Science+Business Media New York 2013

and Campbell (1968) and Asher and Sciarrino (1974). Although addressing the concept of validity more broadly, Wernimont and Campbell (1968) encouraged a general shift from the traditional focus on traits and predispositions to a greater emphasis on observable forms of behavior.

Simulations, in general, can be thought to fall along different points on the sign versus sample continuum. One can think of the Wernimont and Campbell (1968) distinction as a continuum based on content (and face) validity and overlap with job tasks rather than two discrete entities. On the far "sign" side, you have personality inventories; on the far "sample" side, you have work samples; in between, you would have situational judgment tests and interactive simulations.

Based on Wernimont and Campbell's (1968) ideas, selection practitioners have built signs of performance (e.g., personality inventories), samples of performance (e.g., in-basket assessments), and all points in between along the continuum. In fact, over the past 40 years it has become common to see manufacturing assessments at various points on the continuum of content and face validity. However, in the past 5 or 10 years, there has been rapid advancement in the degree to which technology has been leveraged to enhance these signs and samples of performance. Obviously, computerized "signs" have become commonplace. Online "paper-and-pencil" tests are widely used, and technology has allowed these assessments to be administered more efficiently to larger samples.

Technology has provided assessment developers the opportunity to push manufacturing assessments closer and closer to the sample side of the continuum. Enhancements in technology allow assessment developers to change the way we measure behavior. While production simulation work samples have been used as far back as the 1950s, the approach to capturing and evaluating behaviors has evolved dramatically with advances in technology. This chapter will highlight not only technologically advanced testing formats, platforms, and simulations from a candidate perspective, but also, and more importantly, this chapter will address how advances in technology have allowed assessment developers to understand candidate performance in work samples to an amazing degree of specificity, consistency, and objectivity. For example, it is becoming increasingly commonplace to see a "day in the life" production simulation, where job candidates actually perform the job at one or more simulated work stations (e.g., building a vehicle dashboard), that have amazingly complex data collection points occurring automatically "behind the scenes." These types of simulations are able to measure variables, such as fine motor skills, that it would be impossible for a rater to observe. It is possible to leverage technology to automatically measure variables such as processing speed, processing accuracy, task completion order, torque, etc. These variables are critical to success, and until recent technological advances, even the most robust production simulation could not capture these measures. Consequently, in this chapter, we will focus not only on innovative approaches to presenting test content to candidates from various points on the sign versus sample continuum, we will also discuss how technology advances have dramatically impacted approaches to collecting, understanding, and evaluating candidate behaviors. In many ways, the automation of data collection, scoring, and evaluation that occurs "behind the scenes" of today's high-tech, multimedia production simulations represent the most complex and innovative developments in leveraging technology to impact how job candidates are evaluated.

As these complex approaches to measuring candidate behaviors evolve, evidence suggests that these procedures lead to incremental validity. By leveraging technology in simulations, the increase in fidelity and measurement precision is leading to positive outcomes. For example, a growing body of research shows that web-based interactive simulations measuring characteristics such as multitasking, processing speed, and attention to detail have high levels of validity in manufacturing situations (cf. Kinney and O'Connell 2012; Kinney et al. 2009; O'Connell and Reeder 2008).

A number of other potential benefits of simulations over traditional measurement methods (namely, cognitive ability and personality) have been put forth in the academic literature. Although their popularity as selection tools continues to increase (Dudley et al. 2006), personality measures often produce only low to moderate validities when used in isolation (Barrick and Mount 1991; Salgado 1997; Schmitt et al. 1984; Tett et al. 1991). While cognitive ability is widely recognized as one of the best available predictors of overall job performance (Gottfredson 1997; Gottfredson 2002; Murphy 2002; Schmidt 2002), it, too, has its own recognized limitations. For instance, cognitive ability does not appear to predict subfacets of the criterion domain for low-complexity tasks as well as other constructs or methods (see Avis et al. 2002 for a brief review). In fact, evidence exists that interactive simulations, both moderate and high fidelity, add incremental prediction of job performance (Kinney and O'Connell 2012; Reeder et al. 2008) and safety (Kung et al. 2012), above both cognitive ability and personality.

The current chapter presents several examples of multimedia, interactive simulations currently in use by manufacturing organizations across a diverse variety of industries. These unique predictors range from computer-based simulations measuring targeted competencies (e.g., multitasking) to complex multi-work station, interactive production simulations involving the actual equipment and processes encountered on the job. An overview of these simulations will be presented along with research regarding validity in predicting job performance and safety as well as evidence of incremental validity, wherever available. Pros and cons of multimedia production simulations are presented, as well as a discussion of the evolution of production simulations over the past 30 years. Suggestions are also made about how to design and validate these types of multimedia simulations, including a discussion of the challenges faced in developing these types of simulations, such as avoiding the overreliance of cognitive loading, time constraints, realism, and the most appropriate types of dependent variables.

8.1 Manufacturing and the Challenge of Hiring the Best People

The United States has the world's largest manufacturing economy, and roughly 12 million Americans are employed in the manufacturing industry (National Association of Manufacturers 2013). Manufacturing has a long history in America and, while it is often publicized when organizations move plants overseas or outside of the United States, manufacturing is still a major sector of the economy. We work with a broad range of manufacturing companies; they hire large numbers of hourly workers across multiple shifts and are constantly in need of finding good employees who can learn new skills, work both independently as well as part of a team, and work quickly, safely, and accurately. Hourly manufacturing jobs are very attractive to many workers because they do not require higher education, and the company will provide employees with the training they need to perform the job. In addition, compared to other hourly jobs, manufacturing jobs often pay well and offer competitive benefits.

The quality of the product that is manufactured by an organization is inextricably linked to the employees who make it. Ensuring that the best employees are hired helps organizations meet their production quotas and quality standards. Over the past two decades, we have toured literally hundreds of manufacturing facilities and studied the job of a "production worker" around the world. While different products are manufactured, the situation is generally the same. Products are made through a series of interdependent processes. Employees are typically under pressure to complete a specific set of tasks in a set amount of time. Their output is then sent along to someone else for the next step in the manufacturing process. Rarely does a manufacturing process have just one employee produce a complete product from beginning to end. With multiple hands touching the product, it is very important that each employee does his or her job effectively so that a high-quality product exists at the end of the process. From the employees' perspective, these jobs provide attractive wages and benefits, but there are trade-offs. There are often unattractive shift times, the work itself can be monotonous and/or physically challenging, and the work environment can be unpleasant.

The average manufacturing company has the challenge of employing a large number of hourly workers who will show up to work every day and on time, work hard, get along with others, and produce a high-quality product. How do they find these people? What assessment tools make the most sense for doing this? The next sections describe two distinct types of tools: computer-based simulations and multimedia physical simulations.

8.2 Computer-Based Simulations

For purposes of simplicity, we refer to simulations administered by computer, whether they are web-based or not, as computer simulations. Computer-based simulations have clear advantages over paper-and-pencil assessments or even non-simulation-based online assessments. Computer-based simulations provide powerful behavioral samples of job-related performance. As mentioned earlier, whereas an online personality inventory may provide a "sign" of how a candidate may perform on the job, a well-developed simulation can capture actual behaviors on tasks very similar to those that will be encountered on the job. These computer-based simulations are often not as high fidelity as a robust multimedia production simulation (discussed later in this chapter); however, they are easy to administer to a large number of candidates with little to no resource expense in the recruiting process. These easily administered computerized work samples are not only efficient to administer, there are also clear advantages in using these types of measures over the typical online assessment measures in terms of predicting future job performance.

Over the last 20 years, more and more organizations have leveraged technology to impact how they measure the qualifications of candidates. Despite advances in technology, most online assessments are little more than paper-and-pencil tests recreated on a computer screen; however, the potential to leverage technology to measure candidates in amazingly sophisticated ways has existed for many years. While manufacturing jobs often do not take place solely in front of a computer monitor, many production operator behaviors are easily captured by computer simulations. Further, many of these behaviors cannot be reliably measured with less sophisticated approaches to measurement.

In developing or choosing an online behavioral simulation, it is important to think carefully about what types of competencies to measure. Certain competencies are ideal to measure with a computer-based simulation; others are better suited to other modes of measurement. For example, consider three competencies often correlated with success in manufacturing positions: positive attitude, welding/machining skills, and multitasking. Let us look at these three competencies with regard to their best mode of measurement.

Positive attitude (also known as positive affectivity) is a personality trait. Specifically, someone high in positive attitude will likely react positively versus cynically, or view their world optimistically across time, setting, and situation. As such, this competency is not ideally measured in a computer-based simulation. Rather, a typical approach (e.g., online or paper-and-pencil self-report assessment) is more than adequate to capture this job-related individual characteristic. Research has shown that by asking candidates how they tend to behave and react to different situations can lead to a reliable and predictive indicator, or "sign," about whether they will have a positive attitude in a production job (cf. Barrick and Mount 1991). Consequently, building a simulation to measure positive attitude or other personality traits is not typically a worthwhile endeavor.

Similarly, specific skills that require job knowledge are not ideally suited to computer-based simulations. For example, the ability to use a mig/tig welder or the ability to machine a part using a mill is required for many skilled positions in the manufacturing setting. These specific skills, however, are better demonstrated with a more robust, high-fidelity, physical simulation where a candidate may be asked to make a weld or machine a part. Through these hands-on, robust simulations, detailed measurements of job-related knowledge or skill can be collected. A computer-based simulation could ask the candidate questions about their experience with particular equipment or about the actual machines, but a high-fidelity simulation evaluates their level of capability using the actual machine.

We have worked with a number of organizations to develop this exact type of assessment. In every case, candidates have been required to "demonstrate the skill," which is ultimately scored by a trained rater using an objective scoring format. These tend to be excellent ways of evaluating whether someone can actually "do" something as opposed to just "knowing how to do" it. The two are not always the same thing.

Multitasking, on the other hand, is a specific ability ideally suited to measurement in a computer-based simulation. With technological advances over time, the opportunity to encounter multiple tasks in the same general time period has become a key feature in many production jobs. How an employee maneuvers through this constantly shifting sea of tasks necessarily impacts that employee's productivity. Consequently, identifying candidates that are better suited to these demanding and complex environments is critical to maximizing prediction in the selection system. Computer-based simulations of competencies such as multitasking can provide robust, accurate, and predictive measures of someone's ability to perform on the job.

While it is true that multitasking on the shop floor does not occur at a computer terminal, but rather often involves monitoring processes, while checking the quality of the material on the line, and often involves moving from machine to machine and process to process, a computerized multitasking assessment can be valuable. A computer-based simulation does not recreate the exact task; however, a well-developed multitasking simulation can capture the underlying behaviors that make a person successful in these types of positions. If we consider the typical environment on the shop floor where an employee is required to multitask, there are some consistent similarities between the jobs. Many production jobs involve situations where (a) there is a heavy requirement to switch tasks frequently; (b) there is uncertainty about when to switch tasks; and (c) there is time pressure to switch tasks and complete work. While the actual tasks cannot be replicated, these characteristics of a multitasking environment can reliably be reproduced in a computer-based simulation. Consequently, if we can measure (via computer simulation) a person's ability to switch tasks at uncertain intervals under time pressures, we are then able to capture an ability that will translate to on-the-job performance.

So, what does an online, computer-based simulation look like? The key to building a computer-based simulation is specifying the behaviors you are trying to predict. For example, consider multitasking again. In the previous paragraph, the behaviors critical to performance in a multitasking environment were identified; the key to building or selecting the right computer-based simulation is finding an exercise that mimics these behaviors. Figure 8.1 provides a screenshot of one such simulation.

In this simulation, the job candidate must perform a variety of tasks that look similar to a typical production environment. This type of exercise is typically very well received by job candidates because the tasks they are asked to do "look and feel" like what they would expect to see on the shop floor. From a multitasking measurement perspective, all of the key characteristics of a manufacturing multitasking environment are present. First, the candidate is asked to switch tasks frequently. There are two main tasks that the candidate must accomplish. On the left side of the screen, there are a series of gauges that move toward the "red zone." By clicking on the gauge, the candidate can reverse the direction of the needle to keep the gauge in the acceptable range. On the right side of the screen, the candidate must also make as many comparisons of number pairs as possible during the exercise. Consequently, the first condition is satisfied: the candidate must switch tasks frequently.

Fig. 8.1 Screenshot of multitasking simulation

The second condition (uncertainty about when to switch tasks) is created with the gauges. The needles move toward the red zone at uncertain rates of speed. The candidate does not know when they will have to switch tasks to and from the number comparison task to keep the needles out of the red zone. Finally, there is inherent time pressure in this exercise. The candidate is aware that this is a timed task, and that he or she is being evaluated based on the number of comparisons that are made and how effective they are in keeping the needle out of the red zone.

By mimicking the multitasking demands on the shop floor, these computer-based simulations have led organizations to more effectively identify candidates who are not only able to do the tasks that are required of them, but also able to effectively switch among them. This is a key point—job tasks do not exist in isolation; rather, job tasks occur sequentially or even simultaneously. It is not enough to measure whether or not someone can do a particular task; to really understand job performance, measuring how well someone can switch between them is a critical consideration. For example, many people can monitor the quality of product on the line; but it takes a skilled person to monitor quality under demanding *takt* times (also known as cycle times), while also monitoring gauges and evaluating how well machines are running. These computer-based multitasking simulations provide a window into how well a job candidate might manage these demanding environments.

What are the results? Do these computer-based multitasking simulations really work to predict job performance in manufacturing settings? Further, how do these types of assessments predict performance relative to other more commonly used types of assessments for manufacturing jobs?

Until recently, with a few exceptions, not many organizations had high-quality computer-based simulations available to use. The evidence that has been collected, however, is in support of the use of these types of simulations. In a recent series of studies, the authors of this chapter have looked into these research questions. Kinney and O'Connell (2012) reported results not only showing that a computer-based multitasking simulation is an important predictor of manufacturing outcomes, but also, these simulations explain unique performance variance not explained by traditional predictors such as cognitive ability, personality, polychronicity (a personal preference for multitasking), and situational judgment tests.

In the first study, we investigated the criterion-related validity of a measure of polychronicity, cognitive ability, and a computer-based simulation of multitasking using 156 manufacturing operators. When these predictors were correlated with supervisor ratings of task performance, the multitasking simulation was the clear winner of the "horse race." Polychronicity did not predict significant variance in task performance. Cognitive ability was a good predictor ($r = 0.23$; $p < 0.05$); however, multitasking was the strongest predictor ($r = 0.36$; $p < 0.05$). We also investigated the incremental validity of the multitasking simulation in this sample, and it was found that the simulation predictor did contribute unique variance in the performance ratings beyond polychronicity and cognitive ability, respectively (polychronicity and multitasking: $\Delta R^2 = 0.18$; $p < 0.05$; cognitive ability and multitasking: $\Delta R^2 = 0.06$; $p < 0.05$).

While predicting supervisor ratings of performance is important, in a second study Kinney and O'Connell (2012) reported on the impact these simulations have in predicting important objective manufacturing outcomes. In this study, the participants were 901 production team members. The focus was to investigate whether or not these simulations could predict workers' compensation claims over a 12-month period. The results indicated that the multitasking simulation added incremental validity to the prediction of both workers' compensation claims ($\Delta\chi^2 = 24.5$, $df = 3$, $p < 0.001$) and costs ($\Delta R^2 = 0.01$; $p < 0.01$) above situational judgment and a personality composite. Clearly, these results suggest that a well-developed computerized sample of performance can be used as an effective predictor of manufacturing performance.

8.3 Multimedia Production Simulations

As previously discussed, online assessments and computer-based simulations are useful assessment tools for measuring important success competencies for production environments. Online assessments provide a "sign" of what a candidate can do, and computer-based simulations move closer to the "sample" side of measurement but are still moderate in their fidelity. Multimedia production simulations are the epitome of a work sample within the manufacturing realm and provide the highest fidelity measurement. This section of the chapter discusses the benefits of multimedia production simulations within the manufacturing industry, and then specifically discusses how advancements in technology have made it possible to design and implement multimedia production simulations.

The words "production simulation" might conjure an image of a candidate standing at a table and building or assembling a part. This is common in traditional production simulations, where the candidate is typically asked to participate in completing a hands-on task at a station for a defined period of time. The work is completed and then evaluated by an assessor or administrator, and then the candidate moves to another station for another task. In recent years, production workers use computers and computerized machines more often in their jobs, and those same advancements have made it possible to build computer technology into production simulations.

Let us first discuss the value of physical production simulations, computerized or not. Practically speaking, physical production simulations are often used in conjunction with other selection instruments. Before candidates get to this phase, they have often passed other assessment hurdles that evaluated them on experience and personality characteristics. What the production simulation provides is a chance to measure observable, physical performance on a simulated, job-related task. The data from a simulation can provide information on competencies that cannot be effectively obtained through other means such as process monitoring, work tempo, stability of performance over time, physical endurance, fine motor skills, and manual dexterity. Physical production simulations can add significant incremental validity to an already valid selection process because of the extra information they provide (cf. Kung et al. 2012; Reeder et al. 2008).

8.3.1 Benefits of Physical Production Simulations

8.3.1.1 Higher Performance

One of our manufacturing clients was using a physical production simulation as part of their hiring process for hourly assembly workers. Due to resource issues and a push to hire a large number of people in a short period of time, the organization stopped using the simulation. Once the new hires began training, human resources started to receive feedback from the managers on the floor. Suddenly, human resources staff was receiving complaints about the quality of the employees. Managers were frustrated because the new hires were not learning as quickly or performing as well as the previous employees. After a short hiatus, the organization reinstated the production simulation and the quality of their hires dramatically improved. Keep in mind, this organization was still screening individuals using valid non-physical assessment tools including personality, situational judgment, computer-based simulations, and cognitive ability measures. However, the additional validity and screening power that they had come to expect could not be realized without the physical simulation.

8.3.1.2 Reduced Turnover

Turnover is an issue for many manufacturing organizations at the hourly level. Whether the turnover is voluntary or involuntary, it is still costly for the organization. Organizations invest quite a bit of time and money into their employees by giving them all of the training they need to learn how to do the job. If an employee leaves or is terminated before becoming a fully productive employee, the organization does not recoup the money they invested in that individual. In addition, they are losing productivity by having an open position, and they must incur recruitment costs for replacing the worker. Our analysis of turnover in these organizations suggests that most manufacturing turnover, at least on the voluntary side, is related to poor job fit and/or dissatisfaction with the job or organization (cf. Lawrence et al. 2004).

Let us explore poor job fit in more detail. Employees who voluntarily decide to leave the organization shortly after they are hired often mention not liking the type of job task or work environment. Workers talk about the monotony of the tasks; they may not have realized that they would be doing the exact same task (e.g., installing electrical wire harnesses) all day long. Some organizations have workers rotate between several different tasks, while others do not. In addition, manufacturing facilities often have less-than-ideal work environments—they can be noisy, hot or cold, and even dangerous. For example, workers in a stamping facility operate 100-ton presses that stamp white hot metal ingots into axles or other components. There are bins of grey hot parts (over 1,000 °F) going past them all the time, and the sound of the presses reminds one of the T-Rex's footsteps in Jurassic Park. Employees in this plant deal with noise and the possibility of danger at every turn. If they are not paying close attention, they could easily be badly burned or crushed. Some people do not feel comfortable working in that environment every day.

Manufacturing jobs can be physically demanding as well. Tire manufacturers typically have a wide range of positions that require workers to lift 25–50-lb tires hundreds of times throughout a work shift. It is not unusual for employees in such companies to walk the equivalent of over 5 miles and lift thousands of pounds during an 8-h work shift. Over the past 20 years, we have toured a number of automobile manufacturing facilities; their assembly workers are required to crawl in and out of the automobile cabs, stand, crouch, reach overhead, and lift greater than 25-lb parts over and over again on a moving assembly line.

Some workers are drawn to manufacturing jobs for the pay and benefits but realize that the actual work does not fit them. A key benefit of production simulations is that candidates have the opportunity to experience some of the potentially negative aspects of the job prior to accepting a position. In the literature, this is called a realistic job preview (RJP) (cf. Premack and Wanous 1985). An effective production simulation will provide a good enough feel for the job that candidates may self-select out of the hiring process, thus preventing the individual from leaving soon after accepting the job.

Several years ago, we designed a very comprehensive production simulation for a global automobile manufacturer at their plant start-up in the southwestern United States. They were focused on the "fit" part of the production simulation and made sure that the building that housed the production simulation was very similar to the conditions on the plant floor. Summers in that part of the country are very hot and the plant can get very warm—so, the candidates completing the production simulation were expected to complete a full 8-h day of work (four production exercises each lasting about 2 h) in a warm building, doing physically demanding work. We worked with this organization to process tens of thousands of candidates through the production simulation phase of their hiring process. On a regular basis, candidates would decide during the lunch break to withdraw from the process. When asked about it, candidates cited two main reasons for choosing to leave: (1) the tasks were boring, or (2) the tasks were too physically demanding. Because the individuals who participated in the production simulation had already passed several other selection hurdles, the candidates who withdrew from the process very likely would have been hired if the production simulation had not been part of the process. Both the candidates and the employer were grateful to have figured this out ahead of time.

8.3.1.3 Reduced Accidents/Injuries

Despite serious efforts on the part of dedicated safety professionals, injuries and accidents still occur regularly in the manufacturing industry. According to the most recent data from the U.S. Bureau of Labor Statistics[1], slightly more than one-half of the 3.7 million private industry's injury and illness cases reported nationally in 2010 were of a more serious nature that involved days away from work, job transfer or restriction—commonly referred to as DART cases. In other words, over 50 % of all injuries were severe enough to lead to loss of work, restricted duty upon return, and/or transferring out of the original job.

The nature of the jobs creates opportunities for workers to get hurt even though there are extensive safety policies and procedures in place to minimize risk. For example, manufacturing workers regularly work with and around machinery; they climb on ladders, work in confined spaces or underground, or work with hazardous chemicals, to describe just some of the dangerous activities that could lead to injuries. In some environments, a small mistake could lead to serious injuries. Injuries are not always a result of an accident; some of the injuries sustained by manufacturing workers are a result of repetitive motion. Because workers perform the same activity in exactly the same way repeatedly throughout the workday, every day, issues can arise with overuse of some muscles and joints.

Some workers are more able to tolerate the physical activity and follow the safety rules that are required to be safe on the job. Well-designed production simulations replicate the physical tasks required and build proper body positioning and procedures into the instructions and evaluation process. Participants are also required to wear the same personal protective equipment (PPE) during the production simulation that would be required on the job. That may sound trivial, but we were surprised by the number of candidates who actually removed their PPEs during the course of the

[1] USDL-11-1502, October 20, 2011.

production simulation even though they were explicitly instructed to keep them on. Imagine what these individuals would do on the job if they cannot even keep their PPE on for 8 h when there is a proctor watching them the whole time. It is much better to screen them out before they ever get to the shop floor.

Research has shown that using a production simulation as part of a selection process can help companies reduce incidents and injuries in the workplace. For instance, in a recent study, we looked at a sample of 130 manufacturing employees hired using a comprehensive selection process, including a physical production simulation. After employees had been on the job for 1 year, we gathered information on the number of safety incidents in which they had been involved during that time period. Both fine motor skills and work pace measured in the production simulation were significantly related to safety incidents, $r = -0.22$, $p < 0.05$; and $r = -0.27$, $p < 0.05$, respectively (Kung et al. 2012).

8.3.1.4 Positive Candidate Reactions

The reality of high-volume hiring processes, in manufacturing environments and otherwise, is that candidates will likely be screened out based on their performance on non-physical components of the selection process such as personality, cognitive ability, situational judgment, etc. In fact, they will most likely be screened out before they are able to talk with a representative from the hiring organization. Even though this might be the most valid, fair, objective, and efficient means of screening potential employees, this has the potential of leaving them with the feeling that they were not able to demonstrate their true ability.

Candidate reaction research has shown that giving candidates an opportunity to perform leads to more positive reactions regarding the perceived fairness of the selection process in particular and favorability of the company in general (cf. Truxillo and Bauer 2011). Work sample tests such as production simulations tend to elicit positive candidate reactions and are generally perceived as being fair (Robertson and Kandola 1982; Schmidt et al. 1977). Cross-cultural research on candidate perceptions of selection techniques suggests that work sample tests are viewed as among the most favorable within the United States and much of Europe (Bertolino and Steiner 2007; Hausknecht et al. 2004; Marcus 2003; Moscoso and Salgado 2004; Nikolauo and Judge 2007). Such broad findings of positive candidate reactions to work samples may be due to the high level of opportunity to perform, as well as perceptions of job-relatedness by candidates, properties that are not always evident in other types of measurement methods (Nikolaou and Judge 2007).

Consumer-oriented manufacturing companies are particularly concerned about their candidates' reactions to their experience with the company, whether they are hired or not. The automobile facility mentioned earlier was building a new plant and needed to hire about 2,000 assembly workers. They received over 75,000 applications for these positions. Only about 3 % of the people who applied would actually get a position. The local newspaper wrote a story about the hiring process and said it was more difficult to get a job at this facility than it was to get into an Ivy League college. That may or may not be true, but with 97 % of the candidate pool being screened out, the organization wanted to have the fairest and most comprehensive process in place that would not negatively impact their brand in the minds of these candidates. Basically, they still wanted people to look positively on their products even though they were not offered a job. We find this to be a common feeling among large organizations that have large candidate pools.

Granted, many candidates never made it to the production simulation because they were screened out earlier in the process. Nonetheless, well over 20,000 candidates did go through the 8-h physical production process. Informal interviews and follow-up surveys of these individuals revealed that the opportunity to go through the production simulation did have a positive impact on their perceptions of fairness of the entire selection process.

8.3.2 Drawbacks of Production Simulations

After reviewing the many benefits of production simulations, one may ask why every company does not use them. It may not make sense for all manufacturing organizations to use simulations. Listed below are some of the reasons why companies may choose not to include simulations in their selection processes.

8.3.2.1 Expense

Probably the main obstacle for companies is the expense associated with designing and implementing a production simulation. Typically, experts are consulted with regard to the design of the tasks, equipment, and scoring of the simulation. The equipment used in physical production simulations is usually custom fabricated and manufactured using real materials used in the manufacturing process. All equipment must be uniform and ensure that every candidate has the exact same experience, regardless of the equipment being used. Even if a company does have the materials and resources to build a simulation, they still need a place to put it. Physical production simulation equipment can be large and difficult to transport. The company needs to have a space to put the equipment where it can remain for extended periods of time (typically months or even years), and also allow easy access for candidates and administrators. For the automotive plant start-up mentioned earlier, the organization leased a building near the plant to set up 10 complete work stations of three simulated tasks each. The equipment alone covered approximately 2,000 sq ft of space. They hired a full staff of assessors/proctors to process candidates daily for several years until their hiring needs were met. In fact, the production simulation process is still running almost 10 years after start-up because they continue to hire employees into the facility to cover additional shifts, attrition, etc. This might seem like a lot of work, but it should be noted that many of the plants using this robust system routinely win high quality awards by J. D. Power and Associates.

8.3.2.2 Labor Intensive

Because one of the main benefits of the physical production simulation is observable, work-related behavior, trained assessors or administrators are needed to observe and evaluate candidate performance. The more automated the simulation and the scoring, the less an assessor is needed for collecting and evaluating work behavior. Depending on the level of automation and type of activity being evaluated, the ratio of candidates to assessor could range from 2:1 to 20:1. More on the evolution of automation in the production simulation process is discussed later in this chapter. Given the nature and purpose of a production simulation, a typical production simulation lasts for several hours, requiring trained assessors to also be available and dedicated during that time period. Compared to other selection tools such as personality scales, cognitive ability tests, online simulations, etc., the resources needed to run a production simulation are significantly higher.

8.3.2.3 Increased Time to Hire

One of the important metrics tracked by most human resource departments is the time to hire. This is the time it takes for a candidate to apply and move through each phase of the selection process until they are hired—the more steps, the longer it takes. They want to put into place an accurate process that will allow them to process candidates quickly. Including a production simulation in the process adds a lengthy step that requires administrative resources and takes time. In addition, the number of candidates that can be processed at one time is often limited due to the equipment and space limitations. Therefore, companies must factor in the extra time it will take to fill an open position when the production simulation is added to the process. For companies that are focused on hiring the best employees up front, a production simulation definitely helps with that goal. The automobile start-up felt that taking the extra time was important to ensure that they had a high-performing and qualified workforce from the very beginning. As mentioned previously, one organization decided to remove the production simulation from their process because they needed to fill their open positions more quickly and could not budget in the time for it. They felt the impact afterward when they had lower performing employees, poorer job fit, and higher turnover among the new hires.

8.3.2.4 Adverse Impact

As with all selection procedures, it is important to understand how subgroups of candidates perform to determine how its use might affect the diversity of the workforce. With computerized simulations that have a large cognitive component, there would likely be race differences on the performance of the simulation. For physical production simulations, there is often a gender gap, where men pass at a higher rate than women. When the majority group passes at a substantially higher rate than a minority subgroup, then adverse impact exists. While this does not mean that the assessment is discriminating against a subgroup, it does allow for the possibility of a legal challenge. Therefore, it is important for a company to document the design, development, and validation process used to support the production simulation before it is used for decision-making purposes. As long as the tasks included in the production simulation can be clearly linked to tasks required on the job, the likelihood of a successful legal challenge to the process is dramatically reduced. When designing production simulations, it is critical to determine the physical demands required, upon entry into the job, of individuals in the target positions and ensure that the simulation does not exceed those levels. If a job requires most individuals to lift 10-lb weights repeatedly during the shift and the simulation requires them to lift 20-lb weights, it may be difficult to defend that simulation.

Because production simulations are highly face valid and candidates react more positively to them, it makes logical sense to presume that candidates are less likely to challenge the validity of the production simulation as compared to other parts of the selection process. Regardless, physical production simulations are more likely than personality and other non-cognitive and non-physical components to show adverse impact.

8.3.2.5 Limited in Measurement

After reading the literature and seeing the strong arguments for the validity of work samples and simulations, one might conclude that it is the only predictor needed in the selection process. While there are many benefits to simulations, they are limited in the kinds of skills and characteristics they can measure. As mentioned earlier, most manufacturing simulations measure competencies around physical ability, attention to detail, quality focus, multitasking, processing speed, work tempo, and, to some extent, cognitive ability. When we implement a simulation for an organization, it is always in conjunction with other selection measures to ensure that we are getting a comprehensive look at a candidate. The automobile plant alluded to an earlier simulation of an 8-h workday. If that organization had only used the production simulation, they would not only have had to process tens of thousands more candidates through the production simulation, but also they would have obtained less information about their candidates. They would have had a good understanding of who might do well on the job tasks, but they would have been missing information about work ethic, teamwork, positive attitude, responsibility, and leadership potential—all competencies they were also very interested in measuring. Production simulations are best used as one of the final steps in the selection process, after other measures have screened candidates for non-physical competencies.

8.3.3 The Evolution of Production Simulations

As discussed earlier, production simulations have been around for a long time. Our experience with them goes back to the late 1980s but they were used for decades prior to that. This section describes the evolution of production simulations over the past 30 years. The evolution has coincided with two major changes in the workplace. The first has been the changing nature of manufacturing jobs themselves, especially in the United States, Canada, and certain parts of Western Europe. The second has been the incredible expansion of digital technology over that time period.

8.3.3.1 Production Simulations in the 1980s

The idea of providing a RJP during the selection process had already gained acceptance and was quite popular during the 1980s. In manufacturing environments, that often took the form of plant tours during the interview process along with a description of what it was really like to work on the floor. Production simulations greatly expanded the concept of a RJP. Instead of just showing someone where they were going to work or telling them about what it would be like to work there, why not actually have them go through something that would replicate a "day in the life"? As discussed elsewhere in this chapter, the idea then, and now, is to identify the areas that cause new employees the most problems. What are the hardest things for new employees to master? What are the things that cause the biggest complaints or reasons that people cite as reasons for leaving?

Tough, Physical, and Repetitive

One thing that has not changed much over the years is the repetitive nature of most manufacturing jobs. Many organizations like to rotate employees among a group of jobs to increase flexibility in filling in for people who are sick or who leave, as well as to reduce repetitive motion-type injuries, and to some extent improve the job environment. Nonetheless, because of the nature of most production line jobs, they end up rotating from one repetitive task to another one. Some people perform well in repetitive tasks and some do not, whether it is because they have a hard time maintaining focus over time, that they are easily bored, or that they just do not like doing the same thing over and over again.

To simulate the repetitive nature of work, most production simulations, then and now, tend to have job candidates perform several simple tasks for an extended period of time. For instance, one common task in many production simulations, particularly in automotive facilities, is to have candidates take metal tire rims from a rack, carry them for about 10 ft., and then place them onto a metal peg on a grid. It is not a particularly difficult task to learn. For instance, the instructions might tell them to do the following: (Task 1) Red rim onto A4; (Task 2) Blue rim onto B2; (Task 3) White rim on D3, etc. Assume that you have three different colored rims, red, white and blue, and a 4×4 grid going from A1 (lower left corner) to D4 (upper right corner). Most people can perform this task on the first trial. The issue is not that it is difficult to learn, but when you have to lift, carry, and place a 15-lb rim 10 ft. *over and over again* for an hour or two, it is not only tiring but certainly repetitive.

Compared with today's manufacturing jobs, jobs in the 1980s tended to be more physically demanding than they are today. Increasing the level of automation and reducing the physical nature of manufacturing jobs is a trend that has continued steadily since the 1980s. This is particularly true in the United States, Canada, United Kingdom, Germany, and several other countries. As an example, we helped open up a cold-rolled steel facility in the late 1990s that had only 128 full-time operators. It was very technologically advanced and a majority of the work involved monitoring sophisticated equipment and little, if any, physical labor. A similar plant that opened 20 years earlier required over 500 employees performing much more physical, less cognitively demanding tasks. It is often said that more jobs are lost to automation than any other cause, and this is likely to be true. Nowadays, robots perform many of the most repetitive, physically demanding jobs that used to require teams of skilled individuals.

When manufacturing jobs were more physically demanding, the focus of the production simulation tended to be guided by a theme of “make it tough and see if they can handle it.” The purpose was twofold, give candidates a solid feel for the job and see if they can make it through the simulation and measure key predictive competencies for making selection decisions. As the nature of work has evolved, so too has the focus of the production simulation. For instance, when we analyzed the most difficult things for new employees to get used to at several automotive assembly plants, it was less about lifting heavy weights than it was about working in awkward positions.

Specifically, one of the more challenging tasks involves working inside a car chassis and making wiring connections, such as connecting electrical harnesses under the dashboard. It requires a lot of twisting, bending, and fitting into very tight places. Particularly for a larger individual, it is very difficult to do on a regular basis. Because of that, some people have a hard time keeping up with the pace of the work, or they start making mistakes, such as missing connections, not securing connections, or making incorrect connections. Therefore, we focused a lot of our simulation on doing just that sort of activity. It does not require heavy lifting but it does require the candidate to get inside a chassis and make a series of connections. Then they disconnect them, replace the wire harness onto the rack, and then do it again.

Usually, production simulations are designed in stations. We have already discussed two stations that are found in many automotive plants. Those two, tire mounting and wire connections, in and of themselves, might form the basis of a 4-h production simulation. It would not be unusual to have candidates train for 5 min in each activity and then rotate between them every 30 min for the 4-h period. That way, they are performing two very different, yet repetitive, physically demanding activities over an extended period of time.

Replacing Assessors with Automation

One of the clear challenges of production simulations is their resource-heavy nature. The other challenge is how to accurately score them. Fortunately, as computers, digital cameras, video displays, and digital technology in general have become smaller, more powerful, and less expensive, both of those problems can largely be eliminated.

In the 1980s, it was common to have three participants assigned to one assessor, a 3:1 ratio. Now, there are very sophisticated centers that can assess a group of 20 candidates with one assessor, or a 20:1 ratio. The scoring of the simulations in those centers is completely automated and done in real time with much higher levels of accuracy and information than was previously possible.

To provide a better idea of how production simulations have evolved over the past 30 years, we will walk through two different exercises, which we will refer to as "bolt mount" and "weight mount," and discuss how they have gone from being assessor-intensive to almost completely automated.

Low Tech—Assessors Do It All

Traditional production simulations required assessors to: (a) explain and demonstrate to candidates how to perform the tasks; (b) observe candidates as they went through the exercise, usually making notes on a structured rating sheet; (c) score candidates' performance, usually by checking what they had done for a particular phase; (d) direct candidates to different phases, stations, etc.; and (d) make final calculations to assign final simulation or competency scores.

This original version of the "weight mount" exercise is similar to what was described above. Essentially, candidates read an instruction sheet that told them which rim to place on which grid spot, they picked up the appropriate rim, carried it, and placed it on the rack. They continued to do this until the grid was filled up. At that point, the assessor recorded what the candidate did and would then tell them to continue onto the next set of instructions. The candidate then removed the rims from the grid, replaced them on the rack, and began the process anew. Candidates earned points for both speed and accuracy. How many rims was the candidate able to accurately place on the grid over the allotted time period? How many were incorrectly placed? In addition, the assessor typically watched the candidate to make sure that they wore their PPE and followed instructions correctly. Typically, if candidates did not follow established procedures, such as carrying a rim with one hand versus two, or taking off their PPE, then the assessor rated them negatively either on "following instructions" or "safety."

This approach was neither high-tech nor multimedia in any way. Nonetheless, it was effective. For example, based on a study of 153 manufacturing employees, the work tempo component of the production simulation had the highest correlation of any competency in the assessment process with regard to supervisor ratings of work pace ($r = 0.46$; $p < 0.01$) and was also significantly related to overall performance ratings ($r = 0.30$; $p < 0.05$) (O'Connell and Smith 1999).

One challenge that this approach presents is using the assessor time effectively. The assessor was required to monitor two to three candidates, and often these candidates finished their exercises at approximately the same time. In that case, the candidate stood around waiting while the assessor recorded information for another candidate. There is essentially built in "down time," which ultimately takes away from the physical nature of the exercise.

In the “bolt mount” exercise, candidates are presented with a number of bins containing different bolts. Some are black, some silver, some white, etc. They are instructed to put a specific bolt into a spot on a metal plate that might be mounted onto a rack of some type. These plates typically have a grid pattern with a set of pre-drilled holes. There is usually either a number or letter assigned to each hole. So, candidates might be instructed to insert a “black” bolt into hole “A2.” Usually, candidates are mounting these bolts using either a cordless drill or pneumatic drill. The pneumatic drills tend to be a bit heavier and better to simulate, in terms of both feel and sound, what they will experience on the line. The bolt mount exercise is more about manual dexterity and precision than it is about physical endurance.

As far as the assessor requirements, it is similar to the weight mount exercise. After candidates have mounted 20 or so bolts, the assessor comes over and checks their work. They are interested in how many bolts were correctly mounted, how many mistakes were made, and whether the candidate followed instructions appropriately. In addition, they might penalize candidates for stripping the threads of the hole, which tends to occur when bolts are put in haphazardly. The same issues described earlier in terms of the time required for the assessor to record the candidate’s work apply to this exercise.

Automated Multimedia Production Simulations

Our most recent production simulations, those developed since about 2002, have gone to an almost completely automated administration and scoring process. We have leveraged audio, visual, and computer technology to increase the efficiency and accuracy of the physical production simulation. Technology was able to reduce the number of administrative resources needed, increase the realism of the work by reducing downtime, and provide real-time scoring and instant decision making.

Instead of assessors explaining and demonstrating to candidates, in a group, how to perform an activity, candidates now watch a video on a small liquid crystal display (LCD) right at the work station. This ensures consistent instructions for everyone, and if they need to watch it more than once, they can. It also removes the need for an assessor to be involved at this stage. In addition, each participant watches while using headphones, and they are in control of how many times they watch the instructions. Once they feel comfortable with the task, they have control over when they start the exercise and receive the first set of instructions. This also allows candidates to begin the production simulation activities at varying times without disrupting other candidates.

In the bolt mount example described earlier, all of the individual bolt mount plates are self-contained units with pressure actuators that are connected to a computer server. This allows us to place plates throughout the interior of a car or truck chassis so that candidates must sit inside and mount bolts above them, to their side, below them at awkward angles, etc. The same LCD set in front of the windshield of the cab provides all of their instructions. Once they mount a bolt into a hole, the computer senses it and records whether it is correct or not. No assessor is needed. A simulated example of the bolt mount exercise is shown in the figure below (Fig. 8.2).

Fig. 8.2 Simulated bolt mount exercise

A similar process is used for all of the other simulations in the center. This new automated approach provides several key advantages. The first is that one assessor can now proctor a group of 10–20 candidates as opposed to just 2 or 3. The second is that instructions are consistent and provided to candidates in real time. The third is that scoring is now completely objective and the level of information available is significantly higher. In earlier versions, the information gleaned from the simulation was almost exclusively limited to productivity (number of completed tasks) and quality (number of errors). The addition of computer sensors to the production simulation provides information not available with the other designs. For instance, the multimedia production simulation allows us to chart an individual's fatigue level and correlate that with performance on the job. One thing we found was that individuals who had steeply declining curves in terms of productivity and increased errors as time went on, were rated significantly lower by their supervisors. We are currently evaluating additional approaches to better utilize this information to help predict injuries, both acute and those associated with repetitive motion. The final advantage is that all scores are computed in real time and then uploaded to a server where they can be integrated into the candidate's profile, and decisions can be made immediately.

The one real downside of the automation approach is that the initial setup costs for the equipment (e.g., servers) is significantly higher. However, as long as the volume is high and the assessment runs for an extended period of time, the increased accuracy and the dramatically lower administration costs more than offset this one-time cost.

Consider another multimedia physical production simulation that focuses primarily on the monitoring and recalibration of equipment, multitasking, taking measurements, and working quickly and accurately with very few physical demands. This particular simulation was designed for a new plant start-up that had a highly automated manufacturing process. Individuals were hired into one of two primary positions: assembly station operator or machine operator. Although neither of these jobs was particularly physically demanding, making a mistake, such as not ensuring

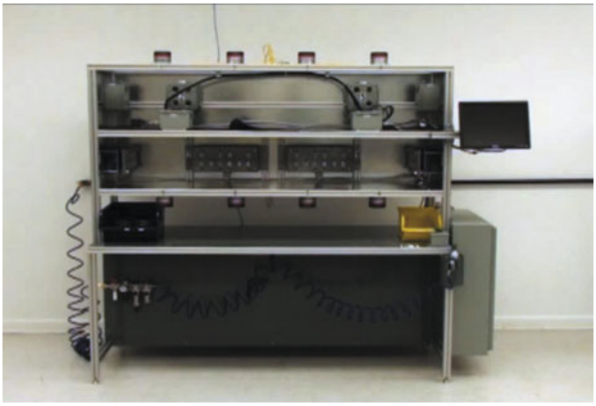


Fig. 8.3 Assembly workstation

that the machines are operating within proper specifications, could cause serious consequences in terms of quality, productivity, and ultimately, costs.

For this particular organization, we designed two distinct workstations. The first mimicked the assembly station. A photo of the multimedia assembly workstation is shown in Fig. 8.3.

Candidates who work on this station are provided with standardized instructions from the video monitor on the right-hand side of the unit, which is also where they receive all of their work instructions during the course of the 2-h simulation. Candidates are required to screw on metal caps, make connections with a wire harness, and screw in and remove bolts using a handheld pneumatic gun. The entire station has built-in sensors throughout, and the computer-based scoring system automatically scores candidates on their pace of work and whether they have followed the instructions appropriately. The three primary functions covered in this particular station accurately simulated approximately 70 % of the core activities required for all production positions at this organization.

The second station was developed primarily to assess performance in the most advanced position on the floor, that of machine operator. While this was still an entry-level position, the cognitive demands required to be successful in the position were higher than in all other positions on the floor. Figure 8.4 provides a quick snapshot of one of the core activities required of candidates in this particular simulation.

This was a truly interactive, multimedia simulation. Candidates moved continuously between three different work stations. Work instructions were provided to them by a flat-screen monitor located between the stations. One of the core activities was to keep the machines within acceptable tolerance limits. They did this by pressing various buttons that moved the readings within various gauges up and down. The individual in this figure is pressing a button to lower the reading on one of the three gauges on the screen. Gauges moved automatically and in an unpredictable manner based on several algorithms built into the system. Candidates needed to stay constantly vigilant to keep them within the tolerance limits. They were also required to take physical measurements and record their readings into the computer. At any

Fig. 8.4 Example of interactive machine operator station

given time, they would be monitoring and adjusting six sets of gauges, resetting their machines and recording product counts that appeared on the screen, and also making and recording measurements of a variety of articles, all based on the instructions presented to them on their computer monitor. As with the other automated simulations described thus far, all scoring was done in real time and automatically by the computers hooked up to these stations.

Reactions both from candidates as well as the organization were very positive regarding the realism of the simulations and, most importantly, to the quality of hires made based on these systems.

8.4 Developing a Multimedia Production Simulation

Developing an effective multimedia physical production simulation takes a lot of time and effort, but companies should experience a return on their investment in terms of higher productivity, reduced turnover, accidents, and injuries. The recommended steps involved are described in detail below:

1. Define the job family—Given the time and expense associated with the design and development of a multimedia production simulation, using it to hire for just one position is costly. Typically, the simulation is used for positions within a larger job family. For instance, it is the norm rather than the exception that a manufacturing plant will assign the title of "production team member" or "production employee" to a broad range of positions covering different functions and departments. This applies to skilled or maintenance positions as well as non-skilled positions. This is a relatively dramatic shift when you compare it to traditional manufacturing facilities and job titles. If you were to look at the number of "job titles" in a traditional automotive facility, you would likely see over a hundred individual job titles. Compare this to newer facilities that may cover those hundred job titles

with three to four primary positions. It still remains that unionized facilities typically have more job titles and non-unionized ones have fewer job titles, although there are plenty of exceptions to that rule. Many newer unionized plants have moved toward broader positions with fewer job titles. This broadening of the job families has salutary implications for both employers and employees. It allows easier cross-training, more flexibility in allocating resources, and more task variety for workers. To use the example of an automotive manufacturer, a production employee in "assembly" is likely putting parts on vehicles as they move along the assembly line. This may include installing the dashboard, installing and securing the steering assembly, electrical wiring harnesses, seats, etc. A production employee in the "stamping" area helps fabricate, or stamp parts out of metal (e.g., doors) that are then put on the car either by robots or by workers in assembly. In this job, there is often a higher level of equipment monitoring, such as stamping presses, required to make sure they are working correctly and that they are properly loaded, etc. Other production employees work in the paint, body weld, final quality inspection, and conveyance (i.e., moving parts and supplies around the plant) departments. They are all in the same job family but do not do the same physical tasks. When candidates apply, they do not apply for a specific department, and they can move around within the organization to other departments throughout their career. With so much task variance within the job family, the production simulation should measure skills needed by all of them. Therefore, it is important at the beginning of the design process to know exactly which jobs will be filled by individuals who complete the production simulation.

2. Analyze—Once you have a good understanding of the jobs involved, it is important to better understand them from the standpoint of the tasks involved, reasons for failure, reasons for turnover, and common injuries or accidents. This information can be valuable when deciding what aspects to build into the design of the simulation.
    a. Job analysis— Learn as much as possible about the jobs in the job family. Tour the facility, observe the jobs, shadow incumbents, work the job yourself, and talk to incumbents and leaders. The more you know, the better the final simulations will be. If the plant is a start-up, visit a different facility with similar jobs or talk to the leadership about the vision for the position.
    b. Turnover analysis—Certain aspects of the job itself can lead to turnover. If available, gather information about why individuals have left the organization—voluntarily and involuntarily. Identify the reasons that speak to job design and determine if the simulation can and should include activities to address them. For one manufacturing company, one of the main reasons for turnover was the monotony of the work. Because of the pay and benefits, individuals who were trained for other careers (e.g., school teachers) were applying for manufacturing positions. Once hired, they had difficulty being tied to the production line for hours without being able to leave for a break. They also missed the intellectual stimulation they were accustomed to in previous jobs. As a result, a production simulation was built that required candidates to do the same task for close to 2 h and then rotate to two other tasks both

for another 2 h each. The job rotation and length of time resembled the job as closely as possible. Instead of screening out individuals without previous manufacturing experience, this company allowed the experience with the production simulation to help make the hiring decision. Turnover was greatly reduced after the production simulation was introduced.

   c. Injury/accident analysis—Knowing as much as possible about common injuries and the types of accidents that have occurred can aid in the design of the production simulation. A Canadian manufacturing facility actually included their occupational therapist as part of the design team. She was responsible for working with engineers and technicians to design the work processes in a way to reduce physical strain and repetitive injuries. Not only did she want to ensure that new hires could perform the physical tasks, she wanted to make sure that candidates did not get hurt while completing the stimulation.

3. Identify minimum physical requirements—As mentioned earlier in the chapter, the complexity and difficulty of simulated tasks should not exceed those required on the job. It is important to identify the minimum requirement for a specific task before building it into the production simulation. For example, it is common for manufacturing workers to lift heavy parts, boxes, or other objects as part of their job. However, not all jobs require heavy lifting or lifting of the same amount of weight. At one manufacturing facility, there were a few jobs that required employees to lift up to 50 lb regularly as part of their job, but most of the jobs never required lifting more than 20 or 25 lb on a regular basis. When designing the production simulation, an exercise was designed where candidates were instructed to move weights back and forth from one location to another. During the design phase, it was decided that the heaviest weight would be 25 lb and it would be the least moved amount of weight by the candidates during the exercise. If the company had chosen to make the heaviest weight 50 lb and instructed candidates to move it often during the exercise, they would have been requiring candidates to work harder than is typical on the job. By doing so, they would have screened out individuals who could have adequately performed the vast majority of the jobs. In addition, because the individuals who would have been screened out were more likely to be women and older workers (both protected classes), there would have been a greater likelihood of a legal challenge. It is difficult to defend using a higher standard than what is required for the most likely target positions. Be aware of this issue when deciding what to require, and be sure to tie the equipment back to the job and have job-related rationale for the decisions being made.
4. Identify common tasks—Similar to the previous step, it is important to identify the tasks most common to the jobs in the job family. You want to simulate the jobs as closely as possible and measure the skills that will get you the best information for predicting a candidate's performance once hired. Because you are looking across jobs within a job family, there will be very few tasks that are present in every single job; however, there will be skills that are common to a large group of them. Like the minimum requirements, be sure to identify tasks that are job related and do not require more of the candidate than the jobs in the plant. All decisions should be documented with job-related rationale.

5. Determine skills and method of evaluation—All the steps up to this point are recommended for production simulation design—multimedia or not. It is at this stage that the multimedia technology makes an impact because it allows the developer an opportunity to easily measure some skills that were difficult, if not impossible, to measure previously (e.g., time per task), and to do so more accurately. In general, production exercises evaluate speed (number of tasks completed) and accuracy (percent completed accurately). Computers can easily calculate those measurements very accurately and provide the ability to look for data trends as well (e.g., stamina or accuracy over time). Our experience has been that it is better to limit the number of competencies measured in the production simulation. Attention to detail (also known as quality orientation) and work tempo (speed) are the most commonly evaluated measures in a production simulation. We have included other competencies, but when we analyzed our results, they typically gravitated to those two primary factors. Other competencies such as multitasking, fine motor skills, following instructions, or safety orientation are also competencies that lend themselves to assessment in such simulations. Nonetheless, it is better to get solid, accurate, and reliable measurement on a few areas than trying to force fit too many competencies into the equation just for the sake of covering more competencies. After identifying the general skills to be measured, the next step is to decide how to score the exercises. Instead of requiring an assessor to observe, record, and score each exercise, a computer does it all. Investigate the technologies that exist to be able to give you the feedback that is needed throughout the exercise. In most cases, simple electric switches that feed data back to a server are adequate for telling the computer if a candidate was right or wrong. Knowing these things ahead of time is important because it will affect the design and fabrication of the exercises.
6. Design—There are many factors to consider when designing a multimedia production simulation. Some of the factors to consider are listed below:
   a. Administration—When designing the exercises, keep in mind how the candidate will receive instructions. Using instructional videos and on-screen instructions provides each individual with clear, consistent instructions on how to complete the exercise. It is also recommended to build in a short practice time for the candidate on each exercise before the scored exercise begins.
   b. Number of exercises—The number of exercises to be developed is often driven by time and budget. If the target job(s) offer job rotation, the simulation should also offer this opportunity. However, if the target job requires workers to complete the same task for the entire workday, the simulation would be more beneficial by not rotating candidates.
   c. Type of exercises—Remember that employees in the organization have been trained on how to complete their job tasks. In some cases, that training lasts for days or weeks to ensure that they are able to effectively perform the tasks. In a production simulation, candidates have a very short period of time to learn the task and then perform it. The tasks should be straightforward enough that every candidate, regardless of knowledge and experience, has the same

understanding of the task at hand. It is not necessary for the production exercise to simulate an exact job within the plant. For example, let us examine the task of fastening screws and bolts. The manual dexterity needed to perform the task is what should be measured in the production simulation. The screws can be fastened onto a metal plate on a table. However, if screws and bolts need to be fastened in awkward positions or overhead, the simulation should build this into the task as well by placing metal plates high and low and in hard-to-reach places. Keep in mind the underlying skill that is being measured, and keep the task simple enough for candidates to learn easily. In addition, once the exercises have been determined, the next step is to investigate the technology available to be able to build a simulation that is able to collect the important data needed to measure the skill or competency of interest. One point to consider when using technological equipment for assessment is that parts can malfunction or break. When designing the system, ensure that there are "checks" in place to identify switches or other equipment not working prior to assessing a candidate, as broken equipment could affect the final score of a candidate.

d. Length of simulation—The amount of time that the candidate should take to complete the entire simulation can be difficult to determine. The shorter the simulation, the more candidates that can be processed and the less time resources spend administering and scoring the assessment. However, if the simulation is too short, then it does not adequately provide candidates a good preview of the actual work. When a task is new, monotony and boredom do not set in right away. It is important to give candidates enough time to get tired of the task before switching them to a new one or ending the simulation.
e. Ergonomics—When building the equipment for the simulation, it is important to allow for adjustments to accommodate different heights. If a candidate is particularly short or tall, making incorrect motions over time on some tasks can lead to injury or put them at a disadvantage in terms of performance. A simple rule is that if the organization offers accommodations on the job, then you should incorporate similar accommodations during the simulation.

7. Pilot test—After the tasks have been designed and the equipment has been fabricated, pilot testing should be conducted to test the accuracy of the design and establish a reference point for scoring candidates. For this step of the process, individuals who are unfamiliar with the simulation should act like candidates and complete the simulation. Some companies have even hired a temporary agency to provide a group of "fake" candidates. Job incumbents are not the best population to use for this step of the process because they have on-the-job experience, which might give them an advantage when completing the exercises. The data gathered from the pilot allows the individuals involved in the design to identify errors and improvements and set scoring (including initial cut scores) for the assessment. The pilot testing is also an ideal time to train administrators and assessors on their roles.
8. Validate—As is often the case, a predictive validation study is recommended because it allows you to examine the relationships between the simulation and

job performance, turnover, safety incidents, and other outcome variables. The drawback is that you may need to wait 2 years or more before conducting such a study. Conducting concurrent validation studies, using incumbents in the job, is also an appropriate strategy. While incumbents may have an advantage over job candidates, we have found that you typically still obtain acceptable levels of variance of performance in the exercise. You will not be able to measure turnover, but you can look at other criteria measures and establish validity linkages. As a solid starting point, production simulations lend themselves to content validation strategies. If designed well, a production simulation epitomizes a content valid approach. This requires drawing clear linkages between tasks on the job to tasks in the simulation. If you are not able to establish the content validity of the production simulation then you probably have not designed a good production simulation.

8.5 Summary

Manufacturing environments require a broad range of skill sets, abilities, and motivations that clearly lend themselves to simulations in general and production simulations in particular. The results are consistently strong in terms of the validity of such simulations. The perceived fairness and candidate reactions to such simulations also appear to be positive.

Another benefit of such simulations, that is not widely discussed or published, is their face validity and perceived fairness by internal stakeholders. On the surface, that may not seem to be that important, but it has a number of beneficial outcomes. The first is that it helps ensure that higher validity solutions, such as comprehensive assessment center/production simulations, are given appropriate consideration in the hiring process compared with less valid, more subjective methods such as hiring manager interviews. While interviews are certainly valid predictors of performance, our experience has been that production managers tend to overemphasize previous manufacturing experience to the detriment of almost anything else. While research does suggest that previous manufacturing experience is moderately related to reduced turnover (cf. O'Connell and Kung 2007) it is only modestly related to performance, especially compared with a well-designed production simulation. During times of plant start-ups, we have found that production and other line managers tend to be overly cautious in accepting candidates onto the shop floor. This is not without reason. It is in their best interest to have the best possible employees in the plant from the moment the plant starts production, and even before. However, in their zeal to "hire the best," they unfortunately tend to fall back on bad habits and implicit biases and ultimately screen out some individuals who would be predicted to be fantastic employees because they "have never worked in manufacturing" or "have worked in manufacturing but not like ours. . . " Because of factors such as financial incentives and lower labor costs, many plants open up in areas that do not have well-established manufacturing bases and, therefore, it is difficult to find individuals who have relevant manufacturing experience. That is not a problem as long as the company has a fair and accurate methodology for screening candidates. The production simulation and other manufacturing simulations provide such an approach. It usually takes some education and time to break old habits, but when production managers see what candidates have to successfully complete in order to make it to the final interview they are more likely to trust the system than go by their "gut feel." This is a huge benefit to the human resource professionals at the plant tasked with staffing the facility.

Leveraging recent technologies and applying them to multimedia production simulations, as well as interactive computer/web-based simulations, is a smart way for organizations to gain predictive power to their selection processes without extra resources. They have proven themselves many times over to be effective, fair, and accurate methodologies for assessing candidates and making predictions of their likelihood of success in a manufacturing environment. The future of these types of simulations is likely to mirror the environments they are designed to simulate. As described in detail earlier in this chapter, manufacturing jobs continue to move away from single task activities and instead are shifting toward those that require more decision making, multitasking, and collaboration. Manual dexterity, processing speed, work pace, stamina, and attention to detail will continue to be important competencies in almost any manufacturing job. Well-designed multimedia production simulations are the most fair and accurate method of measuring such competencies.

References

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519–533.

Avis, J. M., Kudisch, J. D., & Fortunato, V. J. (2002). Examining the incremental validity and adverse impact of cognitive ability and conscientiousness on job performance. *Journal of Business and Psychology, 17,* 87–105.

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

Bertolino, M., & Steiner, D. D. (2007). Fairness reactions to selection methods: An Italian study. International Journal of Selectionand. *Assessment, 15,* 197–205.

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8,* 248–260.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and incremental validity of narrow traits. *Journal of Applied Psychology, 91,* 40–57.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24,* 79–132.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance, 15,* 25–46.

Hausknect, J. P., Day, D. V., & Thomas, S. C. (2004). Application reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57,* 639–683.

Kinney, T., & O'Connell, M. S. (2012). Are high-fidelity multitasking simulations more than just a pretty face? Paper presented at the 26th annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Kinney, T., Reeder, M., & O'Connell, M. S. (2009). Comparing multitasking and cognitive ability as predictors of job performance. Paper presented at the 24th annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Kung, M. C., O'Connell, M. S., Tristan, E., & Dishman, B. (2012). Simulate the job: Predicting accidents using a work sample. Paper presented at the 26th annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Lawrence, A. D., Doverspike, D., & O'Connell, M. S. (2004). An examination of the role job fit plays in selection. Paper presented at the 19th annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

Marcus, B. (2003). Attitudes towards personnel selection methods: A partial replication and extension in a German sample. *Applied Psychology: An International Review, 52,* 515–532.

Moscoso, S., & Salgado, J. F. (2004). Fairness reactions to personnel selection techniques in Spain and Portugal. *International Journal of Selection and Assessment, 12,* 187–196.

Murphy, K. R. (2002). Can conflicting perspectives on the role of g in personnel selection be resolved? *Human Performance, 15,* 173–186.

National Association of Manufacturers (2013). Facts about manufacturing. http://www.nam.org/Statistics-And-Data/Facts-About-Manufacturing/Landing.aspx. Accessed 15 Jan 2013.

Nikolaou, I., & Judge, T. A. (2007). Fairness reactions to personnel selection techniques in Greece: The role of core self-evaluations. *International Journal of Selection and Assessment, 15,* 206–219.

O'Connell, M. S., & Kung, M. C. (2007). The cost of employee turnover. *Industrial Management, 49*(1), 4–20.

O'Connell, M. S., & Reeder, M. (2008). Meta-analysis of the select assessment ®for manufacturing. Technical report. Select International, Inc., Pittsburgh.

O'Connell, M. S., & Smith, M. (1999). Normative and validity results for entry-level assessment scales. Technical report. Select International, Inc., Pittsburgh.

Premack, S. L., & Wanous, J. P. (1985). A meta-analysis of realistic job preview experiments. *Journal of Applied Psychology, 70,* 706–719.

Reeder, M., Kung, M. C., & O'Connell, M. S. (2008). Incremental validity of work sample measures beyond personality and cognitive ability. Paper presented at the 23rd annual meeting of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. *Journal of Occupational Psychology, 55,* 171–183.

Salgado, J. F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology, 82,* 30–43.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15,* 187–210.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30,* 187–197.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703–742.

Truxillo, D. M., & Bauer, T. N. (2011). Applicant reactions to organizations and selection systems. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol 2: Selecting and developing members for the organization. APA handbooks in psychology* (pp. 379–397). Washington, DC: American Psychological Association.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372–376.

Chapter 9
Simulations for Service Roles

Joe LaTorre and Mary Ann Bucklan

9.1 Evolution of Simulations for Service Roles

9.1.1 Market Forces Driving the Need for Service Simulations

The past several decades have seen a significant shift in the American labor market. There was a time, in the not so distant past, that manufacturing jobs dominated the landscape for job seekers in pursuit of entry-level positions. These entry-level manufacturing roles represented a starting point for workers. American workers could then parlay these roles into lasting careers. These manufacturing careers sometimes presented advancement opportunities in management, sales, or other operational areas within the organization.

Competitive advantage in those days was primarily achieved by product features, benefits, and innovation. Whoever could develop the better mousetrap usually had an advantage in the marketplace that could be sustained, at least for a reasonable amount of time. While it is true that the best organizations always valued high levels of customer service, they could excel with adequate service levels, if they did in fact have a superior product. Well, it is safe to say that those days are gone. Yes, you still need an outstanding product to compete but that is just the ticket to get into the show. If you want to be the star, you need to differentiate yourself in other ways, and the best way is through your people. Not just the technical people behind the scenes, or the executives in the board rooms, it is the people who *are* your company to your customers—your customer service providers. They are the ones that form the corporate image with customers, in each moment of truth, that will make or break the organization. Yes, outstanding products are a requirement, but with the technological

J. LaTorre (✉) · M. A. Bucklan
Employment Technologies Corporation,
Winter Park, FL, USA
e-mail: joe.l@etc-easy.com

M. A. Bucklan
e-mail: maryann.b@etc-easy.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection,*
DOI 10.1007/978-1-4614-7681-8_9, © Springer Science+Business Media New York 2013

advancements available today, differentiation based on product alone is short lived. It is the service people who often make the difference between success and failure.

It is also a reality that there are fewer manufacturing positions available (Dymond 2012). Offshoring for cheaper labor, robotics, and automation have reduced the number of manufacturing positions required here in the United States. In contrast, the number of service jobs available remains significant. A plethora of position openings exist in the service sector, in growth industries like health care, retail, banking, and hospitality (Murray 2009). Although entry-level service positions will continue to be available, competition among candidates will be fierce. Hiring organizations are likely to continue to see wave after wave of candidates applying for these entry-level service positions. This will enable hiring organizations to be selective as they strive to fill these positions with new hires who will make a positive impact on service culture and help secure a competitive advantage.

Being selective drives the need for highly accurate assessments that can predict candidate service, skills, and aptitudes. All of these conditions have led to an increased need for service-related simulations.

9.1.2 Automation of Traditional Assessment Center Technology

Simulation technology is not new. For decades it has been used for training and development in the military, transportation, and health care industries, among others. The use of simulations for hiring initially was focused on mid-manager to executive-level positions. These simulations were usually in the form of low-tech managerial assessment centers. These simulations consisted of multiple job-related exercises that required participants to complete work activities that mirrored the responsibilities of the target position. These assessment centers required administrators, role players, and multiple assessors. They also required business travel so participants could reach the location of the assessment center. Although this approach was and still is highly accurate, it is costly and time consuming. Using traditional assessment center methods for a high-volume entry-level customer service role is simply not feasible from a time or cost perspective. The market desire for the accuracy of the assessment center approach at an entry-level assessment price drove the evolution of simulations for service roles. In the late 1980's, we at Employment Technologies Corporation (ETC) began developing some of the first service-related simulations in the industry. We were the founding members of a small think tank charged with finding a way to make the assessment center approach both feasible and affordable for high-volume entry-level service positions.

To accomplish this, we needed to achieve three objectives:

- Eliminate the need for travel.
- Eliminate the need for role players.
- Eliminate the need for assessors.

The solution to all three of these objectives was technology. To eliminate the need for travel, we created the simulation content so it could be electronically delivered to the participant instead of the participant going to the simulation. To eliminate role players, we elected to use video vignettes instead of live role players to depict the work-related situations. To eliminate the need for assessors, we adopted a multiple-choice item format that could be automatically scored without human intervention.

9.1.3 Early Examples of Service-Related Simulations

Our earliest forms of service simulations were created and delivered on VHS video cassettes. These simulations usually included a brief realistic preview of the target position. The job preview would show scenes of the work environment and describe the key responsibilities of the job. The preview would also present the work orientation factors associated with the job and highlight both the pleasant and unpleasant features associated with the job. These job previews help establish clear expectations of the job and help reduce short-term turnover by providing candidates with clear expectations about the job (Wanous 1989). The job previews were followed by in-depth instructions for completing the simulation. We also believed it was important to provide candidates with information about the skills being measured in the simulation and how their performance would be evaluated. Finally, these early simulations included a practice scenario that would enable candidates to gain a clear understanding of how the simulation works before beginning a formal assessment period. The simulation content itself consisted of a series of video vignettes portraying critical situations typically encountered by employees in the target position. At key points in the video, candidates were asked to respond to the situation as if they were actually performing the job. The candidates would review a series of potential responses and select the one they believed to be best. These early video-based simulations were administered at a hiring location using a proctor. The proctor would be required to start the tape and observe candidates complete the simulation. Often, the simulation was administered in a classroom setting with multiple candidates completing the simulation at the same time. Once the tape started, the proctor would ensure that the tape was not stopped or rewound, to ensure a consistent simulation experience for all candidates. In these early days, candidates were provided with paper test booklets that contained a list of the questions and response options, as well as answer sheets to record their responses. These answer sheets were then collected and responses were either entered or scanned into proprietary scoring software. The scoring software would automatically score the responses and provide a feedback report for each participant who completed the simulation.

One example of these early service simulations was our first-generation teller assessment, called the Teller Assessment Program (TAP). TAP was developed and released in 1989. TAP was embraced and successfully used by several of the largest banks in America including Fleet Financial, Huntington Bank, and BB&T. TAP was a video-based simulation consisting of a series of service transactions. Candidates were required to view these transactions and respond as if they were the teller completing the transaction. This simulation included a basic set of banking policies and procedures for the bank that would help candidates respond to the situations within the parameters of the bank. Another early example was the Customer Service Skills Assessment Program (CSSAP), developed and released in 1992. To develop this service-based simulation, ETC partnered with Saturn Corporation. Saturn was interested in redefining the customer experience within their car dealerships. Their goal was to create a service culture unlike anything associated within the automotive industry at the time. To achieve this vision, they would need to recruit and hire top-tier service providers. Saturn turned to us to develop an accurate service predictor for all dealership personnel who could potentially come in contact with customers at the dealership. To address this need, we created a video-based simulation that depicted service-related scenarios across a series of different service positions within the dealership. Although the service roles and the scenarios differed by position, the service skills that were required and measured by the simulation were universal across the dealership. This enabled us to create a single simulation that proved to be a valid and accurate predictor across all entry-level service positions within the dealership. Saturn used this simulation to launch their entire new dealership network. The success of the simulation was evident based on Saturn's outstanding customer satisfaction ratings related to the on-site dealership experience. This simulation was eventually upgraded for on-line delivery, enabling candidates to enter responses on-screen. A sample screenshot from this simulation is provided in Fig. 9.1.

9.1.4 Ongoing Advancements and State-of-the-Art Examples

These early video-based simulations continued to evolve based on the technology available. VHS-delivered simulations transitioned to DVD and ultimately to computer delivered simulations that provided candidates the ability to respond directly on-screen using the keyboard and mouse. The computerized simulation approach essentially eliminated the need for paper support materials. All of our simulations are now paperless. The instructions, support materials, and help functions are all self-contained within the computer-delivered simulation. This further ensures the consistency of the process across all candidates. The emergence of the Internet and the market's desire for delivering unproctored simulations has led us to design all of our simulations for web delivery. This trend is expected to continue into the future. When creating simulations, it is critical to consider the audience, the technology platforms, and the devices that candidates will use to complete the simulation. Market demand for versatile simulations that can be delivered on multiple platforms including handheld devices will continue to grow.

In addition to technology changes, we must also continue to evolve and advance the simulation experience for candidates. The use of virtual environments and gaming techniques will also be key factors that ensure the ongoing success of simulations

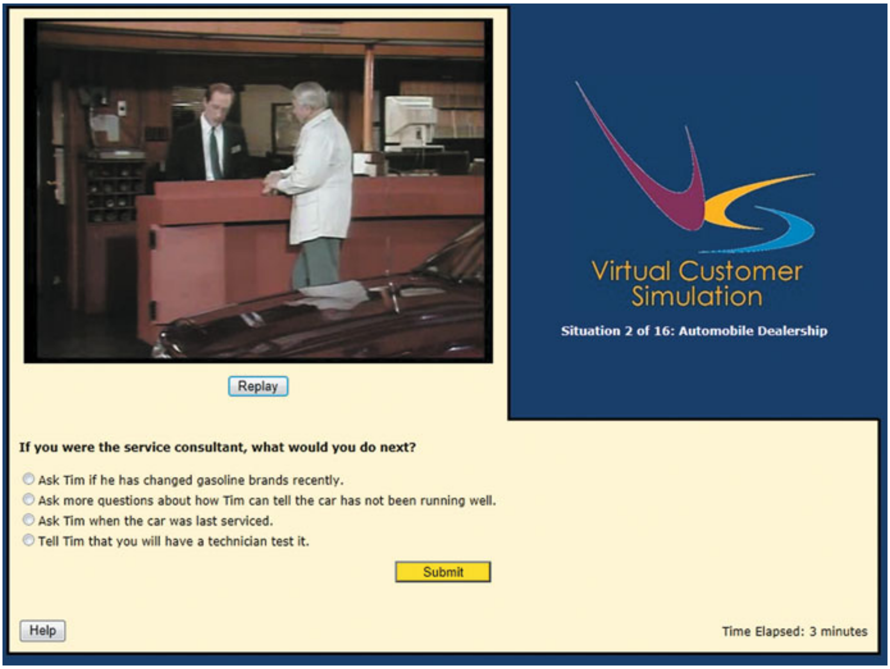


Fig. 9.1 Sample screenshot from early simulation

for service roles. Figure 9.2 provides screenshots from some of our current service simulations that utilize these techniques.

Candidate reaction data that we have gathered indicates that the use of animation and gaming techniques generally improves their perception of preemployment testing, leading to an enjoyable and entertaining experience for the candidate (ETC 2012). This has only positive effects for the hiring organization by expanding recruiting pools and perpetuating a progressive and positive corporate image that is important to young job seekers vying for entry-level service positions.

The technology available to develop and deliver simulations will continue to evolve. As simulation developers, we need to continue to evolve and adapt with it. Expect to be challenged for more mobile delivery options and higher expectations related to the overall sensory experience delivered via service simulations. We can also expect the market to continue to demand accurate, fair, and legally defensible simulations. Hiring organizations want to retain the accuracy of simulations provided, but they will want the simulations to be shorter. This poses new challenges for simulation developers. We will need to leverage all of the lessons learned and blaze new paths in research and development to meet these challenges and continue the expanded use of simulations for hiring service roles.

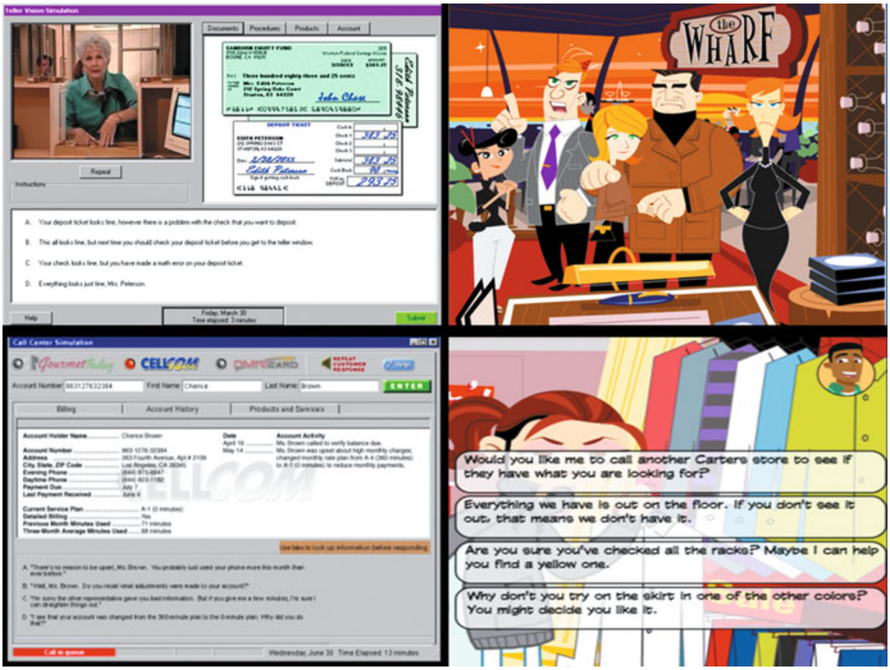


Fig. 9.2 Sample screenshots from current service simulations

9.2 Development Considerations and Processes

9.2.1 Understanding the Target Jobs and the Simulation Audience

When developing simulations for service roles, the first step is to gain an understanding of the target role. To create a realistic and accurate simulation, you must have an in-depth understanding of all the nuances of the job. Our simulation development process always begins with an in-depth job analysis. Our job analysis process employs several data gathering methods. On-site observations are combined with job analysis interviews with current incumbents and supervisors of incumbents in the target role. We use a standard job analysis interview protocol to identify the key elements of the position. The interview is focused on defining the key responsibilities of the target role as well as the specific work tasks performed under each key responsibility. The interviews should also contain questions to identify the knowledge, skills, abilities, and other characteristics (KSAOs) required to successfully perform the work tasks of the target role.

The job analysis effort should also focus on identifying critical incidents experienced by incumbents in the target roles. These critical incidents are the service-related scenarios that define success or failure in the target role. For service roles, these critical incidents should involve key customer interactions that make or break the service experience for the customer. These critical incidents should be used in developing the subsequent simulation script. Creating a script that is based in truth and contains adaptations of real life situations will contribute to the realism and face validity of the simulation. Completing on-site observations is also critical. On-site observations provide the simulation developer with a rich picture of the work environment as well as the context and tools used when performing the target role.

The data gathering process should result in some form of a job profile that documents the information gathered in the process. The job profile should clearly describe the work performed by incumbents in the target position, as well as the KSAOs required by the workers who perform the target role. It is also a good practice to include a test plan for the subsequent simulation. The test plan should define the key characteristics of the simulation, as well as a listing of the types of situations for inclusion in the simulation. This document should clearly define the competencies that will be measured by the simulation, as well as a description of how these competencies will be measured by the test items. These initial job analysis efforts will lay the foundation needed to create an accurate and realistic simulation of the target role.

Finally, it is critical to gain an in-depth understanding of the audience for your service-related simulation. Take a close look at the demographic profile of the candidate pool for the target position. Consider factors like age, experience, education level, and language proficiency. Be sure that you develop and design the simulation at a level that is consistent with the intended audience. If the majority of your candidates are high school graduates with no postgraduate education, do not create a simulation that requires an advanced reading level. Be sure to avoid any cultural issues that may be misunderstood or misinterpreted by a segment of the candidate pool. Your general guiding principle is to design a simulation that is fair for all candidates, regardless of prior work experience. Simulations measuring general service competencies should provide an equal opportunity for all candidates to succeed if they have an understanding of basic customer service skills and a strong service aptitude. Also, be sure to consider the technology platform and devices that are accessible to the candidate pool. For example, do not create a simulation that can only be delivered on one platform if only a small percentage of the candidate pool has access to that platform. It is always best to develop service-related simulations that can reach the largest potential audience across technology platforms and devices. Current development tools that cut across different operating systems provide flexibility for simulation design and allow your simulation to be experienced by any end user that has high-speed Internet access.

9.2.2 Selecting a Multimedia Approach

One of the key decisions when developing service-related simulations is deciding on a multimedia approach to use when presenting the simulation content. The good news is that today's technology provides many options. Rolling video, animation, virtual environments, and audio stimuli are all viable options. When choosing a multimedia approach, there are three main factors to consider:

- The nature of the target role
- The target market(s) for the simulation
- The demographics and preferences of the simulation audience

Consider how the target role is performed. Are the service interactions face-to-face or are they performed over the phone? Your multimedia choices should reflect the actual way in which the target role is performed. If the interactions occur over the phone, then use a multimedia approach that features customer voice only. If job incumbents need to identify customer moods and conditions using audio cues only, then only include audio cues in your simulation. If the interactions occur in a face-to-face environment, then depict the customers visually in your simulation. Use live video or animation to depict your customer scenarios. In general, a good service-related simulation should strive to recreate the actual job conditions as closely as possible. You want to provide candidates with an opportunity to experience job-related situations as if they are actually performing the target role. This will increase the accuracy of your simulation as a predictor of job performance and increase the accuracy of the realistic job preview you are providing to candidates.

It is also essential that you consider your target audience when selecting a multimedia approach. Every service simulation has its own look and feel. Consider the industry and potential target markets for the simulation. For example, the banking and financial industry is still fairly conservative in nature. In this industry, it may not be a good idea to create a simulation that resembles an on-line video game chock full of ringing sound effects and humorous animated characterizations of customer interactions. On the other hand, if you are developing a service simulation in the retail or hospitality industry, the organization may be more amenable to a highly visual gaming style approach. Also, consider the candidates who will be completing the simulation as part of the hiring process. Consider their age and their expectations around multimedia. If possible, you should strive to create service-related simulations that are not just accurate and fair but fully engaging to the audience who will experience them. Lastly, make sure the multimedia approach you select does not restrict the delivery of the simulation to your audience. Do not use a multimedia format that can only be displayed on a single operating system or device. Always seek an approach that will enable you to reach the largest potential audience.

9.2.3 Developing Simulation Content

As this book demonstrates, there are many approaches to developing simulation content. Some approaches are more appropriate for certain end uses than others. The approaches described here are most appropriate for service simulations. As mentioned earlier, almost all simulation development projects begin with a thorough and comprehensive job analysis that includes critical incidents. As you identify critical incidents, go beyond the data gathered from Subject Matter Experts (SMEs). Do not be afraid to rely on your own experience as a source of critical incidents. Unlike a technical or manufacturing simulation, we can call upon personal experience to supplement the critical incidents that are supplied by the SMEs. We have all been customers or clients, and we can all readily recall incidents of good service and poor service. Some of the situations presented in our Virtual Customer—Service simulation are based directly on our test developers' personal experiences with service providers.

During the job analysis, it is essential to identify the competencies that are important for good performance and that underlie effective job performance. Not all competencies are appropriate for inclusion in a job simulation. For example, work ethic is consistently rated by employers as one of the most important qualities in an employee (Hill and Petty 1995). However, it can be difficult to assess this in a 45-minute simulation. This is an example of a competency that is better suited to assessment elsewhere in the selection process.

When considering whether to attempt measurement of a competency in the simulation, be sure that the competency can be specifically defined in behavioral terms. If it cannot be defined in specific behavioral terms, then it cannot be simulated and therefore does not belong in the assessment. Table 9.1 presents an example of general and specific definitions of a competency.

The work performed in service roles varies considerably, but almost always involves interaction with others. Therefore, one basic decision will be how to simulate these interactions. For example, when we developed a simulation for collections agents, the job analysis revealed that the vast majority of these interactions take place over the phone. In this instance, it was particularly important to record realistic audio, using professional voice talent to represent the wide range of emotions and situations that are typically encountered by collections agents in their dealings with consumers. There was no need to employ video- or avatar-based stimuli.

Beyond interacting with customers/clients, service providers may have other unique job tasks. For example, some service providers have to summarize their interactions at the conclusion of a call or visit. This involves accurately recalling and understanding the key issues and effectively and efficiently summarizing them. Be creative with question types and simulation content to capture these responses and measure these competencies. When simulating content, it is important to work closely with the technical team and computer programmers to ensure that design ideas can be effectively translated from a technical standpoint.

Finally, always develop more simulation content than is needed. Some of the content will not work well when the SMEs review it, some may drop out during the validation phase, and some ideas may seem good on paper, but are impossible to simulate.

Table 9.1 Definitions of friendliness

General	Specific
Displays kindness and helpfulness toward others, comes across as amiable and genial, treats each customer with respect, and makes sure they are satisfied with their service	Courteously greets customers, consistently smiles when interacting with customers, uses a pleasant tone of voice and appropriate language when talking to customers, is approachable and easily enters into conversations with customers, displays empathy and consideration for customers' problems, maintains composure in challenging situations, and thanks customers at the end of interactions and lets them know they are valued

9.2.4 Involving Subject Matter Experts in the Design Process

SMEs should be involved during all steps of simulation development. In addition to participating in the job analysis phase, SMEs can be used to review simulation content, specify optimal and preferred/less preferred response options, assist in development of the scoring key, and participate in beta testing the simulation.

Different SMEs will have different perspectives on the service role that is being simulated. SMEs may typically include incumbents (those occupying the target role), supervisors of incumbents, and higher level managers and executives. Each of these SME sources is uniquely positioned to provide important information that can aid in simulation development. Table 9.2 provides a description of SME sources and other considerations.

It is also important to use a diverse mix of SMEs. One obvious goal should be to include a diverse mix of SMEs with respect to gender and ethnicity. But do not forget more subtle indications of diversity, such as SMEs from different geographical settings. During one project when we were working with a large telecommunications organization to develop a simulation for retail store salespeople, we learned that there were several differences between retail sales people in urban versus rural stores. While the underlying competencies necessary for success were consistent across settings, there were large differences in the types of customers, store characteristics, and even the tasks performed.

SMEs, like all of us, are busy, time-challenged professionals. From an SME perspective, time spent working with a simulation developer means time away from his or her "real" job, which could mean lost sales opportunities, delayed responses to e-mails, or time logged out from a work station. Furthermore, in many service roles, there are certain times of the year that are busier than others. During these times, the availability of SMEs may be limited or nonexistent.

SMEs will be much more willing to participate when meetings and interactions are kept as short as feasible. This may require some creative scheduling. For example, a 6-hour meeting may need to be separated into several shorter meetings. Having a large pool of SMEs available will make things easier by decreasing the individual burden on each SME. On the other hand, there is an advantage to having a core group of SMEs who participate throughout the process. An SME who understands the purpose and process of simulation development can be an invaluable partner throughout the project.

Table 9.2 SME sources for simulation development

SME source	Best suited to provide information about	Limitations/considerations
Incumbents	Critical incidents, daily tasks, competencies needed for success, work environment	SMEs for this process should possess good communication skills and be considered high performing in the job
Supervisors	Competencies needed for success, differences between high and low performers, suggestions for simulation content	Supervisors usually provide less detailed descriptions of daily tasks and work activities
Higher level managers and executives	Limitations/shortcomings of current workforce, future changes that may impact the role	This group may be less knowledgeable about day-to-day challenges and work activities
Human resource/training personnel	Technical documentation about the job, training materials	This group may be less knowledgeable about day-to-day challenges and work activities

9.2.5 General Scoring Considerations

The approach to scoring a service-role simulation is inevitably influenced by the fact that service roles tend to be high-volume positions. In many organizations, service employees will account for the largest sector of employment in the organization. Because these types of simulations are typically used early in the selection process as a prescreener, testing volume will likely be high.

Simulation scoring approaches can range from fully automated systems requiring no human interaction or judgment, to work samples that must be individually scored, to a hybrid approach that combines automation with human intervention. For service simulations, fully automated scoring is usually the best choice because of following reasons:

- Results can be available immediately upon completion, thus facilitating shorter candidate cycle times.
- Test administrator training is minimized, and is usually limited to obtaining and interpreting results.
- Scoring consistency is maximized, which can be especially important for organizations with multiple locations.

The simplest automated scoring strategy involves a straightforward scoring of user responses to predefined choices (e.g., multiple-choice questions, extended list questions). However, participant responses do not necessarily need to be predefined to be scored automatically. As discussed in Chap. 5, it is possible to automatically score more complex user-generate responses using sophisticated, decision-based software programs that interpret participants' free responses. Programs utilizing automated voice capture and automated essay scoring are available. In general, utilizing these types of software programs requires a high degree of technical sophistication. Vendor programs should be approached with caution, as there is a great deal of variability in the sophistication and reliability of commercially available software.

Another option is to combine automated scoring with human scoring. Under the right circumstances, this can be a very successful approach. For example, ETC's SkilTrak^{TM} simulation utilizes user-scored evaluations of participant's free-response e-mails. Users are trained and guided to evaluate participant responses using a pre-defined checklist. Based on the checklist responses from the evaluators, the system generates an automated score. This approach, although requiring a small amount of time from the evaluator, maximizes reliability and consistency while still allowing for an element of human judgment.

The issue of timing the simulation should be considered carefully and should be based on the time pressures that exist in the target job. Some service jobs (e.g., grocery store cashier) must be performed under significant time pressures. The simulation design for this type of job should include an element of time pressure. Speed, accuracy, and completion percentages can all be factored into the scoring formula. On the other hand, some service jobs (e.g., retail sales associate) do not have substantial time pressures. For this type of job, the simulation should have a generous time limit that will allow most users to complete the entire program before time expires.

9.2.6 Avoiding Common Pitfalls

With more than two decades of experience in simulation design and development, there have been many lessons learned along the way. The following is a brief summary of common pitfalls and suggestions for strategies to avoid them. Simulations are time-consuming and expensive to produce. Be prepared that some end users may expect an exact duplicate of their current working environment and operating systems. This is not desirable or necessary in order to create a realistic, effective simulation that will remain relevant for as long as possible. Information gained during the job analysis phase can guide the content development to ensure that the simulation targets tasks and competencies that will remain relevant now and well into the future. When designing the look of the simulation, it is better to strive for a more generic look and avoid "dating" the simulation by including anything too trendy. Attempt to avoid trendy expressions and colloquialisms in the dialogue. If live actors are used, strive for a classic look by avoiding clothing and hairstyles that may appear outdated in a few years. This is an area in which avatars or virtual characters have an advantage. Older avatars can often be replaced with newer ones without the need to reengineer the entire simulation.

Equal Employment Opportunity Commission (EEOC) regulations caution test developers to avoid evaluating participants on content that can be easily trained or is readily accessible to employees on the job (EEOC 1978). If successful completion of the simulation requires knowledge of a policy or procedure, participants should have ready access to this information throughout the assessment period.

This can be accomplished by including a location for reference materials within the simulation, giving participants time to review the material prior to beginning the assessment period, and instructing participants to rely only on this information during the simulation.

When designing the introduction to the simulation, there are competing considerations. On one hand, the instructions given to participants need to be clear and comprehensive. If participants are confused, then the process will not yield valid results. Participants, especially those in a high-stakes selection context, may be nervous and somewhat distracted. They may not fully comprehend the instructions immediately. It is very helpful to include a practice component to help participants feel comfortable with the interface prior to beginning the actual assessment period. If a practice component is not included, then the first task or interaction will often function as a practice, which is not the ideal situation if discriminant validity is important.

On the other hand, recruiters and human resources personnel want the shortest possible assessment time. In today's fast-paced world, even candidates will become frustrated and lose interest if the assessment period seems excessive. When viewed as a test budget, more time spent in an introduction means less time available for the actual assessment. There is no single best answer to this dilemma, just a caution to consider these competing concerns when designing the simulation.

9.3 Proving Effectiveness

Simulations, unlike many other types of assessments, have a high degree of "face validity." The connection between the assessment of one's ability to perform the work and the actual performance of the work is often readily apparent. A typing/keyboarding test is a classic example of a test with high face validity. If a participant can accurately input 80 words per minute in a typing assessment, then it seems logical that on-the-job performance of this task would be similar.

Simulations for service roles, however, are far more complex than a simple keyboarding test. Successful performance on the simulation requires a combination of skills and abilities. For instance, most service roles require at least some interpersonal skills, including competencies such as building rapport, persuasion, collaboration, and teamwork. Even if the simulation has the look and feel of the target job, it is still one step removed from actual, live interactions with real people. It does not automatically follow that the competencies necessary to be successful in the simulation are the same competencies that contribute to actual on-the-job success in the service role.

Therefore, for both legal and practical reasons, it is necessary to prove that the simulation is effective for its stated purpose. Proving effectiveness can take many forms, including formal documentation of content validity, statistical methods demonstrating criterion validity, and practical approaches that focus on the return on investment

(ROI) for organizations. When practical, a combination of all of these methods should be employed to document the simulation's effectiveness.

All validity efforts should be documented in a technical validation report. The EEOC's *Uniform Guidelines on Employee Selection Procedures* (1978) provides detailed standards for conducting and documenting validity efforts for selection procedures. The following comments are intended to highlight areas that are particularly important in the context of service-role simulations, but are not intended to be an exhaustive description of all the necessary components and considerations in a validation effort.

9.3.1 Content Validity

Simulations lend themselves well to a content validity approach. The primary purpose of content validation is to ensure that the content of the test (e.g., situations, test questions, response items) accurately represents the content of the target job. The Society for Industrial and Organizational Psychology (SIOP) states "When the selection procedure is designed specifically as a sample of important elements in the work domain, the validation study should provide evidence that the selection procedure samples the important work behaviors, activities, and/or worker KSAOs necessary for performance on-the-job . . . this provides the rationale for the generalization of the results from the validation study to prediction of work behaviors" (SIOP 2003, p. 21).

Since service jobs tend to be heavily interaction-based, ensuring content validity starts with a solid job analysis that includes collecting critical incidents of interactions that are typical in the target job. To the extent that these incidents are represented in the simulation, this approach has the added bonus of providing a realistic preview of the job. While many critical incidents will involve dealing with angry or upset customers, it is important to simulate a variety of types of interactions, including customers who present a neutral or positive affect. Candidates who experience a simulation filled with undesirable interactions may be left with an unrealistically negative impression of the target job.

Working closely with SMEs throughout the development of the simulation will contribute to the content validity of the end product. SMEs should be involved in many steps of the development process and their involvement should be documented in the technical report. If possible, try to collect descriptive information on all SMEs who participate, including job title, tenure, and demographic details. Including these details in the technical report will demonstrate the diversity of SMEs who participated and enhance the legal defensibility of the development process.

A key component of the content validity process is to establish a linkage between the work activities that are simulated and the competencies that are being measured. This can involve a simple matrix that links each of the work activities with the competencies required to perform each activity. This can be done by the test developers, by SMEs, or by both groups.

Survey methodology can be used to obtain and document SME feedback on the technical accuracy, realism, and appropriateness of the simulation content. There are existing survey approaches available to capture SME ratings (e.g., Lawshe 1975). Alternatively, surveys can be custom-designed to fit the project needs. Survey results should be summarized and included in the technical report. These results may include a summary of quantitative ratings of content validity as well as SME comments and feedback.

9.3.2 Criterion Validity

In addition to establishing content validity, the legal defensibility of any assessment will be greatly strengthened by conducting a criterion validity study. Criterion studies can be time-consuming and logistically complicated to complete. However, the advantages of a criterion validity study include the following:

- The validation group can be used to establish the norms and performance benchmarks for the assessment.
- Given large enough sample sizes, the data can be examined for any potential subgroup bias and scoring can be adjusted if necessary.
- Participant reactions (e.g., perceptions on engagement, difficulty level, and realism) can be collected in conjunction with the validation study. These reactions can be useful when promoting the assessment, and can also provide insights to test developers.
- When the simulation is deployed with a larger group of participants, occasionally technical problems that did not appear during the beta testing phase will surface, and these can be addressed prior to its final release.

In a predictive criterion validation study, job candidates experience the simulation, get hired (with or without knowledge of simulation results), get trained, and finally become capable of independent job performance. At this point, job performance is measured and compared with the preemployment simulation results. A disadvantage of the predictive approach is that this process can take months, even considering the relatively high-volume, fast-paced nature of service roles. In the service industry, the time to proficiency from the initial hire to independent performance is usually about 3 months. The advantage of a predictive study is that criterion data can be collected during multiple points in this process. Simulation scores can be compared with a number of data sources, including other assessment results that the candidate experienced during the selection process, performance during training and "nesting" periods, and time to achieve proficiency.

It is usually easier to conduct a concurrent criterion validation study, rather than a predictive study. In a concurrent study, incumbents in the target job complete the simulation, while job performance metrics are simultaneously gathered. The use of a concurrent approach allows researchers to gather and compare test data and performance information from a large sample in a short time frame. The disadvantage of this approach is that simulation scores will be subjected to a restriction of range due to the use of an incumbent population that has already been screened, trained, and coached to perform the target position. There also may be restriction of range in the criterion, due to the attrition of low-performing employees and the reluctance of supervisors to provide low ratings for current employees.

Regardless of whether the study is concurrent or predictive, there are a number of job performance metrics that lend themselves to measurement. Performance measurements can include objective and/or subjective metrics. Objective metrics that are frequently useful for criterion measures in service roles include training scores, average handle time (AHT), call quality monitoring scores, customer satisfaction scores, and quota achievement metrics. The most useful objective metrics will exhibit a range of values and be under the control of the individual employee.

For example, if all contact center agents in the study have availability scores (i.e., percent of time available to take calls) of 96 % or greater, then this metric will probably not be useful in the validation study. Similarly, if an objective metric is largely determined by factors that are outside of the individual employee's control, it is unlikely to show a relationship to the assessment. Although almost all collection agents are tracked on the amount of money collected, this metric is often influenced by so many outside factors that its usefulness in a validation study is limited. For the same reason, team-based metrics that depend on the overall performance of a team are problematic.

Subjective metrics usually consist of performance evaluations provided by supervisors of study participants. Although it is possible to use an existing evaluation, it is often best to develop an instrument specifically for the validation study. The survey should focus specifically on the competencies being measured by the simulation. To increase the reliability of ratings, survey items and response options should be behaviorally based. Finally, supervisors should not have knowledge of the participants' simulation results and should be assured that the ratings will remain confidential and only be used for research purposes.

Raw, uncorrected correlation coefficients from well-designed and well-executed studies are almost always an underestimate of the operational validity of the assessment (Sackett et al. 2008). In order to obtain a more accurate estimate, validity coefficients should be statistically corrected when appropriate. For validation studies that utilize incumbent populations, it is appropriate to correct for restriction of range. For validation studies that rely on a subjective criterion measure, it is also appropriate to correct for unreliability of the criterion. Because the performance measure is typically completed by only one supervisor, this correction estimates the correlation that would be obtained between the test and a composite performance rating of two or more raters.

The U.S. Department of Labor (2000) has provided some guidelines regarding the anticipated usefulness of assessments, given their criterion validity coefficients. Although there are many factors that influence the usefulness of an assessment, these guidelines serve as a useful starting point when evaluating validity coefficients (Table 9.3).

Table 9.3 General guidelines for interpreting validity coefficients

Validity coefficient value	Interpretation
Above 0.35	Very beneficial
0.21–0.35	Likely to be useful
0.11–0.20	Depends on circumstances
Below 0.11	Unlikely to be useful

Table 9.4 Validity results from ETC validation studies

Simulation	Target job	Validity coefficient	*N*
Collection Agent Simulation for Hiring	Collection agent	0.48**	323
Teller Vision	Bank teller	0.47**	370
Virtual Customer—Service	Customer service provider	0.47**	785
Virtual Customer—Sales	Retail sales associate	0.41**	359

**Results significant at $p < 0.01$

Table 9.4 provides some actual validity coefficients from several ETC assessments that target service-related roles. All of these results were obtained from concurrent validation studies involving incumbent service providers, and all validity coefficients were corrected as described previously. These relatively high validity coefficients demonstrate that simulations can be used very effectively in the service sector.

9.3.3 Setting a Minimum Passing Score

A criterion validation study can provide data to assist in the establishment of a minimum passing score for the assessment. Establishing the minimum passing score for an assessment is a complex task that combines both science and experience. In understanding how a minimum passing score is determined, it is first helpful to review the key objectives of preemployment assessments. An effective preemployment tool has three characteristics. It will be valid and show utility in predicting job success, reduce the chance of a legal challenge, and be practical and cost-effective for the organization.

An appropriate minimum passing score will allow the assessment to best meet the above three objectives. If a criterion validity study has shown that the assessment is proven to predict on-the-job performance, then the minimum passing score can be set at a level that best suits business needs and minimizes adverse impact.

9.3.4 Minimizing Adverse Impact

The potential for adverse impact against protected classes is frequently an important consideration when deploying simulations for service roles. There are several strategies during development that can minimize adverse impact. Ensure that the SME group includes a diverse group of individuals who are representative of the potential candidate pool. SMEs who are high-performing incumbents in the target job may be particularly helpful in this regard. For example, SMEs may be able to flag phrases, words, or situations that have the potential for cultural bias.

When collecting demographic information as part of the simulation, it is better to place these questions at the end of the simulation rather than the beginning, to minimize the possibility of stereotype threat. Stereotype threat has been identified as the pressure that a person can feel when he or she is at risk of confirming, or being seen to confirm, a negative stereotype about his or her group (Steele and Davies 2003). Although researchers have found mixed results in support of this theory (Sackett 2003), placing the demographic questions at the end is a prudent and easy step to address this potential concern.

Finally, monitor the reading level of all text-based material that is part of the simulation. This includes policies and procedures, question stems and responses, and background information (e.g., product descriptions, policies, and procedures). The inclusion of text-based information should be directly supported by the job analysis results. While some service roles (e.g., technical help desk associate) rely heavily on text-based information, other service roles (e.g., theme park attendant) may not. As a general rule, the reading level of text-based material should not exceed an 8th grade reading level.

9.3.5 Determining Return on Investment

Return on Investment (ROI), in the context of assessments, is more accurately a form of utility analysis. It is a practical approach to proving the financial impact of using a simulation, and ultimately to proving the simulation's economic value to an organization. Similar to any other financial investment, an ROI calculation considers the costs of using the simulation as compared with the economic gains that can be expected. The goal is to capture the worth to the organization (in dollars) of having higher performing employees who are less likely to be absent, less likely to turn over, and who are more productive.

ROI approaches start with the assumption that the simulation has validity, and many ROI formulas incorporate test statistics (e.g., criterion validity coefficients, descriptive statistics) into the calculations. For this reason, ROI computations are often conducted after the initial validation research has established the simulation's effectiveness from a statistical standpoint.

When compared with validity research, ROI approaches can be accessible to a wider audience. Many people are unfamiliar with advanced statistical concepts and remain unimpressed with validity evidence, even when the evidence is extremely compelling. Using an ROI approach can translate the statistical evidence into economic terms that present a convincing picture of the simulation's effectiveness.

The cost side of the ROI equation should include the ongoing usage costs of administering the assessment. This can include usage fees, recruiter or administrator time, and equipment/computer costs if the simulation is administered on-site. If the simulation has been developed in-house, these development costs should also be included and perhaps prorated over the expected lifespan of the simulation.

The benefit side of the ROI equation is much less straightforward to calculate. There are vast differences in approaches, assumptions, and formulas that have been used to estimate the "return" component of ROI calculations. However, it can be useful to divide the ROI benefits of simulations into calculations of *cost savings* and calculations of *economic opportunities*.

Calculating costs savings is usually the easier of the two calculations. In the high turnover environment associated with many service positions, a preemployment simulation can select better employees and thereby lower turnover. This can lead to significant cost savings by reducing the following:

- Hiring costs, including advertising, recruiting, testing, and interviewing.
- Development costs, including training and onboarding.
- Dissatisfaction costs associated with customer dissatisfaction or low employee morale. For example, empty seats in a call center can lead to longer customer hold times as well as stressed out employees who are expected to pick up the slack.

Estimating the economic opportunities from increased employee performance is often more difficult than estimating the cost savings from lower employee turnover. There are many formulas available, ranging from simple formulas with few variables to those that are more complex. Nevertheless, a preemployment simulation that identifies better employees can lead to higher productivity (e.g., shorter wait times, more customers served), higher customer satisfaction, and potentially higher revenue for the organization.

Schmidt and Hunter (1998) have developed a utility formula that is relatively easy to apply. The formula takes into account three factors: the difference between the new (presumably more valid) selection method and the old selection method; the difference in job performance between a poor employee and a good employee; and the average score on the selection method of those hired as compared with the general candidate pool. We used this formula to estimate ROI when a large multinational rental car organization began using our simulation, Virtual Customer—Sales, to select rental counter personnel. The organization knew that higher performing employees brought more revenue into the organization through the sale of optional rental services. Using conservative estimates for some of the variables, the formula yielded a projected sales increase of $1,280/employee/year; using more generous estimates for some of the variables, yielded a projected sales increase of $2,560/employee/year. When compared with the costs of implementing the simulation, it was immediately clear that the economic benefits far outweighed the costs.

9.3.6 Organizational Communication Strategies to Convey Results

After being immersed in an effort to demonstrate the effectiveness of a simulation, it is tempting to assume that the positive results will be obvious to everyone else. It can be disheartening when the results are met with confusion, skepticism, or worse, disinterest. This can be avoided by tailoring the results to different audiences. Although those with an Industrial Organizational (I-O) or science-related background will care about validity coefficients, others may greet these results with a big yawn.

It is useful to remember that different stakeholders in an organization care about different things. With regard to high-volume service roles, human resource personnel and recruiters will be most interested in how the simulation will help them fill vacancies quickly and easily. They will care about ease of use, throughput of candidates, and screen out rates. They will be most impressed with results that show them how the simulation will efficiently provide them with higher quality candidates.

Managers want employees who will be successful in training and get up to speed more quickly. They also want reliable employees who will not quit or be absent, which can be a huge challenge for managers of service workers. They will respond to results related to training proficiency, absenteeism, turnover, and satisfaction with hire. Executives, who are focusing on the larger picture, want a better workforce that will give them a competitive edge over their competition. This group will respond to ROI calculations that focus on higher productivity, customer satisfaction, and revenue.

When done correctly, communications that include pictures, graphs, and charts can help convey the message clearly and efficiently. Graphic displays should be streamlined. The message should be immediately apparent at first glance. Charts should be limited to only the most pertinent information that can be easily comprehended. In short, a well-crafted graphic will convey a compelling story of cause and effect using numbers and images. More information on this topic can be found in the landmark series of books by Edward Tufte (Tufte and Graves-Morris 1983; Tufte and Weise Moeller 1997).

An expectancy chart is one example of a graphic image that can quickly convey the relationship between simulation performance and on-the-job performance in a nonstatistical way. An expectancy chart shows a participant's predicted performance on some criterion variable, given the participant's score on the simulation. The underlying relationship is based on a regression equation that has been generated using data from a validation study. However, there is no statistical knowledge needed to understand the results of the chart.

In the following example, ETC conducted a blind predictive study with an organization that was using our Call Center Simulation™ assessment. The new employee assessment results were plotted against call quality monitoring scores that were collected during their first few months on the job (see Fig. 9.3). This chart provides two important messages. First, it is apparent that regardless of simulation score, all employees will likely show improvement in call quality during their first several

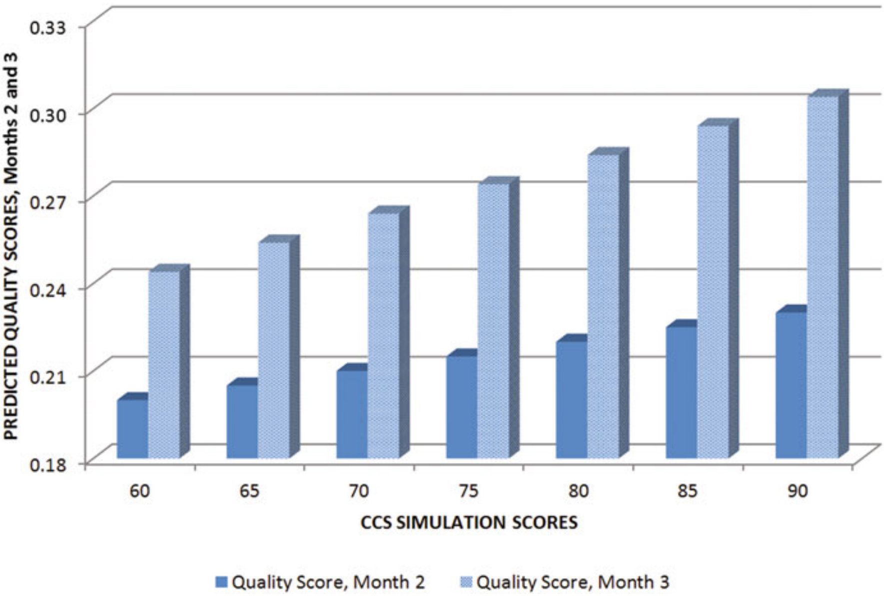


Fig. 9.3 Expectancy chart for Call Center Simulation

months on the job. But more importantly, the improvement will be accelerated for those who achieve higher simulation scores. The chart shows that employees who score higher on this preemployment simulation will get up to speed faster, resulting in positive benefits for the organization.

9.4 Implementation Strategies

9.4.1 Simulation Placement in a Multiple Hurdle Approach

To maximize the effectiveness of simulations as hiring tools for service roles, the manner in which they are used is of great importance. Building a great simulation does not guarantee success if it is implemented incorrectly. There are a variety of implementation strategies to consider, but we will highlight the most common strategies that represent best-in-class practice. These strategies have been honed over more than 20 years of experience with hundreds of hiring organizations, both large and small. A quality simulation will always provide some value regardless of implementation strategy, and there is no one best way. However, we will describe some guiding principles that should be considered. In most cases, a service simulation will be used in conjunction with other assessments as part of an overall selection process. Many times, these selection processes use a multiple-hurdle approach. A multiple-hurdle approach requires candidates to perform at a certain threshold on each assessment to progress to the next step in the process. This is an attractive approach when the hiring organization needs to screen large numbers of candidates to fill the available job openings. Many service roles such as call center agents, retail clerks, bank tellers, and food and beverage staff are high-volume positions. In these examples, the hiring organization will have many openings to fill and a large number of candidates to choose from. Given these conditions, a multiple-hurdle approach is both a practical and fair approach. These multiple-hurdle approaches typically utilize a series of different types of assessments, each designed to measure a unique part of the criteria that define success in the target role. The most effective multiple-hurdle process employs assessments that each contribute to the overall predictive power of the process as a whole.

Each component of the process should be targeted to different aspects of on-the-job success in the target role. Since simulations are often competency-based assessments, they are often combined with résumé and application reviews, prescreening assessments containing minimum qualifications questions, values assessments, personality inventories, and screening interviews. Each of these types of tools provides unique value. The subsequent decisions involve 1) how to best configure these tools to maximize the effectiveness of the process and 2) where should the service simulation be placed?

We will make the assumption that each of the tools in the selection process is valid and predicts some facet of job performance. We will also assume that the service simulation has demonstrated criterion-related validity, based on one or more validation studies. Over the years, we have discovered that simulations for service roles are consistently among the most accurate predictors available (see Table 9.4). In addition, the validity evidence of simulations was supported in a meta-analysis conducted and presented by Frank Landy at a recent SIOP conference (Landy 2008).

These high validity coefficients suggest that in best practice, service simulations should be placed early in the overall selection process. The fact that simulations are computer scored, objective, and unbiased suggest that early placement in the selection process is most beneficial. We suggest a service simulation be placed directly after a résumé/application review and a minimum qualifications prescreening assessment. If a value or personality inventory is part of the process, the simulation can be placed directly before or after such an inventory.

Consider issues like accuracy, cost, and completion time when deciding on the placement of the simulation. A general guideline is to leverage the most valid and unbiased instruments earlier in the process before the candidate comes into direct contact with a recruiter or hiring manager from the organization. Interviews, whether conducted by phone or in person, are still subject to bias and rater error. Our experience suggests that once someone in the hiring organization forms an opinion about the quality of a candidate, there is no turning back. Using these more subjective methods early in the process will ultimately undermine the effectiveness of the simulation. If the simulation is used after candidate contact, recruiters and hiring managers will expect the simulation to confirm their perceptions, however inaccurate those perceptions may be. You will be better served to let the highly valid and objective simulation work for you to screen out unqualified candidates. Using the simulation to prequalify candidates before they speak with recruiters and hiring managers will minimize the likelihood of making bad hires.

This approach is also highly efficient. Much time is still wasted interviewing unqualified candidates. If you can provide interviewers with only prequalified candidates, then you will save significant time and money. This will enable your interviewers to focus their energy on only the most qualified candidates. This will also encourage more in-depth interviewing and better results. Initially, interviewers may be reluctant to embrace this approach, fearing that they are losing control of the hiring process. Be patient and in the long run your interviewers will thank you.

Not all hiring organizations elect to use a multiple-hurdle approach. Some organizations prefer to administer an entire test battery, including a simulation, to all candidates regardless of individual scores on each instrument. Organizations may adopt this approach if their candidate flow is less intense. They may also want their recruiters to maintain more control over the hiring process. In these cases, a service simulation will provide key information that, when combined with other tools in the process, provides a robust picture of each candidate's likelihood for success. When using this approach, hiring organizations should consider all of the information gathered in the process and make hiring decisions based on the totality of the information. To ensure consistency and legal defensibility, the hiring organization should still develop standard written guidelines on using the assessment information to make hiring decisions. This approach still leverages the data gained by service simulations, while providing the hiring organization with more flexibility than a standard multiple hurdle approach.

One of the advantages of service simulations is the wealth of data generated based on candidate performance. A well-designed service simulation should provide an overall score that can be used to make employment decisions. This overall score is based on a total of the cumulative responses across the entire simulation. This score is the most reliable score because it is based on the largest number of items. However, well-designed simulations should provide more than just an overall score. In addition, a service simulation should provide feedback on the various service dimensions that are measured across the simulation. Our service simulations provide information about candidate performance at the service competency level. Most service-related roles can be defined by a set of competencies that includes some form of building rapport, discovering needs, problem solving, and ensuring customer satisfaction. Since these subscores are based on fewer items, they should primarily be used to diagnose the developmental priorities for those candidates who are ultimately hired.

9.4.2 Access to Simulation Results

An important consideration is determining who should have access to simulation scores and when access should be granted. Service simulations are commonly used to make hiring decisions for entry-level service employees. These hiring decisions are usually made jointly by recruiters and hiring managers. In high-volume hiring situations, recruiters are tasked with sourcing candidates, screening them, and then passing qualified candidates to the hiring managers to make final decisions. In our 20 plus years of experience designing and implementing service simulations, we have seen organizations employ many different approaches to disseminating and managing simulation scores. Most of these approaches fall into two categories: limited access to results and open access to results.

Organizations that elect to use a limited access approach do this by carefully managing and controlling the release of simulation scores. They limit the scores that recruiters and managers are allowed to see and when they can see them. Some organizations elect to only share pass or fail results with both recruiters and hiring managers. This approach has some advantages. If candidate flow is heavy, a simple pass or fail result is all that is needed to determine if job candidates should proceed past the service simulation to later parts of the selection process. Limiting access to only pass/fail results also ensures that there is no bias in later parts of the selection process based on simulation scores. Some hiring organizations fear that the release of full simulation scores that usually range between 0 and 100, as well as service dimension subscores, will bias both recruiters and hiring managers in subsequent stages of the hiring process.

Other organizations prefer allowing more open access to service simulation results. An open access approach provides recruiters and hiring managers with total access to service simulation scores. These organizations allow recruiters and managers to review overall numerical scores (e.g., between 0 and 100) and the subscores for each service competency measured in the simulation. This approach enables the hiring organization to use all of the available information to help make accurate hiring decisions that align with the organization's specific needs and hiring objectives. Some of these open access organizations also advocate using the simulation results to drive the content of subsequent interviews in the selection process. These organizations look at the simulation results to identify potential red flags related to competency performance and then target these areas in the interview process. Of course, other hiring organizations use a hybrid approach that may provide total access to recruiters but only pass/fail information to hiring managers (or vice versa). In almost all cases, job candidates do not receive any overall simulation scores.

If hired, some organizations will share the subscore results with candidates as well as hiring managers to identify developmental opportunities and customize the initial training and onboarding process to increase efficiency. Unfortunately, most organizations do not leverage the service simulation data posthire. This is a common mistake. Service simulations can provide a great deal of developmental feedback that can be very useful in the early developmental cycle of an employee. Simulation scores can identify developmental priorities based on simulation results and provide developmental suggestions for improvement. More organizations would benefit from leveraging prehire simulation data for posthire developmental purposes.

In reality, there is no one best way to use and manage the access and reporting of service simulation results. Both limited and open access approaches have their respective advantages. The key is to design service simulations that provide flexible reporting. Flexible reporting will enable the hiring organization to support a variety of approaches. Creating reports that display different types of scores designed for different audiences is beneficial. Also, consider the timing of when results will be shared and with whom. Posthire reports designed for new hires or hiring managers should have a developmental tone that is focused on performance improvement, rather than an emphasis on specific scores used in the hiring decision. In summary, the best practice is to provide options. Options will allow the organization to adopt and implement the strategy that best meets the needs of its business.

9.4.3 Ongoing Consultation and Support to Optimize Results

As simulation developers, our job does not stop once a hiring organization implements a simulation. Our job really begins after implementation. In order to ensure that hiring organizations receive the maximum value from our service simulations, we need to provide ongoing consultation and support to optimize results. There are several areas that require ongoing attention. The first is passing rates of the candidate pool. Hiring organizations look to service simulations to help them screen out unqualified candidates who have limited opportunities for success on the job. We use normative data gathered during the validation and subsequent benchmarking processes to help hiring organizations set minimum passing scores.

Most hiring organizations expect a service simulation to screen out somewhere between 20 and 35 % of the candidate pool. Initial minimum passing scores can be set to accommodate this goal as long as the validation data supports this approach. The validation data should reflect that candidates who score below the minimum passing score are unlikely to achieve satisfactory performance ratings on the job. Of course, EEOC considerations should also be taken into account when setting minimum passing scores so that the potential for adverse impact is minimized. Once set, the performance of the service simulation should be monitored to ensure it is performing as expected. We recommend periodically review passing rates at least once every 6 months. Look for meaningful trends in the data that may impact the ongoing use of the simulation, including

- Are the passing rates similar to the initial expectations established at the time of implementation?
- Are there regional differences in passing rates?
- Are there score differences based on recruiting source, prescreening method, or administration approach?

Service simulations should also include an optional demographics section requesting candidates to identify their race, gender, and age categories. This information will enable you to monitor passing rates by demographic subgroup to monitor the potential for adverse impact. Monitoring simulation scores over time to discover trends provides added value to hiring organizations and will extend the usefulness and impact of the simulation.

Since service simulations are generally among the most accurate predictors of job performance when compared with more traditional screening methods like applications, résumé reviews, and prescreening interviews, an organization may want to use the same simulation across multiple position titles and business units. In these circumstances, you should use a consistent process to determine the suitability of a simulation for additional job titles. If a localized validity study is not feasible, you can use a transfer of validity approach. At ETC, we employ a job match process to ensure the work activities performed by incumbents in the target position match the work activities of the population in the job analysis and original validation study. A short job match survey can be used for this process. The job match survey should include items related to the work activities performed and competencies required for success on the job. A suitable degree of similarity should be established before recommending a service simulation as a hiring tool for any target job that was not included in the original scope of the job analysis and simulation development efforts. Selecting a sound implementation strategy and monitoring simulation performance will ensure success when using service simulations for the hiring of service roles.

In recent years, it has been exciting to see the growing acceptance of computerized simulation technology for hiring. As we began our early work in this area, it was hard to imagine a book of this kind with so many respected authors representing various facets of simulation technology. We are pleased to be a part of it. We hope this chapter has provided some insights based on our collective experience in developing and implementing service-related simulations.

References

Dymond, J. (2012). The decline of US manufacturing jobs and living standards. BBC News. http://www.bbc.co.uk/news/world-us-canada-18992241. Accessed 10 Feb 2013.

Employment Technologies Corporation (ETC). (2012). *Virtual customer service: Development and validation report*. Winter Park: Author.

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38295–38309.

Hill, R. B., & Petty, G. C. (1995). A new look at selected employability skills: A factor analysis of the occupational work ethic. *Journal of Vocational Education Research, 20*(4), 59–73.

Landy, F. (2008). Meta-analysis of electronic customer service assessments using call center simulations. Paper presented at 23rd Annual SIOP Conference, San Francisco.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575.

Murray, S. (2009). Service sector is job-growth engine. The Wall Street Journal. http://online.wsj.com/article/SB126049773496186833.html. Accessed 10 Feb 2013.

Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance, 16*(3), 295–309.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215–227.

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*(2), 262–274.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). College Park: Author.
Steele, C. M., & Davies, P. G. (2003). Stereotype threat and employment testing: A commentary. *Human Performance, 16*(3), 311–326.
Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information* (Vol. 31). Cheshire: Graphics Press.
Tufte, E. R., & Weise Moeller, E. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire: Graphics Press.
U.S. Department of Labor. Employment & Training Administration. (2000). Testing and assessment: An employer's guide to good practices. http://www.onetcenter.org/dl_files/empTestAsse.pdf. Accessed 10 Feb 2013.
Wanous, J. P. (1989). Installing a realistic job preview: Ten tough choices. *Personnel Psychology, 42*(1), 117–134.

Chapter 10
Managerial Simulations

Sara L. Gutierrez and Jolene M. Meyer

Managers are pivotal in their role as the link between individual contributors and higher-level leaders within an organization. Overseeing the majority of the workforce, they are responsible for ensuring effective operations, high performance, and a positive work climate. Finding the talent to fill these key roles cannot be left to chance. It is not enough to simply be strong technically; a manager needs to be able to inspire others and accomplish work through them.

Compared to the role of individual contributor, the role of manager presents unique challenges. There is acceleration in demands for prioritizing tasks and making quick decisions. Managers cannot lose sight of the organizational goals and strategies as they focus on day-to-day operations. High-quality work that got them noticed in the first place must be maintained while demonstrating nimbleness in dealing with new issues. They need to be people-focused: setting aside time for developing their employees, monitoring their performance, and managing disputes as they arise. To be truly effective in leading others, they also need to be an advocate for their team while interfacing with the rest of the organization. Based on numerous job analyses conducted by Corporate Executive Board Company (CEB), the following is a list of competencies and experiences that underlie successful performance in managerial roles:

- Coaching and Development: effectively engaging with direct reports
- Prioritization: identifying and completing critical tasks in the most efficient manner
- Decision Making: identifying issues and drawing conclusions based on these issues
- Monitoring: assessing the effectiveness of a team's performance
- Personality Characteristics, such as Achievement, Influence, Independence, Reliability, and Confidence and Optimism: setting and accomplishing challenging

S. L. Gutierrez (✉) · J. M. Meyer
CEB, Chantilly, VA, USA
e-mail: sara.gutierrez@shl.com

J. M. Meyer
Minneapolis, MN, USA
e-mail: jolene.meyer@shl.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_10,

goals, taking actions and making decisions without direct support from others, directing others in situations that require leadership and leading others toward a group objective, taking responsibility for own actions, and having belief in one's ability to get the job done
- Problem Solving: efficiently and effectively using numbers and analytical reasoning to solve problems

Given the breadth of behaviors and experiences listed earlier, well-designed selection tools are necessary to identify the best person for a managerial position. However, finding the tools that can measure the more complex aspects of the position can be a significant challenge for an organization. Traditional selection tools such as multiple-choice cognitive tests, biodata scales, and personality assessments can be administered to measure problem solving, professional potential, and personality characteristics. However, these types of assessments are not ideal when attempting to measure complex competencies such as coaching and development, prioritization, or monitoring. For example, the extent to which a candidate can effectively prioritize information from numerous sources, make day-to-day strategic decisions, coach and develop employees, and monitor employees' performance are difficult work behaviors to assess utilizing simple multiple-choice or Likert-type response formats. As such, even in combination, traditional selection tools will not thoroughly cover the domain of key competencies needed for success in managerial positions.

These difficult-to-measure competencies have traditionally been assessed in the context of an assessment center (i.e., sets of instruments and techniques used as part of managerial selection to judge the likelihood of a candidate's success as a manager; Cascio 1998). Assessment centers require candidates to role play or perform in-basket exercises to simulate interpersonal or problem-solving tasks that are frequently performed by managers (Bray and Howard 1983). Although utilizing a managerial-focused assessment center may be the optimal choice when attempting to understand a candidate's potential for success in a role that requires tasks such as monitoring employee performance and developing one's employees, organizations do not often find them to be cost effective. Traditional assessment centers rely on human raters to judge the extent to which a candidate's record of behaviors is related to the characteristics required for the job. Many organizations have a high number of manager positions spread across a variety of locations; sending candidates through an assessment center would require at least one assessor to be available at each location in order to conduct the assessment, a requirement that is quite costly.

An alternative approach, both more cost effective and time efficient, is to supplement traditional selection tools with state-of-the-art simulations designed to assess the more complex core competencies required for these important positions. Simulations offer assessments that appear more job-relevant to candidates than traditional selection tools, such as text-based assessments. The increased face validity is rooted in the fact that simulation item formats allow the information to be presented in a manner more similar to how the information would be experienced in daily life, which is a more authentic presentation of information to the candidate. For example, rather than reading a situational judgment item that describes an employee who is upset about a co-worker stealing his sales opportunity, a simulation item format will use a video to convey not only the message but also visual cues such as body language, and verbal cues such as voice inflection. As the candidate can better envision him or herself within the situation, improved measurement of these competencies may be realized. In other words, simulation content may provide a more direct presentation of information to candidates, leading to more accurate measurement and therefore stronger reliability and validity of scores (Zenisky and Sireci 2002).

Simulations may also hold the potential for better measurement of these constructs due to the positive reactions that are likely to be elicited from candidates. Measurement practitioners have shown that negative reactions to a test, such as viewing the content as irrelevant, can lead to poor motivation to do well (Macan et al. 1994). When candidates have a decreased motivation to exert effort, their test scores will not accurately reflect their ability on that construct. Since simulations have been shown to elicit positive reactions by candidates (Richman-Hirsch et al. 2000; Shotland et al. 1998), administering a simulation may result in increased motivation of candidates and lead to scores that more accurately reflect candidates' true ability on the constructs of interest.

Finally, although assessment centers are job relevant and can offer much in the way of both content and criterion-related validity evidence (Winfred et al. 2006; Hermelin et al. 2007), these methods tend to be time-consuming and costly. They are often ineffective when the organization is global and manager positions are spread across a number of locations. Technological advances have allowed employers to bring much of this assessment content into an online format in which technology is leveraged to assess candidates via interactive, media-rich simulations that simulate a "day-in-the-life" of a manager while measuring job-relevant traits and abilities. The use of simulated assessment content contains immense potential for the automation and objective measurement of the core competencies needed to succeed in managerial roles that are difficult to assess otherwise. Utilizing simulations allows for the assessment of more candidates less expensively, with fewer resources, and does not require that candidates travel to a particular location to complete the assessment.

Although great resource-related benefits can be realized using simulated assessments, these assessments must also be held to the same psychometric standards as traditional selection tools—they must be reliable and valid predictors of performance. In order to develop high-quality simulation content for managerial roles that is likely to be valid and reliable, an understanding of the work behaviors and competencies that are necessary to perform the job well must first be achieved. The purpose of the current chapter is to provide a detailed framework of the development work that was conducted to successfully design and implement two managerial simulations, (1) In-box Simulation, and (2) Coaching Simulation, that have been shown to be predictive of success in managerial roles.

10.1 An Overview of the Managerial Simulations

10.1.1 The Inbox Simulation

The Inbox Simulation consists of two sections. The first section measures prioritization, while the second measures decision making and monitoring. The interface of this simulation is designed to closely mirror the day-to-day managerial experience. In the assessment, candidates assume the role of a leader and are provided with background information and tools that are needed to complete the assessment, including a job description and organizational chart. Candidates are asked to prioritize demands from across the organization and their team, identify critical tasks, and ensure that those tasks are completed. Information is presented to candidates in the form of email, voicemail, phone calls, visitors, and calendar reminders. Emails are presented with text, the phone calls and voicemails through audio files, and the visitors to the office appear on screen in video format. Situations encountered in the Inbox Simulation include determining what resources employees need to do their jobs, responding to urgent questions from higher-level management, doing "just in time" problem solving, and working with managers of other departments or organizations. All questions are multiple-choice in nature. Examples of this interface can be seen in Figs. 10.1 and 10.2.

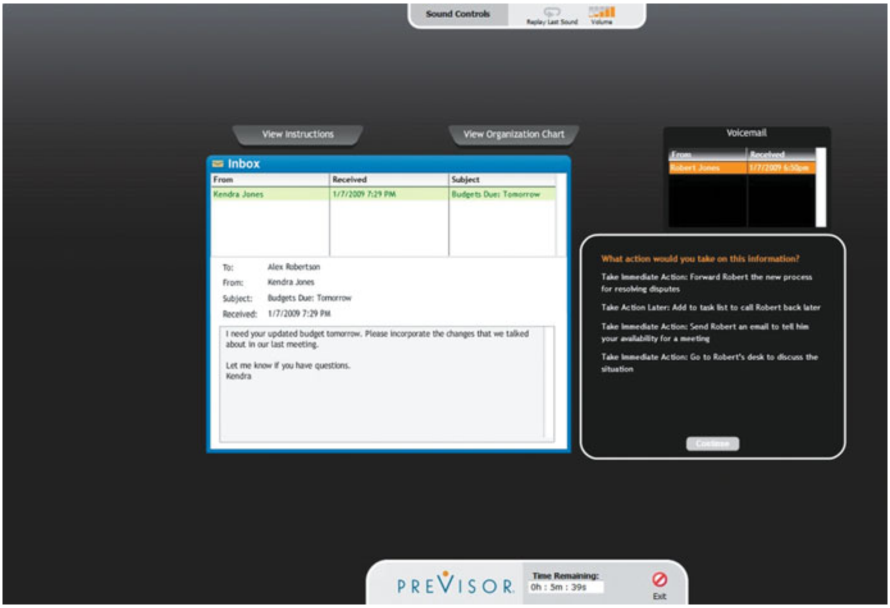


Fig. 10.1 Inbox simulation interface—displaying email and voicemail stimuli

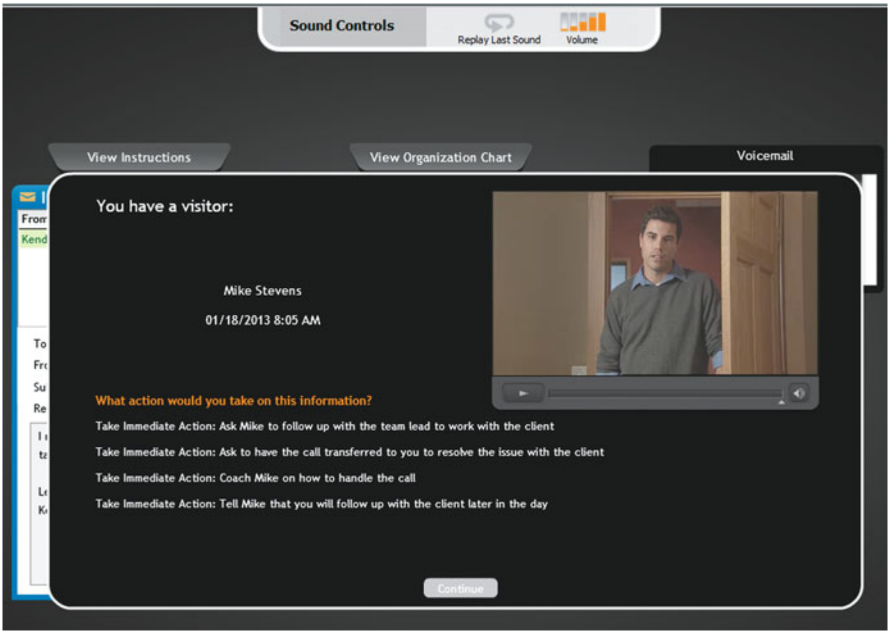


Fig. 10.2 Inbox simulation interface—displaying visitor to manager's office

10.1.2 The Coaching Simulation

The Coaching Simulation consists of various scenarios designed to measure coaching effectiveness. These scenarios contain fictitious subordinates, and candidates are measured on how they approach these conversations. Candidates are provided with opportunities to show how they would manage employees, display supportive leadership behaviors such as listening, probing, encouraging and empowering, and be directive and assertive when necessary. Scores on the Coaching Simulation are designed to predict the candidate's likelihood of engaging in such behaviors on the job. The simulation design is similar to a situational judgment test. Candidates are presented with video-based scenarios along with alternative courses of action. They are then asked to select which response options are most and least effective given the situation. An example of this interface can be seen in Fig. 10.3.

The following section is devoted to providing an in-depth review of the development efforts involved in designing and implementing the simulations described earlier in the chapter.

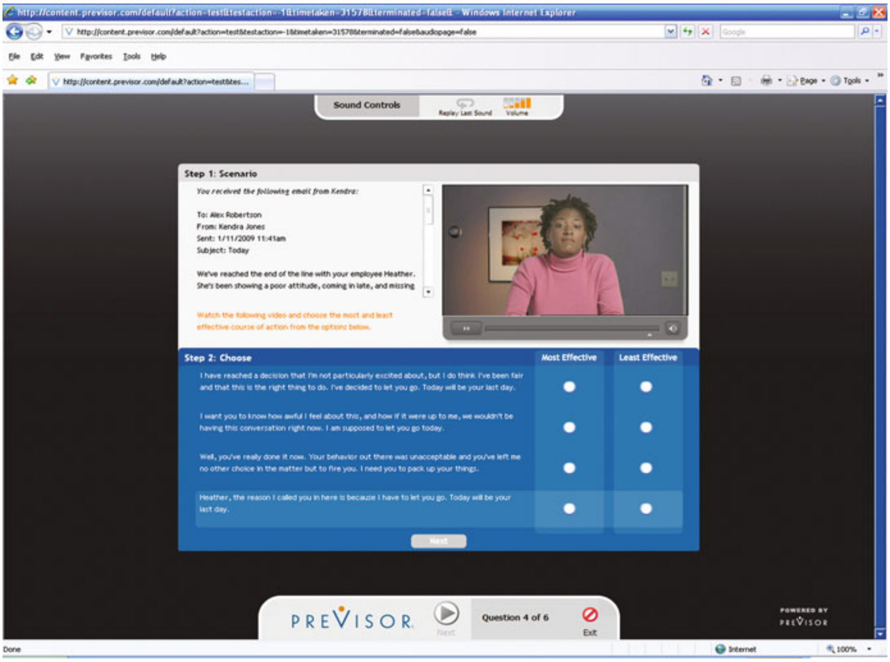


Fig. 10.3 Coaching simulation interface

10.2 Simulation Development and Implementation

10.2.1 Job Analysis

Development of the Inbox Simulation and the Coaching Simulation began with a series of job analyses involving managerial roles across multiple organizations within a variety of industries, including Telecommunications, Retail, Financial Services, Insurance, Healthcare, and Medical Services. As the simulations were designed to be applied universally to most industries, it was important to have a combination of industries represented in this phase of development. The goal of the job analysis was to gather information about the present and future job requirements of manager positions in order to develop a set of highly predictive and face valid simulation content. The approach to the job analyses involved four activities not only to comprehensively study the managerial roles, but also to focus specifically on the development of the simulations: (1) job observations of incumbents, (2) focus groups, (3) interviews with incumbents, their managers, and executive leadership, and (4) job analyses questionnaires. The multiple methods of gathering input were utilized to obtain perspective and feedback from a variety of levels. These four activities are important parts of any job analysis and test development process, but even more so for simulation development. As one of the main goals of utilizing simulations is a more realistic presentation of information, test developers had to truly understand the context in which these managers were working. The test development team included five Industrial–Organizational Psychologists who held advanced degrees, all with extensive experience in the area of selection and assessment. Through job observations, onsite focus groups, and interviews, test developers were able to obtain important situational, environmental, and background information that helped shape the basis for the simulations.

10.2.1.1 Job Observations

The goals of the job observations were to (1) directly observe the actions performed by the managers, and (2) gather critical incidents about the types of issues presented to these leaders on a day-to-day basis. The critical incidents gathered in this portion of the job analysis process became the topic matter for the managerial simulation content.

The job observation process included a walk-through of a typical day on the job, followed by the job analysts observing the interactions between the managers and his/her team members as the manager performed basic job activities. Additionally, at the end of the observation period, the analyst asked the manager a set of questions regarding the tasks that had been observed. Following each observation, job analysts rated the importance of a variety of competencies in terms of their job relevance to the work behaviors performed by the manager. In addition, to aid in the development of the Inbox Simulation, job observations were used to determine the breakdown of time spent on tasks/activities performed by the incumbent. The job observations were typically completed at the same time as the interview, so additional questions were asked to gather critical incidents for the Inbox Simulation. Critical incidents focused on the following topics:

- Typical distractions that occur during the course of the day that require switching attention from one task to another
- Methods used to prioritize work when multiple things/people are competing for attention
- Methods and frequency of monitoring work being completed by subordinates
- Tasks that are typically delegated to direct reports

Furthermore, certain topics were the focus for the development of the Coaching Simulation. Questions for the Coaching Simulation targeted the following topics:

- Frequency of coaching poor-performing employees
- Common performance issues that require coaching, such as compensation, customer complaints, and interpersonal disagreements
- Most difficult problems to coach
- Important things to do when coaching an employee and common mistakes an ineffective or new manager might make

Finally, for both managerial simulations, participants were asked to rate incidents that were important, common, difficult, and challenging to handle. Then, they were asked to provide examples of how one might effectively or ineffectively respond to such incidents.

10.2.1.2 Interviews

The goal of the interview portion of the job analysis was to obtain information about the overall mission and operating procedures within the participating organizations. In addition to the interviews conducted with the managers, interviews were conducted with individuals in director, trainer, and HR manager roles. Interview protocols were structured to elicit the following information:

- Overall mission of the organization and the role that front line managers and supervisors play in executing on the mission of the organization
- Description of the performance ratings and metrics collected
- Current selection process
- Knowledge, skills, abilities, and other characteristics needed to be a successful manager
- Anticipated changes to the business over the next 5 years
- Characteristics that separate top managers from those who are just 'average' performers

Some of this interview protocol was not directly relevant to the design of the simulation content per se (i.e., mission of the organization, anticipated change to the business). However, it was important for the test developers to gather such information in order to ensure that the simulations were supported by other job and organizational-relevant content that would be later incorporated in the final assessment solution.

10.2.1.3 Focus Groups

Focus groups were conducted with incumbents in manager roles and consisted of two phases: (1) continue to gather critical incident information, and (2) obtain feedback on potential test questions for the new managerial simulations. The goal, then, of conducting focus groups was to gather additional material for scenarios that could be used within the two new simulations. Specifically, critical incidents were gathered relating to work behaviors that were known to be difficult to assess with traditional selection tools, such as coaching and developing employees, monitoring, decision making, and prioritizing on the job. Additionally, the development team had created potential test questions prior to the start of the focus group process, based on the information gathered in the job observations and interviews. Focus group participants were asked to evaluate these potential test questions in terms of job relevance and appropriateness/plausibility of response options, and were also asked to generate ideas for alternative response options.

10.2.1.4 Job Analysis Questionnaires

Upon the completion of the job observations, focus groups, and interviews, the resulting information was compiled and reviewed. Analysts developed a comprehensive list of work behaviors that represented what typical managers do in the course of a work day. From this list, an extensive conceptual model was created that contained the competencies and work behaviors that led to successful performance of the requirements of the roles and was the basis for the job analysis questionnaire. As the model was designed to be encompassing all competencies required in these roles, this questionnaire addressed work behaviors outside of the behaviors that the two managerial simulations were targeted to measure.

The questionnaire consisted of job task and activity statements relevant to entry-level leadership roles, such as: "*Strive to achieve departmental objectives despite challenges.*" These statements were organized into 26 work behavior dimensions, nine of which were to be mapped to content in the two new simulations: Handling Conflict, Building Relationships, Developing Employees, Analyzing Problems, Identifying and Considering Alternatives, Building Teams, Planning and Monitoring Progress, Prioritizing Work Demands, and Monitoring Against Goals. Job task and activity statements were placed on a five-point scale that ranged from 1 (Unimportant) to 5 (Critically Important). High-performing job incumbents from across participating organizations completed the questionnaire. Results indicated that the conceptual model consisting of 26 work behavior dimensions aligned with the role of manager. This confirmed that the extensive four-prong approach to the job analysis of the manager's position led to the identification and confirmation of the work behavior dimensions required for success in these positions.

The goal throughout the development of both simulations was to be able to assess a portion of competencies traditionally measured in an assessment center context using internet-based simulations. The phases of the job analysis described earlier were imperative to gain insight into the true "day-to-day" context in which managers functioned and served to aid the development of realistic and appropriate item content for the simulations. The critical incidents collected during the job analysis process directly served as the stimuli for the assessment. The process of transforming the critical incidents gathered during the job analysis into functioning simulation content is described in the next section.

10.2.2 Content Development

The development of the two simulations is described next.

10.2.2.1 Inbox Simulation

Stimulus Material Development

The Inbox Simulation was divided into two sections. The first section of the simulation was designed to focus on candidates' ability to prioritize information, and is related to the Prioritizing competency identified through the job analysis. The emails, phone calls, voicemails, and visitor stimuli were written to vary in terms of their urgency of response. Focus groups conducted during the job analysis helped to refine and revise the stimuli. The second section of the Inbox Simulation was designed to consist of "mini in-basket exercises." These exercises were designed to measure candidates' judgment with respect to monitoring employees' performance and decision making in the context of receiving information from multiple sources. The items are related to the Monitoring and Decision Making competencies identified through the job analysis. Related sets of content targeting specific themes were created, such as planning/administration, monitoring employee output, and responding to customer issues.

Item and Scoring Development

The format for all items in the Inbox Simulation is similar to situational judgment items. Candidates are presented with information through the various stimulus materials discussed earlier, and are asked to make judgments about the information. A range of response options were written for an item that would discriminate those with high levels of the competency from those with low levels of the competency.

Subject matter experts (SMEs) provided feedback on these response options by rating the relevance of the items to the dimension it was intended to measure, providing input on the quality/plausibility of the response options, and indicating the extent to which each response option would relate to successful outcomes. The SME rating/feedback process contributed to the creation of a priori scoring that was empirically tested during the criterion-related validation phase of development.

10.2.2.2 Coaching Simulation

Stimulus Material Development

Similar to the Inbox Simulation, content development of the Coaching Simulation began with the job analysis. The development goals included utilizing the information obtained through job analysis to design a highly predictive and face valid simulation by understanding the extent to which coaching was part of the job, and the nature and types of coaching that were performed.

Critical incidents gathered during the focus groups included an employee confiding in their manager about personal issues at home impacting their performance, providing critical feedback to an employee about poor job performance, having an employee make excuses and push back when critical feedback is given, resolving conflict between two employees, and having a junior-level employee complete his portion of the project and go home without telling the project lead. These critical incidents were used for developing the scenarios in the simulations. Each scenario designed for the simulation consisted of a hybrid of the incidents that were documented during job analysis. Site visits occurring later in the job analysis process were used to refine the scenarios and to develop new response options.

Item and Scoring Development

Four professional coaches were asked to rate the response options on their effectiveness. The coach read the background to each situation, viewed the video, and then rated the effectiveness of the responses. During the validation, data were collected that were later used to empirically investigate the usefulness of the score key provided by the effectiveness ratings of SME raters during development. The empirical scoring method was determined to be superior at predicting successful outcomes than the SME effectiveness ratings, and was thus implemented in the final version of the assessment.

10.2.3 Final Validation Versions of the Assessments

10.2.3.1 Creation of Scripts and Test Stimuli

For both managerial simulations, the stimulus materials, test questions, scoring, and dimensions were organized and developed into test scripts. These scripts served as the basis for the technical development of the simulations as well as the scripts for professional acting and voice talent. The test questions were created by engaging with a professional audio/video production company to record the audio and video portions of the assessments. Following the audio/video production, files and test scripts were handed off to a programmer for creation of the assessments.

Once programming was complete, the final beta versions of the managerial simulations existed on an online testing platform and were ready to be included in the validation work conducted by partnering organizations. The online testing platform allowed the managerial simulations to be deployed anywhere with an Internet connection, at any time of day.

10.2.3.2 Concurrent Validation

The Inbox Simulation and the Coaching Simulation were validated within a criterion-related study design with data collected by a consortium of six organizations. Incumbents working in the role of manager within the consortium organizations were asked to complete a set of assessment content that included the new simulation content along with problem-solving (cognitive ability) measures, a writing ability test, personality scales, and biodata scales. For each incumbent, the direct manager was asked to complete a job performance rating form. The form included a variety of items designed to obtain as much information about the incumbent's job performance as possible. For example, the form included 27 performance dimension ratings—items that aligned with the job analysis work behavior dimensions such as Decision Making and Managing Talent. Seven items referring to the incumbent's global/overall job performance were also included (e.g., incumbent's long term career potential, the manager's desire to rehire this incumbent, general effectiveness as a manager). From these items, several performance composites were formed to serve as the primary criteria for the validation. The data provided by the consortium were used to examine the statistical relationships between scores on the new simulation content and ratings of job performance.

Prior to investigating the test-level validities of the new simulation content, item-level analyses were performed for both the Inbox Simulation and the Coaching Simulation. Final forms of the assessments were created based on extensive review of the items and how they functioned both at the item- and the test-level. Based on the incumbent data provided during the study, scores were tabulated for each incumbent on these final forms of the assessment. These scores were then used for the examination of the validity of the simulation content.

Given the multiple samples of data provided by the consortium, the validation results were examined in the context of meta-analysis. Meta-analysis in this case allowed for the synthesis of information from multiple studies that used the same content in a variety of settings to judge the overall value of the new content. The results of the meta-analysis indicated that scores on both simulations were substantially related to various types of supervisor ratings of job performance, and provided support for the use of the new managerial simulations to assess some of the core competencies required to perform well in front line manager and supervisory roles.

10.2.4 Implementation

Following the development and validation of the media-rich assessments, a number of clients have implemented these assessments into their organizations' hiring systems. A case study for a large retailer that has implemented the two managerial simulations into their selection process for Store Manager and Assistant Store Manager roles is discussed further.

10.2.4.1 Case Study: Retail Store Managers and Assistant Store Managers

Understanding the Role

For this particular organization, meeting customer needs and expectations was the most important goal of the business. It was up to the Store Managers and Assistant

Store Managers to model this organizational core value in their everyday actions, leading their team members to do the same. Individuals in this role were expected to be ambassadors for the brand. These roles were critical to the organizations' success as Store Managers were not simply team leaders or sales leaders, but were business leaders who were held responsible for the revenue and profitability of their branch.

The work behaviors expected in these positions actively encompassed many of the competencies outlined earlier in this chapter. Of specific interest was the Store Managers' ability to coach and develop the branch sales team. Through an internal survey, coaching and development was identified as a weak spot within the organization. Individuals in these roles needed to be adept at monitoring and assessing the effectiveness of a team's performance. Additionally, they needed to be able to prioritize competing tasks, identify issues, and draw conclusions in order to address such issues.

Understanding the Organizational Goals

Prior to implementation of the managerial simulation content, the hiring process involved a resume review, an interview, background check, and finally a reference check. This process was resource intensive, requiring multiple human touch points for every candidate applying for the role. As the Store Manager role gained visibility within the organization and more accountability for reaching revenue targets was placed on those in the role, the organization strived to improve the quality and standardization of their current hiring process. The organization desired to increase the objective data used to inform the selection process, relying less on subjective decisions. In particular, they wanted to be able to assess one's ability to coach and develop subordinates in a less subjective manner.

Due to the decentralized nature of the organization, resource constraints (both budget and staff) limited the feasibility of the use of assessment centers to inform the selection process. Although paper and pencil assessment may have been an economical solution for the decentralized company, paper and pencil tests would not have been able to effectively cover the range of competencies required for success in these very important roles. Given the requirements of the role within the organization, along with the organizational goals to improve upon the current selection system/process, implementing the two managerial simulations offered an opportunity to assess core competencies and work behaviors that might not be assessed otherwise.

Implementation

CEB recommends the use of a comprehensive assessment program in order to measure the knowledge, skill, ability, or other characteristics of an individual. We feel that organizations should take a 'whole person' approach when implementing an assessment program, utilizing a variety of test types and content to measure the set of competencies relevant to a job. This approach allows for a better understanding of an individual's characteristics, and therefore will increase the likelihood of hiring candidates who will perform well on the job. In order to effectively implement a 'whole person' approach, an organization must take the time to design an effective assessment battery (otherwise referred to as a "solution") for that particular job role.

Based on Federal guidelines and professional standards, we approach the design and implementation of assessment solutions for clients in the following manner:

1. Perform a job analysis to understand and document the job requirements: this step allows for the identification of the primary work activities, competencies, knowledge, skills, and abilities required for successful performance on the job within a specific organization.
2. Assemble a tailored assessment solution: select a set of well-developed assessments that measure the core work behavior dimensions and competencies identified by the job analysis.
3. Collect validation evidence showing the ability of the assessment scores to predict performance in the job role.
4. Implement a final version of the tailored solution: utilize the validation evidence to determine the most predictive set of content and calculate an overall score, based on the underlying competency scores, which can be used for easy decision making.

The described process was followed by the focal organization of this case study to design a relevant assessment solution for their roles. This assessment solution was then implemented into the hiring process for the Store Manager and Assistant Store Manager roles and included the following flow: a high-level resume screen, part one of an assessment solution (this section does not contain the simulation content for security purposes as it is administered unproctored in the location of the candidate's choice), an onsite interview and part two of the assessment solution (this is proctored, and includes the simulations), and finally a background check. The goal of the implementation was to put the assessment content as early in the process as possible—acting to screen candidates out prior to bringing them onsite for an interview. Implementing these technology-based assessments has reduced the amount of human judgment in the hiring process by effectively providing objective data points regarding key competencies. The ability to coach or monitor employees was previously determined through the interview process, whereas, assessment scores now exist to inform the selection decisions that are made. Additionally, the organizational goal to streamline the recruiting process was also met.

The effect of implementing such technology-laden assessment content did not increase the number of technological problems or user complaints that this organization typically encountered, beyond what was normally seen for traditional assessment content. When issues do arise, the problem is most often due to the logistical constraints under which the proctored tests must be delivered, and not with the managerial simulations themselves. Within the organization, the managerial simulations continue to receive positive reactions from business leaders due to not only their effectiveness in predicting successful candidates for these extremely important roles, but also the increased face validity of the assessment due to the media-rich simulations.

10.3 Conclusion

Simulations can meet an important need when selecting the best individual for a management role; without them, important and hard-to-measure competencies may not be effectively assessed. This chapter reviewed the development of two managerial simulations that can effectively replace two traditional assessment center exercises—the direct report role-play and the in-basket—with multimedia-based versions that can be delivered via the Internet. These simulations are most often supplemented with other types of online content such as personality, cognitive ability, and biodata assessments, or with in-person assessment exercises that are not currently available online. The future of the assessment of candidates for managerial roles may look very much like a traditional assessment center, but one that is delivered and scored completely online. Today, face-to-face interviews and other exercises such as the leaderless group discussion and presentation exercise, are still very often delivered in person. The promise of simulation technology is that it will offer a choice to an organization, to either hire managers with in-person assessment centers or to opt for a completely online version that is just as effective, more efficient, and less expensive than the traditional in-person assessment center. As a field, work in the area of simulation development is far from complete. Assessment professionals must look towards the future and answer the question, "What is next for this type of simulation?"

As technology advances, additional consideration should be explored and leveraged during the design and development of simulation assessment. For example, to the best extent possible, the ease of use should be optimized. This may include changes to the question type, the interface, and how the test is deployed (proctored vs. unproctored). With advanced technological capabilities, future versions of managerial simulations may include the use of customization to enhance the face validity and representativeness of the simulation to the organization employing it. Avatars could be used instead of live actors to increase the ability for customization. By changing the avatar and color scheme behind the simulation interface, a simulation could be made to have the 'look and feel' of the specific organization for which the candidate is taking the test. Additionally, the use of dynamic/random administration of items to increase the test security of the assessment could be considered.

References

Bray, D. W., & Howard, A. (1983). Personality and the assessment center method. In C. D. Spielberger, & J. N. Butcher (Eds.), *Advances in personality assessment* (Vol. 3, pp. 1–34). NJ: Erlbaum.

Cascio, W. F. (1998). *Applied psychology in human resource management* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centers for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15,* 405–411.
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.
Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85,* 880–887.
Shotland, A., Alliger, G. M., & Sales, T. (1998). Face validity in the context of personnel selection: A multimedia approach. *International Journal of Selection and Assessment, 6,* 124–130.
Winfred Jr., W. A., Day, E. A., McNelly, T. L., & Edens, P. S. (2006). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125–154.
Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15,* 337–362.

Chapter 11
Tracing Cognition with Assessment Center Simulations: Using Technology to See in the Dark

Brett W. Guidry, Deborah E. Rupp and Martin Lanik

11.1 Introduction

Both employees and organizations alike face challenges in the current, post-recession business environment. In December 2007, Asia, Europe, and North America entered a major economic recession (National Bureau of Economic Research 2008). For the U.S., it was the largest economic downturn since the great depression of the 1930s. The sheer magnitude of the economic shift imposed changes in organizational strategies, structure, and operations that many leaders were simply not equipped to handle. For instance, in July 2011, Borders Group, the second-largest brick-and-mortar bookstore chain in the U.S., announced that it would close its remaining 399 stores and lay off 10,700 employees (Spector and Trachtenberg 2011). What went wrong with the company that once set the trend for book superstores with coffee bars and sophisticated sales and inventory tracking systems? For one thing, it failed to adjust to twenty-first-century online shopping trends and growing customer demands for e-readers—demands answered by the competition (e.g., iPad, Kindle, Nook).

Stories of similar recent corporate failures worldwide are plentiful, including the bankruptcy of Hungary's largest airline, Malev (Simon 2012), the slow deflation of once dominant technology firms like RIM, Dell, and Yahoo (Evans 2012; Miller and Kopytoff 2011; Worthen 2012), and even the wrecking of a Carnival cruise ship under the command of a careless captain (McGregor 2012). Such events have led many scholars and practitioners to borrow the term VUCA (volatile, uncertain,

B. W. Guidry (✉) · D. E. Rupp
Purdue University, West Lafayette, IN, USA
e-mail: bguidry@psych.purdue.edu

D. E. Rupp
e-mail: ruppd@purdue.edu

M. Lanik
Global Assessor Pool, Denver, CO, USA
e-mail: martin.lanik@globalassessorpool.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_11, © Springer Science+Business Media New York 2013

complex, and ambiguous) from the U.S. Army War College to describe the post-recession economy. As today's business leaders encounter unanticipated and complex situations, the decisions they make can result in a competitive advantage, as well as an extreme loss for the firm (Corporate Leadership Council 2010).

Although not all business failure is the fault of incompetent leadership, it can be argued that the *decisions* made by managers, executives, and even influences outside of management roles essentially serve as the catalyst for most organizational outcomes. Decision making in organizations is more important than ever before, and thus our ability to optimally assess decision making is critical for effective selection, promotion, training diagnosis, and employee development. Recognizing this fact, this chapter focuses on proposing novel ways to get "under the hood" and leverage technology to measure previously unobservable decision strategies. Herein we argue that an evaluation of candidates' concealed thought processes can enhance myriad personnel processes. We further argue that these technologies might be best retrofit onto existing *assessment center* exercises.

The Assessment Center (AC) method, a compilation of simulation-based exercises used for assessment and development of personnel in organizations, has enjoyed a long, successful history in assessing the decision-making proficiency of managers and leaders (Thornton and Rupp 2006). However, the high costs associated with development and implementation of ACs have limited their application in organizations. In 2005, the Corporate Leadership Council reported that ACs could cost up to hundreds of thousands of dollars to design and then several thousand dollars per participant to operate. Largely motivated by cost reduction and better approximation of today's jobs, numerous technological innovations to the AC method have surfaced during the past decade. Indeed, Hughes et al. (2012) reported that 57 % of current ACs worldwide employ some form of technology, and 18–30 % of respondents expect to introduce technology features into their ACs in the next two years. Although many consulting firms and organizations have been moving AC simulations to online platforms, most key functions remain unchanged—common dimensions are observed, recorded, classified, and rated by trained assessors to predict future performance, diagnose training needs, and aid in developmental feedback. Technology features in ACs have mainly focused on aiding assessors in the observation and rating process (Reynolds and Rupp 2010; Rupp et al. 2008). In this chapter, we argue that technology has a far greater place in ACs than used currently, especially considering the increased stakes of decision making in organizations.

For simplicity, throughout this chapter, we focus specifically on in-basket/inbox exercises to illustrate new ways in which technology can be used to provide additional information about candidates' standing on important dimensions. We feel that this particular choice has the most far reaching generalizability, in that out of all the operational ACs worldwide, 42 % are estimated to contain at least one inbox exercise and 98 % are estimated to assess at least one dimension related to decision making (e.g., information seeking, problem solving, etc.; Hughes et al. 2012). Regardless, this illustration could easily be extended to other exercises.

Traditionally, in-basket exercises present candidates a set of hypothetical paper memos and correspondence that has accumulated over some period of time. Candidates sort through the information they receive, prioritize activities, make decisions, and respond to pressing issues. Over time, in-basket exercises have evolved, allowing candidates to access email, voicemail, and electronic calendars. Whereas this may be seen as one form of technological improvement over a purely paper-based exercise, our chapter is focused on technology enhancements that are beyond the mere computerization of administrative processes. That is, we focus on the additional behaviors that can be captured via technology after exercises have been computerized in this way. We then pull from the process tracing and decision-making literatures to provide the reader with a framework for organizing these currently untapped candidate behaviors. Finally, we discuss how these new data can be incorporated with traditional assessor ratings for gaining additional insights about candidates' proficiency in complex decision making.

11.2 Technology in Assessment Center Simulations: Past and Future

Although technology has been part of employment testing for several decades (Reynolds and Dickter 2010), its adoption by AC simulations is fairly recent. The delayed adoption of technology in ACs may be attributed to the relative complexity of simulating various processes and mechanisms employees use on the job while maintaining open-ended responding and interactivity with live assessors, i.e., features that distinguish the method from other types of assessment (International Task Force 2009). For instance, the market has seen a surge in *computerized situational judgment tests* (SJTs) that move the candidate into a game-like environment, present test questions as video clips or avatar animations, limit candidates' responses to a few preset options, and automatically score this multiple-choice test (e.g., simulearn.net). What distinguishes AC simulations from these sorts of assessments is ACs' open-ended format that elicits candidates' actual behaviors while responding to the stimuli. For example, in ACs, the candidates respond to stimuli by having a meeting, giving a presentation, preparing a report, or corresponding with others via phone, email, or memoranda, all in real time. In AC simulations, candidates are not constrained to select among behavioral response options, but rather *actually* engage in free will behavior. Candidates' responses are reviewed by trained assessors who classify behaviors into dimensions or competencies of interest and arrive at competency ratings.

Again, it is estimated that over half of ACs around the world use some sort of technology, the most common being to streamline assessor processes (Hughes et al. 2012; Reynolds and Rupp 2010). This includes recording behavioral observations directly into a computer system or database, automatically integrating scores across assessors, exercises, and dimensions, automating or semiautomating report generation, and scheduling. These upgrades are generally seen as having a moderately positive impact on the efficiency of AC operations. Other technological upgrades include computerizing aspects of exercises, as discussed above. Since many leadership roles require electronic communication, having candidates complete a computerized inbox equipped with email, calendars, and other professional and social media tools (as opposed to paper memos) is not only a natural and high-fidelity extension of the method, but it is also positively correlated with perceptions of ROI and participant reactions (Gibbons et al. 2013).

Despite such upgrades, a particular challenge still posed by the inbox exercise is that assessors can often only access the final product of candidates' decision making (i.e., content of sent emails), while insight into the actual *processes* candidates employ to arrive at their decisions remains within a "black box." That is, determining how candidates arrive at the set of decisions, communications, and other output at the end of the exercise is traditionally *inferred* by assessors based on limited information. In the paper version of the exercise assessors may have, for example, reviewed notes that led the candidate to his/her decisions, yet such notes are seldom captured in the electronic versions of such exercises. Of course, assessors have the option of conducting follow-up interviews with candidates to gain insight into their decision process. However, it is common practice to train assessors to be skeptical of any reflections on decision making (Thornton and Rupp 2006) because people are not necessarily accurate in their understanding of their own decision-making processes (e.g., Kirschenbaum 1992; Nisbett and Wilson 1977; Riedl et al. 2008). Thus, if technology could be used to capture actual decision behaviors in real time, this would likely enhance the assessment of dimensions related to problem solving and decision making.

Today's technology is capable of capturing the intermediate steps the candidates take to arrive at decisions and could enhance observation and rating processes by expanding the assessment beyond the focus on the outputs of the exercise. For example, data like keystrokes; mouse clicks; the content and timing of started, drafted, sent, or scrapped emails; timing and names of accessed files; created calendar appointments; or requested information provide an incredibly rich set of behavioral data and allow assessors to trace the path candidates took while moving through the exercise. As we discuss next, theories about judgment and decision making (JDM) from cognitive psychology provide frameworks for organizing this plethora of data, pointing to techniques that may allow for enhanced conclusions about candidates' decision-making acumen.

11.3 Detecting and Interpreting Decision-making Processes in Assessment Centers

Early judgment and decision-making research examined *what* decisions people made based on specific inputs rather than the process of *how* people made those decisions (e.g., Farr 1973). For example, researchers have examined and manipulated input order (e.g., Payne, 1976; 1982; Rowe 1967), input valence (favorable or unfavorable; e.g., Miller and Rowe 1967), and input variety (e.g., Rigney and DeBow 1967). Broadly referred to as *policy capturing*, this paradigm highlights what inputs individuals focus on (i.e., information cues) when arriving at a decision, as well as the differential weights placed on inputs when solving problems.

The policy capturing paradigm assumes a linear relationship between inputs and outputs in decision making and omits more nuanced components of the actual *process* of making decisions (Sherer et al. 1987). This method is similar to the way AC exercises like the inbox are currently scored. For example, candidates are presented with a number of decisions to make and certain information on which those decisions should be made. Working backward from the final decision, assessors determine performance on dimensions such as information seeking or problem solving in addition to more readily detectable dimensions such as (written) communication.

While early research on decision making tended to focus on relating inputs to outputs in a linear fashion, researchers ultimately began shifting toward studying the intricacies of the decision *process* itself (Koop and Johnson 2011). Some (Payne et al. 1978; Svenson 1979) argued that cognitive processes cannot be completely understood by focusing exclusively on the input–output relationship. Similarly, others (Onken et al. 1985; Patrick and James 2004; Thorngate 1980) pointed out that when people work in a complex environment (such as those simulated within an AC exercise), multiple strategies can lead to the same outcome, and these strategies are often difficult to observe. In terms of the inbox exercise, one must consider the behavioral data regarding the *path* to a decision in addition to *what* decision was ultimately reached to fully capture the decision-making proficiency of candidates.

The *process tracing* method, which is designed to uncover the actual process of decision making, can aid the AC practitioner in assessing dimensions such as problem solving and decision making. Instead of relying on manipulating inputs and comparing them with various outputs in the inbox exercise, more data can be collected during the decision-making process (Svenson 1979). Software can easily capture what emails the candidates open, in what order, how long they read them, what other information they access, and even mouse/cursor movements—and such software is publically available [e.g., Mouselab (Payne et al. 1993); ISLab (Cook and Swain 1993), ISCube (Tabatabai 1998); MouseTrace (Jasper and Shapiro 2002); DecisionTracer (Riedl et al. 2008)]. For example, a typical modern AC inbox exercise probably has a scenario in which a decision must be made about an employee. Although two candidates might arrive at the same decision, the efficiency or expertise of the decision can be judged based on this new, rich set of behavioral data.

In addition, such software can be used to measure response latencies—a very precise measure (often in milliseconds) of the actual time spent between actions such as typing or clicking—and can be combined with nonevasive physiological response measurements such as skin conductance (e.g., Electrodermal Activity Skin Conductance (EDA) Analysis Software, Mindware Technologies 2013). This process would allow assessors to measure the cognitive and emotional regulatory processes indicative of the cognitive load and stress facing the individual at certain points of the exercise. Specifically, skin conductivity measures—measures of minute changes in perspiration—provide information about physiological excitement of the individual. Skin conductance can be easily measured via a noninvasive sensor placed on the hand of the candidate and, although multifaceted, has been linked to cognitive processes like one's level of attention given to the task, habituation to any stimulus that is presented, and the extent of exerted cognitive effort (Figner and Murphy 2011). Within the judgment and decision-making literature, researchers frequently use skin conductivity measures to capture people's emotional responses in problem solving situations.

Finally, eye tracking hardware and software has also advanced to a level where it can feasibly be incorporated into AC programs to collect search and reaction time data during problem solving activities (e.g., Duchowski 2007). Almost every decision involves the acquisition of visual information (Russo 2011), and this is particularly true of an inbox. Several studies have revealed how eye fixations, eye movements, and even pupil dilations are related to information processing and subsequent decisions that the individual makes (Tao-Yi Wang 2011). Eye-tracking data can be integrated with the process tracing methods discussed above. Knowledge about what candidates are looking at during an exercise, combined with information about what documents they open and other keystrokes, could provide a more detailed account of their decisional processes. This objective decisional data can be combined with assessor observations (e.g., the content of sent emails) to produce a richer and possibly more accurate assessment of candidates. This new behavioral data could also provide an additional source of information to be used for the delivery of targeted, more effective developmental feedback (Klein 1997). In the following sections, we will first lay out the specific decisional processes that such technology might allow us to capture. Then, we will discuss how exactly such data might be used in an assessment center context.

11.4 Three Components of Decision Making Relevant to AC Simulations

We have identified three general components of decision making that could be captured via process tracing within AC exercises: (1) how candidates seek information, (2) how candidates process information, and (3) how candidates integrate information to ultimately come to a final decision. Although these processes are often not sequential or even independent, for simplicity, we review each process separately.

11.4.1 How Candidates Seek Information

When people encounter a problem to be solved, they first seek data to inform their decision process. For example, Edwards and Fasolo (2001) listed 19 steps to making a decision, the first three of which are to identify options, identify possible outcomes of each option, and identify criteria with which to evaluate the outcome. In common decision tasks, such as choosing a car, each step plays out very predictably. Buyers compare different models of different manufacturers on specific and well-defined attributes (e.g., price, fuel economy, safety) as well as more subjective attributes (e.g., ride, handling, styling). Budgets immediately eliminate some selection choices, and other options would be weighted according to buyers' specific needs or preferences (e.g., taking a hit on fuel economy because a larger vehicle is needed). While some of the better defined attributes are readily available from the manufacturer (e.g., EPA fuel economy estimates), it is trickier to make subjective judgments like handling. Fortunately, with a car, one can always test drive the top choices and gain first-hand experience rather than relying on the opinions of others.

Unlike car buyers, the decision criteria that candidates might apply to "solving" AC exercises (or that managers and executives may apply to complex organizational problems) are less clear, or at least not explicitly stated. To further complicate things, there are often no specific instructions for where to obtain additional information—and test drives are out of the question. For example, an inbox exercise might present a problem with an underperforming subordinate, including multiple emails regarding this same issue. It is up to the candidate to seek out all the relevant emails and any other necessary information, including employee policies, official documentation of past employee behavior, or even information about the people sending the emails in the first place (i.e., their level of expertise in the matter).

Once people are presented a problem, they generally adopt an immediate strategy as to how to approach it (Ford et al. 1989). This overall approach has direct effects on many subsequent behaviors, such as how one seeks new information to address the problem. For example, if a candidate sees the problem as an emergency, he or she will approach it differently than an average situation. The selected approach will vary based on candidates' previous experiences and how the problem is classified (Ford et al. 1989; Walsh 1995). Because the approach candidates select is difficult to capture directly, it is inferred from other behaviors associated with information acquisition. For example, to determine whether someone sees a situation as an emergency, previous behavioral observations of individuals in an emergency situation would need to be matched to a candidate's current approach. Failure to engage in such behaviors would mean that the candidate (a) did not recognize cues to an emergency situation, or (b) recognized the situation as an emergency, but did not respond appropriately. In an AC context, either conclusion is important for an overall prediction of job performance, but pinpointing the reasons for poor performance could significantly improve developmental feedback and goal setting for performance improvement.

Aside from emergency situations, studies have shown that people adopt general patterns as to how they seek information when confronted with a problem to be solved or a decision to be made (Ford et al. 1989). At the highest level, one could examine whether information seeking behavior occurs at all. For example, researchers have examined whether people rely on past experience when confronted with a new problem or whether they rely exclusively on the data at hand (Walsh 1995). In other words, is someone making a decision without actually looking further into the situation? Such lack of information seeking suggests that the candidate assumed the situation to be similar to a previous experience and therefore did not need new information. Alternatively, the candidate may have assigned a low priority to the decision and skipped information seeking that would lead to a more accurate decision. Due to the lower perceived priority, the candidate may be more comfortable making a "good enough" decision without expending the time or effort to collect additional information. Effectively, the absence of information seeking signals how candidates interpret and ultimately approach a particular problem, which could have implications for their on-the-job performance.

Conversely, the presence of information seeking behaviors yields itself to examination of the specific patterns of the behavior. For example, AC practitioners can evaluate whether or not candidates are systematic/targeted in their information seeking (Ford et al. 1989; Walsh 1995). Many studies have indicated that experts tend to employ more active, targeted searches, while novices tend to evaluate information in the order it is received (e.g., Kirschenbaum 1992). Similarly, experts ultimately use *less* information when making a decision, though this is likely due to the efficiencies that past experience provides. Determining how targeted or efficient a particular candidate's search pattern was has implications for judgment on his or her level of expertise in information seeking. Broad and erratic search patterns suggest that candidates (1) may have struggled with how to handle the situation or (2) viewed the decision as high priority and needed as much information as possible to be correct (Ford et al. 1989). Even before considering the specific reason, the general pattern of information seeking behavior can provide insight into candidates' cognitive processes that would otherwise be missing from their evaluation. Once combined with the exercise output (i.e., decisions actually made), as well as information about candidates' tenure in high-responsibility positions and past experience with situations common to the challenges simulated in the exercise, assessors can draw conclusions regarding the effectiveness of candidates' search strategies that add a distinct level of specificity to the overall evaluation.

Although much of the judgment and decision making literature focuses on the cognitive processes of experts vs. novices (Klein 1997), previous experience is not the only factor that affects how information searches unfold. For example, Jacoby et al. (1984), in their multitrial study of stock analysts, found that high performers (as judged by the output of the exercise) examined the same amount of information across trials, but low performers examined less information as trials progressed. In this study, search patterns were not affected by participants' level of expertise but instead by other factors like "running out of steam" or mistakenly becoming comfortable processing less information over time. Collecting objective processing data alongside traditional assessor ratings might enable an assessment team to tease apart experience-based decisional strategies (which do truly add efficiency to decision processes) from grit and perseverance on decisional tasks over time, particularly when the stakes of the outcomes become increasingly high. Such information would be very useful in both high-stakes selection contexts as well as for enriching developmental feedback for individuals in long-term, high-pressure work environments involving complex decision making.

Although information seeking is fundamental to both AC exercises and managerial jobs, it is merely the tip of the iceberg. That is, candidates engage in a multitude of cognitive processes during and after they seek information. As such, we next discuss two other components of the decision-making process: how candidates process information, and how they combine the information to ultimately make a decision.

11.4.2 How Candidates Process Information

Several researchers have pointed out the distinction between analytical decisions—involving information processed in a controlled, step-by-step fashion; and more automatic decisions—where information is processed passively and decisions reached more intuitively (e.g., Ferreira et al. 2006; Highhouse 1997). Although Ferreira et al. (2006) argued that no decision is actually "process pure," and many other researchers believe that most decisions blend analysis with intuition (Hammond 1996), each is worth considering separately. Analytic, controlled processes are deliberate and typically rely on a small number of "cues" or bits of information (Hammond et al. 1997). Similar to "bottom-up" processing (Walsh 1995), people essentially rely on available information to make any decision, including how to interpret or categorize the information being processed. In such cases, the situation is considered novel enough that previous experience (or heuristics) cannot be readily applied. Without the mental shortcuts afforded by heuristics, controlled, bottom-up approaches typically require more energy and time while also being naturally limiting on how much information can be utilized (i.e., fewer bits of information are used). While this type of intensive analysis often results in few errors, when errors do occur they tend to be significant. Conversely, intuitive decision-makers often utilize large amounts of information but have little cognitive awareness of the process. Consequently, analytic decision-makers tend to have lower confidence in their decisions than intuitive decision-makers.

Intuitive decision making primarily functions through the use of heuristics, which allow decision makers to unconsciously reduce the amount of information they process (Perrin et al. 2001). Instead of relying on specific pieces of information, heuristics enable decision makers the use of generalities or approximations. This process is similar to that of "top-down" information processing (Abelson and Black 1986), where a decision maker relies on previous knowledge while making sense of new information (Hayes-Roth 1977). For example, during car-buying decisions, one may simply remember that a prospect's fuel economy is "good" rather than remembering the exact miles per gallon. The judgment of whether a car's fuel efficiency is "good" would be based on previous knowledge about automotive fuel efficiencies. Such an approximation helps the buyer make quick decision without the burden of detailed data.

People use heuristics on an everyday basis to simplify the decision-making process, and these generally efficient heuristics often result in reasonable assumptions and accurate decisions (Thorngate 1980). However, overreliance on heuristics, or the use of heuristics in the wrong situations, can result in poor or incorrect decisions (Highhouse 1997; Sims and Gioia 1986). Using the car-buying example, outdated information about fuel efficiency might lead one to classify a car as "very good" when, based on current standards, efficiency is only average. Similarly, a faulty heuristic might cause one to classify *all* vehicle fuel economy as "good," washing out any variability in the category and effectively making it useless. While the preceding example can occur in cases of limited information, a similar breakdown in heuristics can occur when individuals have too much information. We can think of at least two instances within AC contexts where such an excess of information may come into play (while also serving as useful information for assessors). One is the attraction effect (Highhouse 1997; Huber et al. 1982), in which the presence of a seemingly irrelevant option affects perceptions of other options. For example, in a simulation exercise where a candidate must make personnel selection decisions, the decision between two relatively more qualified candidates may be influenced by the presence of a third "decoy" candidate, even though the decoy is clearly less qualified for the position.

Another information overload example that could be particularly relevant to an AC context is the dilution effect (Highhouse 1997; Nisbett et al. 1981; Zukier 1982), whereby clearly nonrelevant or nondiagnostic information about an individual affects one's impression of that individual. Within an AC exercise, this might involve having information about a problem employee that goes beyond work performance (e.g., hobbies) which in turn impacts how that employee is handled (vs. not having such information). Importantly, many of these errors in judgment are intensified when individuals know they can be held personally accountable for the decision (Simonson 1989; Tetlock and Boettger 1989). Such accountability is something organizations often strive for in the decisions of their managers and something that would be implied in an AC context. In other words, because candidates would be aware that their performance is being evaluated within an AC, there would be added motivation to seek out as much information as possible when making decisions.

In the AC context, these kinds of surplus information situations are commonplace as candidates often have access to more information than they need to make a decision. Ultimately, some candidates may fall into a trap and spend most of their time on tasks deemed unimportant. Candidates' pattern of information searching and processing, and information included in the final decisions can help identify the type of information processing strategy that was employed. In a dynamic exercise, a new problem requiring new information could be presented later in the exercise. Individuals with relatively lower proficiency on this component of problem solving would be less likely to seek new information to correctly address a problem than relatively more proficient candidates, as low performers may fail to recognize that the new problem requires new information. While candidates do not necessarily need an *entirely* accurate understanding of their environment, the ability to distinguish critical tasks and relevant information from irrelevant distractors is crucial to their success in the AC and ultimately on the job (Starbuck and Milliken 1988). By capturing information attended to, time devoted to information, reaction times, and other process tracing behaviors, assessors would be enabled to objectively map candidates search and application of such information.

11.4.3 How Candidates Integrate Information to Arrive at a Decision

The third major aspect of judgment and decision making that is relevant to AC exercises is how candidates integrate information to arrive at a decision. Within this literature, researchers have identified two major integration strategies (Ford et al. 1989). Candidates adhering to *compensatory* or *linear strategies* consider all the available (or acquired) information, assign rough weights, and use the information to make a decision. Given the vast amount of information being integrated, the compensatory method is very complex and cognitively demanding on the decision maker

(Einhorn and Hogarth 1981). Conversely, *noncompensatory integration strategies* are similar to a multiple hurdle approach, in which early information is considered prior to additional information searches or utilization of latter information (Ford et al. 1989). Noncompensatory strategies are often less cognitively demanding since information processing happens in a sequence; i.e., only one subset of information is acquired and processed at a time, and processing end points are clearly defined. For example, when a candidate attempts to decide whether or not to fire an employee for insubordination, he or she may base this decision on specific guidelines available during the AC (e.g., a conduct policy whereby three documented reports of insubordination in a year results in termination). The candidate may not consider any other information (e.g., performance records, other employee/manager opinions), even if they are made available within the context of the exercise.

In reality, people often use both compensatory and noncompensatory strategies when making decisions (Ford et al. 1989). However, the more complex a task, the more likely it is that individuals will use noncompensatory strategies. If a task is particularly complex, people are more inclined to simplify the situation by looking for an easy, all-or-nothing cutoff so they can move on to the next issue. Some researchers even suggested that as a task becomes less complex, individuals move from a noncompensatory to a compensatory strategy (e.g., Bettman and Park 1980). For example, when people are searching for a house, they may limit the initial number of options using a broad cutoff like price or location. Once other criteria have been met (e.g., size, number of bedrooms), they may employ a compensatory strategy to choose from a more manageable set of alternatives.

In the AC context, assessors can make judgments about the candidates' compensatory versus noncompensatory strategy based on the particular decisions reached and behavioral patterns in information searching. For example, when candidates work through an inbox exercise, they may engage in noncompensatory integration strategy at first to reduce the initial number of decisions (particularly the decisions deemed less important). Later, they may switch to compensatory integration strategy for other decisions. Assessors can capture this switch in integration strategy by following what documents or emails candidates opened or reviewed at any given time, and what actions they took after opening those emails. In our previous employee insubordination example, a candidate may go directly to the company handbook and move straight to the section regarding an employee code of conduct. Assessors would make the interpretation that the candidate is searching for a specific guideline on which to base a decision. If the candidate immediately begins implementing a decision after reviewing the particular policy (e.g., drafting an email to the insubordinate employee), the candidate likely utilized a noncompensatory strategy. On the other hand, if the candidate continues to collect information after reviewing the employee conduct policy and then makes a decision, the assessor would interpret the actions as adhering to a compensatory strategy. In other cases, a candidate may continue to collect vast amount of data but ultimately base the decision on a single bit of information (such as the formal rules about employee conduct), which would be deemed an inefficient combination of compensatory and noncompensatory integration strategies.

11.4.3.1 Integrating New and Conflicting Information

In the dynamic environment that ACs strive to simulate, how candidates deal with new, conflicting information is also of value to assessors. Soll and Larrick (2009) suggested that when decision makers receive conflicting information, they either average the information or they emphasize one piece and ignore the rest. Although averaging conflicting information often increases accuracy, most people choose one of the two conflicting sides.

One must consider several variables when determining which strategy (choosing vs. averaging) is most appropriate. First, the effectiveness of a strategy is contingent on the context in which it is applied. Averaging several pieces of information when many of them come from individuals unfamiliar with a particular situation might not be a good strategy. Conversely, if there are too many pieces of seemingly relevant information, relying on a single, trusted, expert opinion might be a prudent decision. Second, simply because a more basic strategy is available in a given situation does not mean that individuals will use the more basic strategy. For example, even though it may be more basic to select the opinion of a single expert, an individual may decide to attempt to average (or even perform a weighted average) of several bits of information. Third, despite having a repertoire of effective strategies for various situations, there is no guarantee that the strategies will be effectively adapted for the specific situation, or even applied at the correct time.

According to Soll and Larrick (2009), choosing is more accurate than averaging if the person can identify the "best" information through choosing. For example, one piece of information may be coming from a seasoned employee while a seemingly conflicting piece of information is coming from a new hire. If the situation mandated two different courses of action depending on which information was given more weight (or even if the two were averaged vs. picking one piece), inferences could be made regarding how the candidate processed the new information. In most situations, subject matter experts can determine which decision is most accurate and which piece of information most valid. By matching a candidate's behavior to that of experts, a new stream of behavioral data could be tapped. The AC practitioner can investigate how candidates change their decisions when new, conflicting information arrives, and whether they base any changes on the source's credibility and/or specific content of the new information.

11.4.4 Other Relevant Elements of Judgment and Decision Making

Other decisional components related to information acquisition, processing, and integration also exist, which could be captured and used in assessment center contexts. According to Ford et al. (1989), when tasks are complex, people tend to engage in more shallow searches, vary more in their search patterns, and, on average, spend less time searching for information. Task complexity is also related to the general depth of information searches, which suggests that the more complex a task, the more likely people are to rely on noncompensatory strategies to reduce their cognitive effort. On the other hand, the more important it is for the person to make the correct decision or the more accountable one is for a particular decision, the more complex the decision strategy may become (Walsh 1995). Typically, the decision maker perceives that such complex strategies lead to more accurate results. This more complex strategy could take the form of a compensatory decision strategy, which would require much more detailed information as well as more cognitive processing.

From an individual difference perspective, the more knowledgeable or experienced people are, the more likely they are to use a complex strategy (Ford et al. 1989). This reliance on more complex strategies is easier for experts as they are able to absorb information quickly due to their familiarity with a particular topic. Finally, experts within a particular area are better at remembering relevant information than novices (Chase and Simon 1973), and therefore experts require less cognitive effort than novices to make a decision. Experts, who possess more previous knowledge or experience than novices, have more cognitive capacity to accommodate complex tasks that requires a compensatory integration strategy. These subtle variations in behavior could be captured through process tracing, and interpreted alongside knowledge of candidates' past experience, to enhance AC ratings and the type of feedback given to candidates.

11.5 Practical Implications and Next Steps

After considering the wealth of behavioral data that can be captured through process tracing, a new breed of AC exercise begins to emerge. Although many of the behavioral data collection methods discussed in this chapter could potentially be added to current AC exercises, in order to take full advantage of the new data, specific components of the exercise would have to be redesigned such that differences in candidates' behavior would be meaningful. Although the overall feel and general functionality of a redesigned inbox exercise would remain the same, the behavioral data related to decision-making processes would provide more valuable insight into candidate performance.

One of the first steps toward a redesigned assessment center exercise is to consider specific situations or cues that would trigger the behavioral response of interest (Thornton and Mueller-Hanson 2004). Utilizing situational cues within exercises to elicit important behaviors has been discussed at length in the AC literature (e.g., Lievens 2008; Lievens and Schollaert 2011; Lievens et al. 2009). Essentially, if a particular dimension is to be measured in an exercise, the exercise must afford the opportunity for the behaviors that are representative of that dimension to emerge (and assessors must be able to *capture* the behavior). For example, a dimension such as "leadership" could more easily manifest in a leaderless group discussion than a simulated interaction with a client. In this vein, candidates working within an inbox exercise must have the opportunity to meaningfully engage in the decisional behaviors we have described in this chapter.

A very simple example of an inbox design change that would feed into the process tracing method we are proposing would involve how candidates access the information they need to complete the exercise. Traditionally, candidates are provided information in physical form (e.g., binders with company or employee information) prior to beginning the exercise or AC (Thornton and Mueller-Hanson 2004; Thornton and Rupp 2006). This procedure allows candidates to learn whatever baseline information is necessary for them to be fully immersed in the simulation (e.g., the position they will assume, general company information). While in some cases, this information has been digitized (i.e., an electronic version of the information is available to download), it is not necessarily possible to track actual information seeking behavior as we have described. Instead of such static environments, candidates would login to a program that would provide a dynamic, digital environment in which to seek out any information they would normally be provided ahead of time. Any behaviors the candidate engages in within this digital environment could be recorded for future interpretation (e.g., opening a file). By adding simple eye tracking hardware and software, the fidelity of the simulation can be maintained while capturing important behavioral data for assessors to review (i.e., exactly what a candidate is looking at or reading at any given moment).

This type of controlled digital environment can be maintained throughout the exercise such that other similar behaviors can be passively tracked while candidates are working. In order to bring these suggestions to life, a potential inbox exercise, adapted to incorporate the new behavioral data we have been discussing, will be described. We will include how the experience would unfold for the candidate and what the data would look like for the assessor.

11.5.1 The Next Generation Simulation

As with a typical inbox exercise, the candidate in our redesigned simulation would receive background information about the company and various people involved in the simulation. If candidates are reviewing the information in a controlled environment (e.g., an onsite or offsite assessment facility), eye tracking can be used in conjunction with the simulation environment to evaluate how they review this initial data. The software can be programmed to capture and flag broad behaviors such as thoroughness (e.g., did all the information appear to be reviewed), repetitiveness (e.g., was information reviewed multiple times), or erratic patterns (e.g., did the candidate jump from one piece of information to another in a nonintuitive order). Taken alone, such information would not necessarily mean anything; however, when considered with more typical exercise outputs during the rating process, assessors would be able to take a more clinical approach in reaching candidates' dimension ratings. In other words, by considering more holistically how the unique candidate—with all of his or her unique qualifications and past experiences—approaches the exercises, it could be determined whether the strategy taken was maximally efficient considering these contextual variables.

After being presented (and reviewing) any initial information necessary for the AC exercise (usually within a set time), candidates would begin the simulation. Again, the virtual environment in which they are working (checking/responding to emails, accessing new/old information, calendars, etc.) is designed such that even subtle behaviors like eye movements, response latencies, and skin conductance can be easily logged by software running in the background.

This special virtual environment takes on an increasingly important role as the exercise unfolds. For example, in our hypothetical inbox, candidates have been given one hour to review how the simulation works as well as baseline information about the company, several employees, and ongoing projects. The exercise then begins with four emails waiting in their inbox. Two of these emails have been objectively determined through studies with experts to be unimportant (i.e., they should be passed along for a virtual assistant to deal with). The other two have been designed to fall clearly into order of priority. One of these (Email A) requires immediate attention. Through software tracking, assessors would be provided with two pieces of information: (1) thoroughness/pattern of reviewing these four initial emails, and (2) properly identifying those to be passed along and those requiring immediate attention.

The behavioral data being collected now begins to take shape in the context of the simulation. Assessors will have a report of how the initial information was reviewed as well as behavioral patterns from the beginning of the simulation. Inefficient or erroneous behavior (e.g., not properly delegating, not identifying the email that requires immediate attention), could be automatically flagged to be reviewed by assessors at the end of the exercise (during the evaluation process). Even more complex behavior patterns could be reviewed automatically. For example, the software could detect that Email A was not addressed first, and automatically check the initial information search pattern from before the simulation started. If an abnormal (or incomplete) information search pattern was detected, particularly if it was recorded that a candidate skipped a very specific section that would have been pertinent to the start of the exercise, the behavior would be flagged in an automatically generated final report. For example, "Candidate failed to identify most important initial email. This is likely the result of relevant preliminary information not being reviewed." In a slightly different scenario, a candidate may simply start responding to email in the order they are presented in the inbox. In such a case, the automatically generated report might read, "Candidate replied to emails sequentially with no initial review—likely failed to prioritize tasks prior to beginning work." Consequently, within just the first few minutes of starting the simulation, automatically compiled data from process tracing methods are already beginning to provide additional information to the assessors.

Again, our hypothetical simulation has been designed around the decision making data afforded by process tracing, and several other "cues" await the candidate. Going back to the initial set of emails, another email (Email B) might require a specific bit of information from employee records (which are accessible now that the simulation has started, but was not prior to the simulation beginning). Since Email A was objectively more important, an ideal candidate would have recognized that fact and responded to it first; however, even if a candidate responded to Email B first, the way in which the email is handled is still subject to evaluation. For example, because Email B requires a specific bit of information located in employee records, the way a candidate accesses that information could be recorded, automatically analyzed, and flagged if necessary. While the content of the email sent by the candidate (the output of the exercise) would either contain the appropriate information or not, the process of information gathering could be important to assessors. In one instance, the candidate may neglect to search for the necessary information at all. In this case, responding to Email B would trigger an automatic flag to the effect of "Candidate failed to acquire necessary information to complete Email B," which would be provided to assessors at the end of the exercise. As a result of this extra piece of behavioral information, even if a candidate "guessed correctly" on how to address an issue, the lack of information seeking could provide relevant information for a specific dimension score. If a candidate were to spend a disproportionate amount of time seeking out the information necessary for responding to Email B, this could also be indicated to assessors as "Candidate spent X % more time than the average[1] searching for information."

Just as the initial information review process was assessed prior to the simulation beginning, new information seeking behavior could be flagged in the same way. For example, in addition to spending a large amount of time seeking out information for Email B, the output to assessors could also read "Candidate displayed a disorganized search pattern," or "Candidate reviewed the same information X more times than the average candidate." When combined with initial information review, a more thorough candidate profile begins to emerge. While two candidates might respond to Email B in the same way, one may quickly and efficiently read the email, go directly to the relevant information, respond accordingly, and move on. Another candidate, on the other hand, might read the email, begin searching for information (possibly even missing it in the first pass), finally shift from information seeking back to responding to the email, then move on.

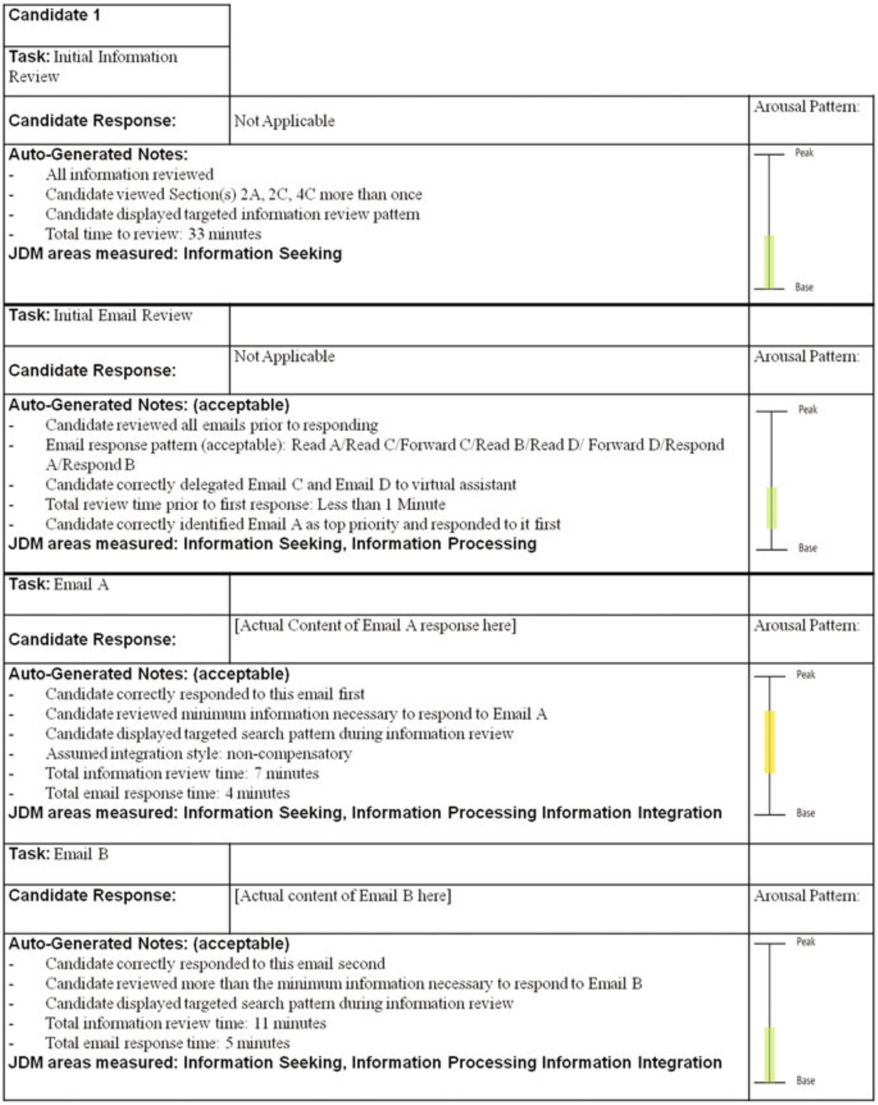
Moving even deeper into information that can be obtained and processed within our hypothetical inbox, various other dynamic situations can be created to elicit particular judgment and decision making behaviors. Continuing through our inbox simulation, the candidate has now addressed four emails (two being delegated). Immediately after clearing the inbox, another email would arrive, which involves some general task. While working on this particular email, yet another will arrive. It will be a response to Email B. Now the candidate is in a situation where new information regarding an old situation will be presented, and he/she will also be tasked with seeking out additional information. This particular scenario is designed such that the new information comes from an employee that would be objectively considered an expert on the situation (as per the presimulation materials). Because of the additional behavioral data that has been collected throughout the simulation, the candidates' ultimate response to this new information can now be more closely scrutinized. For example, it can be flagged whether or not the candidate reviewed the particular employee profile/record that identifies the employee as an expert. As usual, search patterns for new information can also be evaluated. At the end of the exercise, assessors would now have not only the candidate's official email response

[1] We are also implying that over time, or by virtue of incumbent validity studies, a normative sample could be collected with which to provide feedback on decisional strategies relative to the average employee, average expert, and the like. We also recommend benchmarking such data against industry, role, etc.

to the problem, but a number of behavioral flags to aid in the interpretation of the final product (e.g., “Candidate failed to read necessary employee information,” “Candidate failed to utilize necessary employee information,” “Candidate failed to recognize person X as an expert,” “Candidate displayed an erratic search pattern for new information,” “Candidate failed to request necessary information”).

Depending on exactly how the simulation is designed, at any point (or at multiple points) information could be arranged to test a candidate’s information integration strategy (linear vs. compensatory). Since the simulation is designed to detect what a candidate is looking at/reading at any given time, the response to a particular problem (an email) could have different outcomes depending on the integration strategy that is used. For example, throughout the exercise, another email might suddenly show up in the candidates’ inbox (Email E). This email also requires the candidate to seek out new information. Because information seeking is still being tracked, software can determine at what point a candidate stops looking for information and actually responds. For example, as the candidate is working through some new information in order to properly respond to Email E, they will come across a piece of information that would allow them to make an objective “all or nothing” decision. In other words, using this particular piece of information, the situation can be handled and no other information is necessary. However, the candidate could continue collecting information and ultimately arrive at a different decision by integrating several other pieces of information. Since the assessors will see and rate the final product (response to Email E) at the end of the simulation, they will now also know what information the candidate reviewed in arriving at the decision. An example that would differentiate candidates would be reading information that allows for an all or nothing decision, but continuing to collect more information, then still making a decision based on the all or nothing data. In addition to the final email with the decision made by the candidate, assessors would also have information such as “Candidate acquired only necessary information in responding to Email E,” or “Candidate acquired more than adequate information in responding to Email E.” In such a case, two candidates who made a decision based on all or nothing data might have otherwise been treated equally but instead could potentially be differentiated based on how much information they reviewed to make the decision. Similarly, if any information that could have helped in the decision was skipped or not used, this could also be brought to the assessor’s attention in a similar fashion.

Another component that should not be neglected during the AC exercise is potential insight into a candidate’s regulatory processing. With physiological data available primarily through skin conductivity measures (in addition to pulse, blood pressure, etc.), a general pattern of reaction (i.e., physiological arousal) to various components of the exercise could also be incorporated into dimension scores or feedback. For example, consider a scenario where a candidate is waiting for the simulation to begin, reading through information, and maintaining a baseline level of arousal. At the start, however, the candidate has a distinct spike in arousal just before responding to an email that should have been delegated. This same spike in arousal does not occur when appropriately responding to another email. Although assessors would need to be properly trained on how to interpret such data, for *this* candidate, the arousal spikes could indicate a level of uncertainty related to delegating. While the



Candidate 1		
Task: Initial Information Review		
Candidate Response:	Not Applicable	Arousal Pattern:
Auto-Generated Notes: - All information reviewed - Candidate viewed Section(s) 2A, 2C, 4C more than once - Candidate displayed targeted information review pattern - Total time to review: 33 minutes **JDM areas measured: Information Seeking**		Peak Base
Task: Initial Email Review		
Candidate Response:	Not Applicable	Arousal Pattern:
Auto-Generated Notes: (acceptable) - Candidate reviewed all emails prior to responding - Email response pattern (acceptable): Read A/Read C/Forward C/Read B/Read D/ Forward D/Respond A/Respond B - Candidate correctly delegated Email C and Email D to virtual assistant - Total review time prior to first response: Less than 1 Minute - Candidate correctly identified Email A as top priority and responded to it first **JDM areas measured: Information Seeking, Information Processing**		Peak Base
Task: Email A		
Candidate Response:	[Actual Content of Email A response here]	Arousal Pattern:
Auto-Generated Notes: (acceptable) - Candidate correctly responded to this email first - Candidate reviewed minimum information necessary to respond to Email A - Candidate displayed targeted search pattern during information review - Assumed integration style: non-compensatory - Total information review time: 7 minutes - Total email response time: 4 minutes **JDM areas measured: Information Seeking, Information Processing Information Integration**		Peak Base
Task: Email B		
Candidate Response:	[Actual content of Email B here]	Arousal Pattern:
Auto-Generated Notes: (acceptable) - Candidate correctly responded to this email second - Candidate reviewed more than the minimum information necessary to respond to Email B - Candidate displayed targeted search pattern during information review - Total information review time: 11 minutes - Total email response time: 5 minutes **JDM areas measured: Information Seeking, Information Processing Information Integration**		Peak Base

Fig. 11.1 Sample behavioral report, Candidate 1

implications that this type of arousal could easily vary between candidates, when taken as a piece of a much broader picture of the candidate that is being revealed, we can see the valuable insight provided by incorporating such information.

Figures 11.1 and 11.2 provide a snapshot of what some sample simulation output might look like. Candidate 1 is considered an expert with many years of experience, whereas Candidate 2, while experienced, has been in the field about half as long as Candidate 1. As we will discuss below, it is essential that objectively captured behavioral data is interpreted more holistically, with details of the candidate in mind.

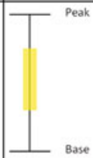
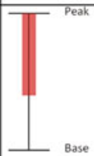
Candidate 2		
Task: Initial Information Review		
Candidate Response:	Not Applicable	Arousal Pattern:
Auto-Generated Notes: (caution*) - All information reviewed - Candidate viewed Section(s) 1B, 1C, 2A, 2B, 2C, 3A, 3B, 3C, 4A, 4C more than once - Candidate displayed erratic information review pattern* - Total time to review: 60 minutes **JDM areas measured: Information Seeking**		
Task: Initial Email Review		
Candidate Response:	Not Applicable	Arousal Pattern:
Auto-Generated Notes: (caution*) - Candidate did not review all emails prior to responding* - Email response pattern: Read D/Forward D/Read A/Respond A/Read C/Respond C/Read B/Respond B* - Candidate correctly delegated Email D to virtual assistant - Candidate failed to delegate Email C* - Total review time prior to first response: 2 Minutes **JDM areas measured: Information Seeking, Information Processing Information Integration**		
Task: Email A		
Candidate Response:	[Actual Content of Email A response here]	Arousal Pattern:
Auto-Generated Notes: (caution*) - Candidate correctly responded to this email first - Candidate reviewed more than minimum information necessary to respond to Email A* - Candidate displayed erratic search pattern during information review* - Assumed integration style: compensatory* - Total information review time: 18 minutes - Total email response time: 7 minutes **JDM areas measured: Information Seeking, Information Processing Information Integration**		
Task: Email B		
Candidate Response:	[Actual content of Email B here]	Arousal Pattern:
Auto-Generated Notes: (caution*) - Candidate responded to this email third* - Candidate reviewed more than the minimum information necessary to respond to Email B* - Candidate displayed targeted search pattern during information review - Total information review time: 11 minutes - Total email response time: 5 minutes **JDM areas measured: Information Seeking, Information Processing Information Integration**		

Fig. 11.2 Sample behavioral report, Candidate 2

This is because the effectiveness and efficiency of decisional strategies is often dependent on experience, practice, and familiarity with similar situations (Highhouse 1997; Kirschenbaum 1992; Walsh 1995).

11.5.2 Returning to a Clinical Approach

As the candidates work their way through the simulation, various behaviors will be continuously monitored and logged, and will in turn trigger behavioral flags that

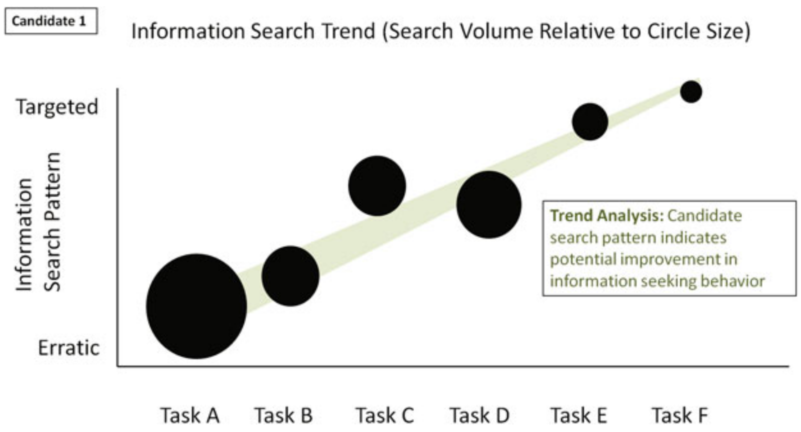


Fig. 11.3 Sample behavioral graph, Candidate 1

would be displayed in a final report. Along with actual responses, assessors would be better equipped to evaluate not just actions but the decision processes underlying those actions. The data could then be combined to aid in more accurate assessment of information integration over time. Finally, this method could allow assessors to pinpoint the exact point in which decisional effectiveness or efficiency was compromised (e.g., initial review, new information search, the processing of new information, decisional commitment and implementation). For example, given the particular response to Email B combined with what information was reviewed, it could be determined that Candidate 2 was engaged in a compensatory integration strategy when making a decision. Assuming the best strategy for that particular email was noncompensatory (as determined by experts, or as suggested by the normative data available), this suggests possible deficiencies in decision making for Candidate 2. If, on the other hand, Candidate 2 made the same decision regarding Email B as Candidate 1, but also spent time reviewing more information than necessary, it could potentially be inferred that Candidate 2 initially attempted to use a compensatory strategy, but ultimately decided on a noncompensatory one. In the latter instance, without the added behavioral information, the two candidates would have appeared equal in their response to Email B.

Essentially, assessors are armed with more information that provides access not only to candidates' work, but to their *working and decisional style*. By automatically processing this new stream of data and translating it into simple behavioral flags, trained assessors can begin to glimpse and interpret how a particular candidate *experienced* the simulation. Important behavioral trends could be readily identified that would have otherwise gone unnoticed. For example, suppose, as the simulation progresses, a candidate shifts from an erratic to a targeted search pattern while generally decreasing the amount of information that is reviewed prior to making a decision. This data could be presented graphically to aid the assessor in evaluating candidates more holistically. Figs. 11.3, 11.4, and 11.5 provide snapshots of what this might look like.

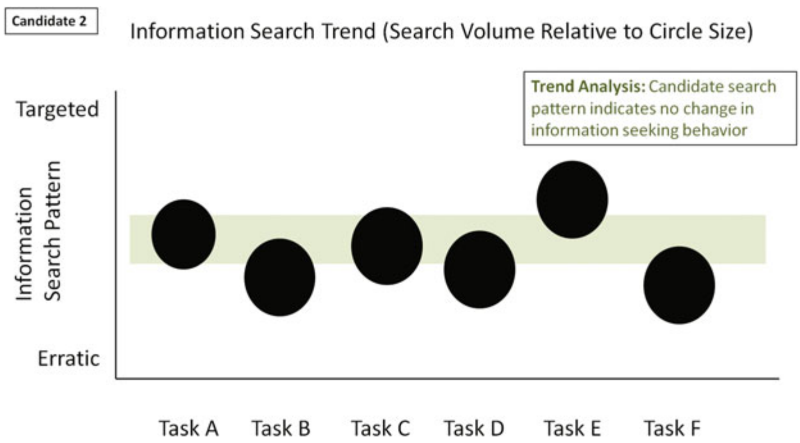


Fig. 11.4 Sample behavioral graph, Candidate 2

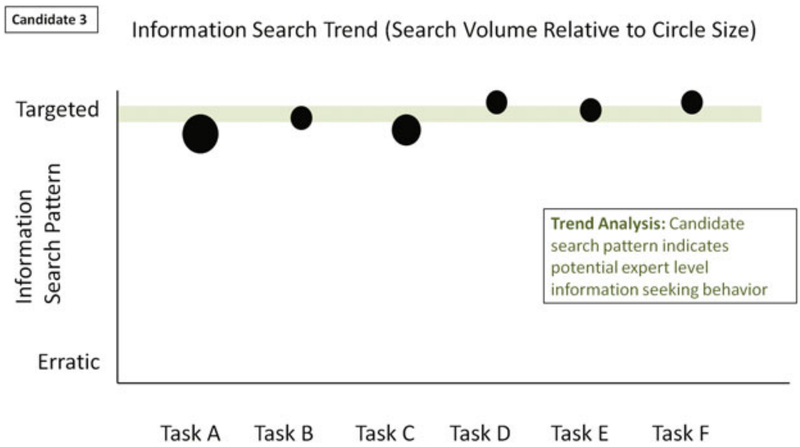


Fig. 11.5 Sample behavioral graph, Candidate 3

Using this type of output, assessors could be in a better position to evaluate how candidates perform on dimensions like judgment and decision making within an AC exercise. For example, imagine a scenario in which the output (email responses) of the three candidates is identical. With the additional behavioral data, a more subtle but important ranking emerges (Candidate 3, then Candidate 1, then Candidate 2). Such subtle behaviors have previously been beyond the reach of most AC simulations or, at the very least, would have required more assessor power to capture. By taking advantage of current technology and knowledge of meaningful differences in how people make decisions, we can potentially revolutionize assessment center simulations.

11.5.3 Implications for Research and Practice (e.g., Candidate Feedback)

By applying process tracing methods to ACs, we can gain new insights into the decisional process in a realistic work environment. Perrin et al. (2001) pointed out the need for more decision making research in real-world environments, and Highhouse (1997) also called for further applied research into the conditions underlying decisional errors. One of the hallmarks of the assessment center method is the focus on simulating real-world activities. Because AC simulations have relatively high fidelity and are also very standardized, ACs provide a living laboratory for uncovering specific insights into leader decision making, which can assist us in developing simulation exercises allowing for the capture of objective behavior relative to decisional processes.

If particular decisional strategies are shown to be more effective than others in a given situation, or among candidates with different sorts of experiences and individual differences, AC feedback could enhance candidates' development by comparing the decision paths one took with approaches shown to be more effective given their specific job and background. For example, if a candidate failed to recognize the uniqueness of a particular situation and instead relied too much on heuristics, the candidate could be cautioned when approaching a similar situation in the future. Conversely, if a candidate spent too much time on an objectively unimportant decision, training or coaching could be provided to the candidate on how to gauge problems in terms of their relative importance, or possibly on time management during the decision-making process.

An important aspect of utilizing the additional information provided by process tracing is the additional time or cost for the AC practitioner. Although upfront costs could be high (software design, additional hardware, new training for assessors), the final product is a wealth of new and potentially vital information. Furthermore, this type of process tracing has virtually no effect on candidates because their natural behavior is not affected by the collection of additional decisional data.

A natural next step after collecting, analyzing, and implementing process tracing data would be to test if the additional data captured via process training leads to incremental validity in both the prediction of job performance and the effectiveness of developmental feedback. It might also be useful to determine levels of discriminant validity with other constructs such as intelligence. Although decision-making ability and intelligence seem to overlap, it is mostly due to resistance to decision-making biases, such as information overload scenarios discussed earlier (Bruine de Bruin et al. 2007; Highhouse 1997). Analysis of decisional strategies can help identify candidates with the kind of experience and ability needed for a given job (Klein 1997). Even with the high validity coefficients (e.g., 71; Dilchert and Ones 2009) achieved by combining AC ratings with other measures (i.e., cognitive ability and personality tests), differences in decisional processes could help identify *true* experts (i.e., those that have the capability to leverage their intelligence, personality, and past experience to optimally navigate new complex environments, and make difficult decisions under pressure).

Another interesting avenue is determining what effects (if any) the capture and use of decisional strategy data has on adverse impact. Although current ACs fare well with mitigating adverse impact, particularly when compared with intelligence measures (for an exception, see Dean et al. 2008), incorporating decisional indices into AC ratings could increase the cognitive-loading of AC scores.

Gaining insight into how candidates make decisions has also implications for providing feedback, training, and coaching. Klein (1997) suggested that decision training may provide another tool to aid people in making better decisions. Previous work on this type of training has yielded a range of positive (Mann et al. 1991), neutral (Zakay and Wooler 1984), or even negative results (e.g., Driskill et al. 1994, Johnston et al. 1997). However, much of the research speaks to the difficulty in teaching a specific, often rigid strategy that is difficult or unwise to implement in a dynamic environment (Klein 1997). Adaptability is an important aspect of utilizing the best decisional strategy and making good decisions. Such adaptability is often observed in the decisional style of experts. Therefore, instead of teaching a particular strategy for making decisions, it seems beneficial to provide feedback of currently used and alternative strategies, along with discussion of what strategies may be best in the candidate's specific work environment (Klein 1997).

A final benefit that our enhanced assessment center framework provides goes beyond the candidates themselves and looks at how *assessors* make decisions. We have suggested that the added behavioral information collected by process tracing should be incorporated with other traditional assessment center ratings in a clinical fashion. However, it is impossible to ignore the fact that the assessors themselves take on a role which is not unlike the role of the candidate, in that they too must collect and synthesize information, and they must make decisions about how proficient candidates are on behavioral dimensions (and sometimes overall). Fortunately, just as a simulation can be created to recreate the work environment of a typical manager, the same can be done for an assessor (perhaps during frame of reference training, e.g., Schleicher et al. 2002). By applying process tracing methods to *assessor* training and evaluation, we might be able to further improve the validity of AC ratings by increasing the interrater reliability of assessors.[2]

11.6 Conclusion

Our purpose in applying process tracing to AC exercises is determining how leaders make decisions and if variability in decision processes affects performance in a meaningful way. Through the use of passive, noninvasive data collection methods afforded by current technology, candidates' behaviors that were previously difficult to observe provide the AC practitioner with insights relevant to assessment and development. Since technology can easily capture decisional data, it is possible that subtle behavioral patterns can shed light on candidates' decision processes. More

[2] The authors wish to thank George C. Thornton, III for his input on this matter.

importantly, by increasing the amount of behavioral data collected and analyzed, other dimensions that could not be previously assessed in ACs can be measured, which may result in more accurate ratings. Furthermore, the assessment of decision strategies, coupled with insights uncovered by our systematic analysis of person and situational determinants, allows for developmental feedback that focuses on the choice and use of relevant decisional strategies. By applying process tracing methods to ACs, we can discover how cognitive strategies affect decisions, how those decisions affect dimension scores, and above all, how decision strategies help predict and improve decision making in organizations, especially during volatile, uncertain, complex, and ambiguous times.

Acknowledgments This research was supported by the SIOP Foundation's Douglas W. Bray and Ann Howard Research Grant. The authors would like to thank Carolyn Jagacinski and James LeBreton for helpful feedback.

References

Abelson, R. P., & Black, J. B. (1986). Introduction. In J. A. Galambos, R. P. Abelson & J. B. Black (Eds.), *Knowledge structures* (pp. 1–18). Hillsdale: Lawrence Erlbaum Associates.

Bettman, J. R., & Park, C. W. (1980). Effects of prior knowledge and experience and phase of the choice process on consumer decision processes: A protocol analysis. *Journal of Consumer Research, 7*(3), 234–248.

Bruine de Bruin, W., Parker, A. M., & Fischoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92,* 938–956.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Cook, G. J., & Swain, M. R. (1993). A computerized approach to decision process tracing for decision support systems design. *Decision Science, 24,* 683–697.

Corporate Leadership Council (2005). Trends in developmental assessment centers. http://www.scribd.com/doc/57017184/Trends-in-Developmental-Assessment-Centers. Accessed 1 Dec 2011.

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology, 93*(3), 685.

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment, 17*(3), 254–270.

Driskell, J. E., Salas, E., & Hall, J. K. (1994). The effect of vigilant and hypervigilant decision training on performance. In Annual Meeting of the Society of Industrial and Organizational Psychology, Nashville.

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice* (2nd Edition). London: Springer.

Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual Review of Psychology, 52*(1), 581–606.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology, 32,* 53–88.

Evans, M. (2012). The key lessons from RIM's struggles. Forbes. From http://www.forbes.com/sites/markevans/2012/07/04/key-lessons-from-rim/. Accessed 1 Mar 2013.

Farr, J. L. (1973). Response requirements and primacy-recency effects in a simulated selection interview. *Journal of Applied Psychology, 57*(3), 228–232.

Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *Journal of Personality and Social Psychology, 91*(5), 797.

Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. In M. Schulte-Mecklenbeck, A. Kuehberger & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research: A critical review and user's guide* (pp. 163–184). New York: Psychology Press.

Ford, J. K., Schmitt, N., Schechtman, S. L., Hults, B. M., & Doherty, M. L. (1989). Process tracing methods: Contributions, problems, and neglected research questions. *Organizational Behavior and Human Decision Processes, 43,* 75–117.

Gibbons, A. M., Hughes, D., Riley, P., Thornton, G. C., & Sanchez, D. (2013). Is the future here? Current uses and perceived benefits of technology in operational assessment centers Unpublished Manuscript.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. USA: Oxford University Press.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1997). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections, and controversies (pp. 144–180). Cambridge series on judgment and decision making*. Cambridge, UK: Cambridge University Press.

Hayes-Roth, B. (1977). Evolution of cognitive structures and processes. *Psychological Review, 84*(3), 260.

Highhouse, S. (1997). Understanding and improving job-finalist choice: The relevance of behavioral decision research. *Human Resource Management Review, 7*(4), 449–470.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9,* 90–98.

Hughes, D., Riley, P., Shalfrooshan, A., Gibbons, A., Thornton, G. (2012). A global survey of assessment center practices. A research report by a & dc and Colorado State University.

International Task Force on Assessment Center Guidelines (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17,* 243–253.

Jacoby, J., Mazursky, D., Troutman, T., Kuss, A. (1984). When feedback is ignored: Disutility of outcome feedback. *Journal of Applied Psychology, 69*(3), 531–545.

Jasper, J. D., & Shapiro, J. (2002). MouseTrace: A better mousetrap for catching decision processes. *Behavior Research Methods, Instruments, & Computers, 34,* 364–374.

Johnston, J. H., Driskell, J. E., & Salas, E. (1997). Vigilant and hypervigilant decision making. *Journal of Applied Psychology, 82*(4), 614–622.

Kirschenbaum, S. S. (1992). Influence of experience on information-gathering strategies. *Journal of Applied Psychology, 77*(3), 343–352.

Klein, J. D. (1997). ETR & D-Development: An analysis of content and survey of future direction. *Educational Technology Research and Development, 45*(3), 57–62.

Koop, G. J., & Johnson, J. G. (2011). Response dynamics: A new window in the decision process. *Judgment and Decision Making, 6*(8), 750–758.

Lievens, F. (2008). What does exercise-based assessment really mean? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(1), 112–115.

Lievens, F., & Schollaert, E. (2011). Reshaping exercise design in assessment centers: Theory, practice, and research. In N. Povah & G. C. Thornton III (Eds.), *Assessment centres and global talent management (pp. 47–60)*. Surrey, UK: Gower.

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. *Research in Personnel and Human Resources Management, 28,* 99–152.

Mann, L., Harmoni, R., & Power, C. (1991). The GOFER course in decision making. In J. Baron & R. V. Brown (Eds.), *Teaching decision making to adolescents* (pp. 61–78). Hillsdale: Lawrence Erlbaum Associates, Inc.

McGregor, J. (2012). In cruise ship sinking, leadership failures from captain to Carnival CEO. The Washington Post Online. http://www.washingtonpost.com/blogs/post-leadership/post/in-cruise-ship-sinking-leadership-failures-from-captain-to-carnival-ceo/2011/04/01/gIQACNtkAQ_blog.html. Accessed 1 Feb 2012.

Miller, C. C., & Kopytoff, V. G. (2011). Once a leader, Yahoo now struggles to find its way. The New York Times Online. http://www.nytimes.com/2011/09/08/technology/once-a-leader-yahoo-now-struggles-to-find-its-way.html?pagewanted=all&_r=0. Accessed 1 Mar 2012.

Miller, J. W., & Rowe, P. M. (1967). Influence of favorable and unfavorable information upon assessment decisions. *Journal of Applied Psychology, 51*(5), 432–435.

National Bureau of Economic Research. (2008). Determination of the December 2007 peak in economic activity. National Bureau of Economic Research. From http://www.nber.org/cycles/dec2008.pdf. Accessed 1 Nov 2011.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13*(2), 248–277.

Onken, J., Hastie, R., & Revelle, W. (1985). Individual differences in the use of simplification strategies in a complex decision-making task. *Journal of Experimental Psychology: Human Perception and Performance, 11*(1), 14–27.

Patrick, J., & James, N. (2004). Process tracing of complex cognitive work tasks. *Journal of Occupational and Organizational Psychology, 77,* 259–280.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An informational search and protocol analysis. *Organizational Behavior and Human Performance, 16,* 366–387.

Payne, J. W. (1982). Contingent decision behavior. *Psychological Bulletin, 92*(2), 382–402.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Payne, J. W., Braunstein, M. L., & Carroll, J. S. (1978). Exploring predecisional behavior: An alternative approach to decision research. *Organizational Behavior and Human Performance, 22*(1), 17–44.

Perrin, B. M., Barnett, B. J., Walrath, L., & Grossman, J. D. (2001). Information order and outcome framing: An assessment of judgment bias in a naturalistic decision-making context. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 43*(2), 227–238.

Povah, N., & Thornton, G. C. III. (2011). *Assessment centers and global talent management. Surrey,* UK: Gower Publishing Limited.

Reynolds, D. H., & Dickter, D. N. (2010). Technology and employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection*. New York: Psychology Press.

Reynolds, D. H., & Rupp, D. E. (2010). Advances in technology-facilitated assessment. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent* (pp. 609–641). San Francisco: Jossey-Bass.

Riedl, R., Brandstätter, E., & Roithmayr, R. (2008). Identifying decision strategies: A process- and outcome-based classification method. *Behavioral Research Methods, 40*(3), 795–807.

Rigney, J. W., & DeBow, C. H. (1967). Multidimensional scaling analysis of decision strategies in threat evaluation. *Journal of Applied Psychology, 51*(4), 305–310.

Rowe, P. M. (1967). Order effects in assessment decisions. *Journal of Applied Psychology, 51*(2), 170–173.

Rupp, D. E., Gibbons, A. M., & Snyder, L. A. (2008). The role of technology in enabling third-generation training and development. *Industrial and Organizational Psychology, 1*(4), 496–500.

Russo, J. E. (2011). Eye fixation as a process trace. In M. Schulte-Mecklenbeck, A. Kühberger & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research: A critical review and user's guide* (pp. 43–64). New York: Psychology Press.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for Frame-of-Reference Training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*(4), 735–746.

Sherer, P. D., Schwab, D. P., & Heneman, H. G. III (1987). Managerial salary-raise decisions: A policy-capturing approach. *Personnel Psychology, 40,* 27–38.

Simon, Z. (2012). Malev airline stops services after 66 years as Hungary moves to cut losses. Bloomberg Online. From http://www.bloomberg.com/news/2012–02-03/malev-stops-flying-after-66-years-as-hungary-cuts-its-losses.html. Accessed 1 Mar 2012.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research, 16,* 158–174.

Sims, H. P., & Gioia, D. A. (1986). *The thinking organization*. San Francisco: Jossey-Bass.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory,and Cognition, 35*(3), 780.

Spector, M., & Trachtenberg, J. A. (2011). Borders forced to liquidate, close all stores. The Wall Street Journal Online. http://online.wsj.com/article/SB10001424052702303661904576454353768550280.html. Accessed 1 Jan 2013.

Starbuck, W. H., & Milliken, F. (1988). Executives' perceptual filters: What they notice and how they make sense. In D. Hambrick (Ed.), *The executive effect: Concepts and methods of studying top managers (pp. 35–65)*. Greenwich, CT: JAI Press.

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance, 23,* 86–112.

Tabatabai, M. (1998). Investigation of decision making process: A hypermedia approach. *Interacting With Computers, 9,* 385–396.

Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology, 57*(3), 388.

Thorngate, W. (1980). Efficient decision heuristics. *Behavioral Science, 25*(3), 219–225.

Thornton, G. C. III, & Mueller-Hanson, R. (2004). *Developing organizational simulations: A guide for practitioners and students. Series in applied psychology*. Mahwah: Lawrence Erlbaum Associates.

Thornton, G. C. III, & Rupp, D. E. (2006). Assessment centers in human resource management: Strategies for prediction, diagnosis, and development. Mahawah: Lawrence Erlbaum Associates, Inc.

Walsh, J. P. (1995). Managerial and organizational cognition: Notes from a trip down memory lane. *Organizational Science, 6*(3), 280–321.

Wang, J. T. Y. (2011). Pupil dilation and eye tracking. In M. Schulte-Mecklenbeck, A. Kühberger & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research: A critical review and user's guide* (pp. 185–204). New York: Psychology Press.

Worthen, B. (2012). H-P, Dell struggle as buyers shun PCs. The Wall Street Journal Online. http://online.wsj.com/article/SB10000872396390444812704577605703329715394.html. Accessed 1 Mar 2012.

Zakay, D., & Wooler, S. (1984). Time pressure, training and decision effectiveness. *Ergonomics, 27*(3), 273–284.

Zukier, H. (1982). The dilution effect: The role of the correlation and the dispersion of predictor variables in the use of nondiagnostic information. *Journal of Personality and Social Psychology, 43*(6), 1163.

Chapter 12
Future Directions

Michael Fetzer

The previous chapters cover a wide range of topics relevant to simulations used for personnel selection purposes, and represent the most quintessential and timely collection of research and information in this area to date. Taken together, it is clear that our field has come a long way, even more so in the past decade than any previous decade. The technology to develop and deliver simulation-based assessments has become more accessible to both creators and end-users of simulations in terms of cost and level of effort. In tandem, the power and sophistication of the tools involved with simulation development and deployment has increased at an exponential pace. Using a high-end personal computer (PC), one person can now do things in a week that only 10 years ago would have taken an entire team, multiple servers, and months to produce. Thinking about what we will be able to do 10 years from now is as exciting as it is mind-bending. Towards this end, this chapter covers three broad characteristics that will likely, with varying degrees, underlie all future simulations designed for talent measurement purposes: increased complexity, enhanced realism, and higher engagement. These three aspects are not mutually exclusive, nor are they mutually inclusive.

12.1 Increased Complexity

We are beginning to see an increase in the complexity of the simulations available for personnel selection, but we have only scratched the surface. Simulation complexity is best thought of in terms of: (1) the external (candidate) experience, namely what the candidate sees, hears, and does during the course of the simulation, and (2) the internal design, including how the simulation progresses and is scored. In some cases, one area of complexity can be increased without impacting the other

M. Fetzer (✉)
CEB Alpharetta, GA, USA
e-mail: Michael.Fetzer@shl.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8_12,

(see Guidry et al. 2013, this volume, for a prime example of how to increase the complexity of the internal design without significantly impacting the candidate experience). In most cases, however, these two areas will be inexorably linked.

For the candidate, the complexity of the job and the complexity of the simulation will become more aligned. We have made tremendous advancements with multimedia and other simulation technologies that enable us to deliver an assessment experience that is more like the target job than ever before. The use of video and computer animation have increased simulation complexity due to the fact that candidates now have to process multiple cues (visual/auditory, body language, tone of voice, etc.) and multiple inputs (e.g., dynamic gauges in a manufacturing simulation, email/voicemail/calendar/visitors in a managerial inbox simulation). However, most simulations are still very linear in nature, where candidate responses/inputs at one stage of the simulation don't impact the experience in subsequent stages of the simulation. There are benefits to providing a consistent (linear) simulation experience across candidates, and linear simulations have proven to be strong predictors of job performance.

At the intersection of the external candidate experience and the internal design lies the main element of simulation complexity that will see significant advances in the not-so-distant future. Through various methods, including (but not limited to) branching, adaptive, and non-linear programming (Sydell et al. 2013, this volume), as well as continuing advances in the field of artificial intelligence (e.g., Bhatia and Mago 2012; Lucci and Kopec 2012), simulations will begin to look and act more like the jobs they are simulating. In essence, we will see an ever-increasing number of different "paths" a candidate can take during the course of a simulation, and an exponential increase in the number of different outcomes possible. This intersection will provide further enhancements to simulation realism (as discussed in the latter sections), as real life is rarely, if ever, a linear experience.

In addition to an increasingly nonlinear experience, advances in technology will have a tremendous impact on the internal design of simulations with regards to the types of input possible as well as the amount of data that can be collected and scored. It is possible, even today, to collect data from a number of different input sources, including mouse clicks/movement, microphones, webcams, touchscreens, pen/graphics tablets, eye tracking devices, motion capture systems (hand, facial, full body), biofeedback devices (to measure brain activity, skin conductance, heartbeat, respiration rate, etc.), and of course, the keyboard. Although it is technically possible to "hook up" a candidate to all of these devices, there needs to be careful consideration with regards to the relevance and ultimate value of the data produced by these input methods. Theoretical linkages between what is being measured and job-related KSAOS must be established (as exemplified in Guidry et al., this volume) before we start leveraging some of these devices for simulation purposes. Once these linkages have been established, the complexity of simulation internal design and, in most cases candidate experience, will advance at a rapid pace.

Finally, it almost goes without saying, but caution should be taken during the simulation development phase to determine the incremental value of increasing simulation complexity. In other words, keep in mind the concept of Maslow's Hammer, which eloquently captures the propensity to see everything as a nail, if all you have is a hammer (Maslow 1966). I'm not suggesting that simulation complexity should not be increased, but no small amount of thought should be spent on the degree to which complexity should be increased with regards to the measurement and validity gains achieved.

12.2 Enhanced Realism

All of the multimedia simulations discussed in the previous chapters (and nearly all others available today) have one thing in common: they are more realistic than what was available 10 (or even 5) years ago. Advances in video production and computer-generated animation have provided many opportunities to enhance simulation realism on many levels. In general, the vast majority of simulations used in personnel selection context will become even more realistic in the future. In other words, they will become more like the jobs they are intended to aid in the identification of the candidates who would eventually perform best in those particular roles (notable exceptions are discussed in the next section). This should be no surprise, as both conceptually and empirically, the best predictor of future job performance is a sample of work performed (e.g., Asher and Sciarrino 1974; Hunter and Hunter 1984; Schmidt and Hunter 1998; Wernimont and Campbell 1968).

We have already seen a great deal of progress in this area, and several chapters in this book provide ample supporting evidence. For example, Gutierrez and Meyer (2013) described the development of realistic coaching and inbox simulations for managerial roles, given the extent of time most managers spend responding to emails, voicemails, and critical incidents. Holland and Lambert (2013) detailed the realistic nature of some of today's call center simulations, utilizing simulated caller audio clips, multiple systems/screens to access relevant information, and even the methods used to document the outcomes of caller interactions and responses to caller inquiries. O'Connell et al. (2013) outlined the increased fidelity of manufacturing simulations, as most manufacturing environments are becoming increasingly automated and can be represented more realistically through computer-based simulations. For the service sector, LaTorre and Bucklan (2013) highlighted the degree of realism that has been embedded within simulations used to select retail sales, customer service, and even financial services employees. Finally, Barr and Coughlin (2013) indicated the extent to which software application simulations are becoming more representative of the work tasks involved in the jobs they are used to evaluate candidate qualifications.

The increase in simulation complexity, described in the previous section, will be a primary driver in the increase in simulation realism. Simulations that "look, feel, and act" like real life will become increasingly possible with advances in candidate experience and internal design complexity. However, a more complex simulation is not necessarily more realistic. To fully realize the potential for enhanced simulation realism, we will continue to draw from advances in the entertainment industry (video games, movies, etc.). Many of the same tools and processes available to the likes of Ubisoft™, Electronic Arts™, and Pixar Animation Studios™ are now accessible to simulation developers. Consequently, the rise of companies in the serious games industry has been a direct result. The application of game mechanics and technologies, begrudgingly known to most as *gamification*, has been primarily focused in the area of training (more on this in the latter sections). However, we are starting to see an influx of this approach in the design of personnel selection simulations, and I fully expect this trend to continue and expand.

If there are any doubts about how realistic simulations can become with regards to the multimedia and animation techniques available even today, I would encourage the interested reader to explore some of the most popular video games on the market (e.g., Battlefield 4, Metro 2033, Crysis 3, Project C.A.R.S, Star Wars 1313); hit the latest full-length animated movie at your local theater and pay close attention to the detail in the background, props, characters, and motions, or simply go online and watch *The Third & The Seventh*, an animated short film by Alex Roman that could fool many into believing it is completely "live action" video.

12.3 Higher Engagement

As the level of simulation complexity and realism continues to climb, I fully expect the level of engagement (the degree to which candidates become "immersed" in the simulation) to reach new heights. The Holy Grail for developers of simulations used for personnel selection purposes is to *achieve a threshold of engagement where the candidate forgets he/she is being assessed*, thus exhibiting true behaviors rather than those that might be displayed when the candidate is trying to second guess the simulation. The methods and technologies that will be used to enhance complexity and realism will propel us towards this lofty goal, but one additional element can help us reach this goal more quickly—the "fun" factor. This element is so critical that we may even be able to reach our goal without significant increases in realism or complexity (more on this in the next section).

The fun factor encompasses the move towards embedding certain elements of game mechanics into our simulations. Drawing from the research of Shute and Ke (2012), these elements can include: interactive problem solving, setting specific goals and/or rules, adaptive challenges, candidate control, feedback, uncertainty, and sensory stimuli. Not surprisingly, these elements are also reflected in the work of Csikszentmihalyi (1990) with regards to the state of "flow," Coleridge (1906) and the "suspension of disbelief," and subsequently in Ermi and Mayra's (2005) model of immersion.

Now, I realize that there are only a few of us out there that are lucky enough to be able to consider our job fun, so I'm not suggesting that a job must be fun in order for a simulation of that job to be considered fun in its own right. What I am suggesting is that there are certain game elements that can be built into even the most mundane of job simulations in order to increase engagement, especially in a personnel selection situation. A simple test of the "fun factor" degree is whether or not someone (other than the simulation developer) who has completed it once wants to go back for a second turn. Consider the number of times people play Angry Birds™ and you'll understand how the fun factor can be invaluable.

As alluded to earlier, there may be situations where highly complex and/or realistic job simulations are not needed. There are games in use today that look nothing like a job, but could be used to evaluate certain candidate qualifications. For example, Nintendo's *Big Brain Academy* or Microsoft's *Brain Challenge* are essentially cognitive ability tests in an entertainment setting. I would hypothesize players who are successful in leading teams in massively multiplayer online role-playing games (MMORPGs) like *World of Warcraft* and *Global Agenda* are more likely to be effective leaders in a corporate setting compared to those players who are not successful. In other words, certain skills that are useful in environments that are not (by any means) job realistic may in fact transfer to job performance. Other examples may include project management, multitasking, and strategic thinking skills required to be successful in games such as *SimCity*, *Roller Coaster Tycoon*, and *Age of Empires*.

I'm not suggesting that organizations utilize existing video games for personnel selection purposes. My point is that games that have been designed purely for entertainment purposes and are not "face valid" from a job realism perspective may in fact be measuring the same KSAOs we are attempting to evaluate through high fidelity simulations. Thus, we need to explore the potential for implementing "serious games" that are not at all realistic compared to the job, but measure the relevant KSAOs and possess a higher degree of the fun factor than a straight job simulation. Increasing the degree of engagement by producing an assessment that is "fun" is something we can all aspire to, and the best source of inspiration may be sitting right in our living room. Some early examples of this approach are starting to emerge, including the work in progress by Personnel Decisions Research Institutes (PDRI), the Corporate Executive Board (CEB), and Knack.it.

12.4 Conclusion

Simulations used for personnel selection purposes are significantly more advanced than their predecessors in use a decade ago, and it is safe to say we will be making the same statement 10 years from now. There is much work to be done with regards to the standard psychometric, validation, and other research required to support the increasing use of simulations for selection, and the leap in complexity of simulation design and scoring will challenge us to keep this work grounded in solid science. However, as we enhance the complexity, realism, and engagement factor of these simulations, we move closer and closer to assessments that provide more accurate measurement, deeper insight, and ultimately higher levels of validity that we have been able to achieve in the past, across all methods of assessment. It is a very exciting time, to be sure.

References

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519–533.

Barr, R., & Coughlin, C. (2013). Designing and implementing software and computer skills simulations. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Bhatia, N., & Mago, V. K. (2012). *Cross-disciplinary applications of artificial intelligence and pattern recognition: Advancing technologies*. PA: Business Science Reference.

Coleridge, S. (1906). *Biographia literaria*. New York: E.P. Dutton.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row.

Ermi, L., & Mayra, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. In S. Castell & J. Jenson (Ed.). Print proceedings. Changing views: Worlds in play, digital games research association conference proceedings. Vancouver, British Columbia.

Guidry, B. W., Rupp, D. E., & Lanik, M. (2013). Tracing cognition with assessment center simulations: Using technology to see in the dark. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Gutierrez, S. L., & Meyer, J. M. (2013). Managerial simulations. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Holland, B., & Lambert, D. (2013). How to measure contact center skills using multimedia simulations. In M. S. Fetzer and K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

LaTorre, J., & Bucklan, M. A. (2013). *Simulations for service roles*. In M. S. Fetzer and K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Lucci, S., & Kopec, D. (2012). *Artificial intelligence in the 21st century*. VA: Mercury Learning and Information.

Maslow, A. (1966). *The psychology of science*. New York: Harper & Row.

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.

O'Connell, M., Lawrence, A., & Kinney, T. (2013). Show me you can do it: The use of interactive simulations in manufacturing settings. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.). Assessment in game-based learning: Foundations, innovations, and perspectives. New York: Springer.

Sydell, E., Ferrell, J., Carpenter, J., Frost, C., & Brodbeck, C. C. (2013). Simulation scoring. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection*. New York: Springer.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, 52, 372–376.

Index

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection*,
DOI 10.1007/978-1-4614-7681-8,