Wiley Handbooks in Education

THE HANDBOOK OF COGNITION AND ASSESSMENT

FRAMEWORKS, METHODOLOGIES, AND APPLICATIONS

Edited By
ANDRÉ A. RUPP
AND JACQUELINE P. LEIGHTON

WILEY Blackwell

The Handbook of Cognition and Assessment

The Wiley Handbooks in Education offer a capacious and comprehensive overview of higher education in a global context. These state-of-the-art volumes offer a magisterial overview of every sector, sub-field and facet of the discipline – from reform and foundations to K–12 learning and literacy. The *Handbooks* also engage with topics and themes dominating today's educational agenda – mentoring, technology, adult and continuing education, college access, race and educational attainment. Showcasing the very best scholarship that the discipline has to offer, *The Wiley Handbooks in Education* will set the intellectual agenda for scholars, students, researchers for years to come.

The Wiley Handbook of Learning Technology
Edited by Nick Rushby and Daniel W. Surry

The Handbook of Cognition and Assessment
Edited by André A. Rupp and Jacqueline P. Leighton

The Handbook of Cognition and Assessment

Frameworks, Methodologies, and Applications

Edited by

André A. Rupp and Jacqueline P. Leighton

WILEY Blackwell

This edition first published 2017

Registered Office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Catalog Number: 2016036147

ISBN Hardback: 9781118956571

A catalogue record for this book is available from the British Library.

Cover image: Claire Sower, Pool of Dreams (2016) / Claire Sower, Falling in Love (2016)

Set in 10.5/12.5pt Minion by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

To Brooke, my only possible soulmate and the most loving, inspiring, and simply fabulous partner I could ever hope for as well as to Jean-Marie, my truly amazing, compassionate, and delightful son - you two are my family and will always be immensely loved!

André A. Rupp

To my husband and best friend, Greg.

Jacqueline P. Leighton

Contents

Notes on Contributors

Jessica J. Andrews is an Associate Research Scientist in the Computational Psychometrics Research Center at Educational Testing Service (ETS) in Princeton, NJ. She received her Ph.D. in Learning Sciences at Northwestern University. Her research examines the cognitive processes underlying collaborative learning, and the use of technological environments (e.g., simulations, learning management systems) in supporting student learning and assessing individuals' cognitive and noncognitive (e.g., collaborative) skills.

Ryan S. Baker is Associate Professor of Cognitive Studies at Teachers College, Columbia University, and Program Coordinator of TC's Masters of Learning Analytics. He earned his Ph.D in Human-Computer Interaction from Carnegie Mellon University. Dr. Baker was previously Assistant Professor of Psychology and the Learning Sciences at Worcester Polytechnic Institute, and served as the first Technical Director of the Pittsburgh Science of Learning Center DataShop, the largest public repository for data on the interaction between learners and educational software. He was the founding president of the International Educational Data Mining Society, and is currently Associate Editor of the *Journal of Educational Data Mining*. He has taught two MOOCs, Big Data and Education (twice), and (co-taught) Data, Analytics, and Learning. His research combines educational data mining and quantitative field observation methods to better understand how students respond to educational software, and how these responses impact their learning. He studies these issues within intelligent tutors, simulations, multi-user virtual environments, MOOCs, and educational games.

John T. Behrens is Vice President, Advanced Computing & Data Science Lab at Pearson and Adjunct Assistant Research Professor in the Department of Psychology at the University of Notre Dame. He develops and studies learning and assessment systems that integrate advances in the learning, computing, and data sciences. He has written extensively about the use of evidence-centered design to guide development of

complex educational systems as well as about the foundational logics of data analysis/data science and the methodological impacts of the digital revolution.

Isaac I. Bejar holds the title of Principal Research Scientist with Educational Testing Service (ETS) in Princeton, NJ. He is interested in improving methods of testing by incorporating advances in psychometric theory, cognitive psychology, natural language processing, and computer technology. He was a member of the editorial board and advisory board of *Applied Psychological Measurement* from 1981 to 1989, and was awarded the ETS Research Scientist Award in 2000. He published *Cognitive and Psychometric Analysis of Analogical Problem Solving* and co-edited *Automated Scoring of Complex Tasks in Computer-Based Testing*.

Laine Bradshaw is an Assistant Professor of Quantitative Methodology in the Educational Psychology Department in the College of Education at the University of Georgia (UGA). Her primary research focuses on advancing multidimensional psychometric methodology to support the diagnostic assessment of complex knowledge structures for educational purposes. With a Master's degree in Mathematics Education, she is also active in collaborations on interdisciplinary assessment development projects that require tailoring psychometrics to cognitive theories. Her work has been published in journals such as *Psychometrika* and *Educational Measurement: Issues and Practice*. Her early career program of research was recently recognized by the National Council of Measurement in Education's Jason Millman Award.

Gabrielle Cayton-Hodges is a Research Scientist in the Learning Sciences Group at Educational Testing Service (ETS) in Princeton, NJ. She earned her BS degree in Brain and Cognitive Sciences from MIT and her PhD in Mathematics, Science, Technology, and Engineering Education from Tufts University. Gabrielle's specialty is mathematical cognition and elementary mathematics education, focusing on the application of cognitive and learning sciences to mathematics assessment and the use of technology to support innovative approaches to gathering evidence about what students know and can do. She has a specific expertise in student understandings of numerical concepts such as place value and the use of multiple representations in mathematics and has also spent several years studying early algebra and learning progressions in the understanding of area and volume.

Sun-Joo Cho is an Assistant Professor at Peabody College, Vanderbilt University.
Her research topics include generalized latent variable modeling and its parameter estimation, with a focus on item response modeling.

Ying Cui is an Associate Professor at the University of Alberta. Her research interests include cognitive diagnostic assessment, person fit analysis, and applied statistical methods.

Paul De Boeck is Professor of Quantitative Psychology at The Ohio State University and emeritus from the KU Leuven (Belgium). He is especially interested in how psychometric models can be redefined as explanatory models or supplemented with explanatory components for applications in psychology and education.

Kristen E. DiCerbo's research program centers on digital technologies in learning and assessment, particularly on the use of data generated from interactions to inform instructional decisions. She is the Vice President of Education Research at Pearson and has conducted qualitative and quantitative investigations of games and simulations, particularly focusing on the identification and accumulation of evidence. She previously worked as an educational researcher at Cisco and as a school psychologist. She holds doctorate and master's degrees in Educational Psychology from Arizona State University.

Susan Embretson is Professor of Psychology at the Georgia Institute of Technology. Previously, she was Professor at the *University of Kansas*. Her research concerns integrating cognitive theory into psychometric item response theory models and into the design of measurement tasks. She has been recognized for this research, including the Career Contribution Award (2013) and the Technical and Scientific Contribution Award (1994–1997) from the *National Council on Measurement and Education*; the Distinguished Lifetime Achievement Award (2011) from the *American Educational Research Association: Assessment and Cognition*; and the Distinguished Scientist Award from *American Psychological Association Division (5) for Measurement, Evaluation and Statistics* for research and theory on item generation from cognitive theory. Embretson has also served as president for three societies in her area of specialization.

Gary Feng is a Research Scientist in the Research and Development division at Educational Testing Service (ETS) in Princeton, NJ. He works in the Cognitive, Accessibility, and Technology Sciences Center. He received his PhD in Developmental Psychology and MS in Statistics from the University of Illinois at Champaign-Urbana. Before joining ETS, he was a faculty member at Duke University and held visiting and research positions at the University of Michigan and the University of Potsdam, Germany. He is broadly interested in the acquisition of reading skills and neurocognitive processes in reading. His past work uses eye-tracking to examine cognitive processes of skilled and developing readers across different cultures. Gary contributes to the development of innovative literacy assessments.

Steve Ferrara was Vice President for Performance Assessment and led the Center for Next Generation Learning and Performance in Pearson's Research and Innovation Network. Steve conducts psychometric research and designs large scale and formative assessments and automated language learning systems. He specializes in principled design, development, implementation, and validation of performance assessments and in research content, cognitive, and linguistic response demands placed on examinees and predicts technical characteristics of items. Steve earned an MEd in Special Education from Boston State College and an EdS in Program Evaluation and a PhD in Educational Psychology and measurement from Stanford University.

Mark J. Gierl is Professor of Educational Psychology and the Director of the Centre for Research in Applied Measurement and Evaluation (CRAME) at the University of Alberta. His specialization is educational and psychological testing, with an emphasis on the application of cognitive principles to assessment practices. Professor Gierl's current research is focused on automatic item generation and automated essay scoring.

His research is funded by the Medical Council of Canada, Elsevier, ACT Inc., and the Social Sciences and Humanities Research Council of Canada. He holds the Tier I Canada Research Chair in Educational Measurement.

Janice D. Gobert is a Professor of Learning Sciences and Educational Psychology at Rutgers. Formerly, she was the Co-director of the Learning Sciences and Technologies Program at Worcester polytechnic Institute. Her specialty is in technology-based with visualizations and simulations in scientific domains; her research areas are: intelligent tutoring systems for science, skill acquisition, performance assessment via log files, learning with visualizations, learner characteristics, and epistemology. She is also the Founding CEO of a start-up company named Apprendis (www.apprendis.com), whose flagship products are Inq-ITS and Inq-Blotter, both described in the chapter.

José P. González-Brenes is a Research Scientist in the Center for Digital Data, Analytics & Adaptive Learning at Pearson. He investigates methods of machine learning to make education faster, better, and less expensive. His work has been nominated for best paper awards in the International Educational Data Mining and the Special Interest Group of Dialogue Systems conferences. He is the happy first-prize winner of international data mining competition involving over 350 teams. His postgraduate training includes a PhD in Computer Science from Carnegie Mellon University and an IMBA in Technology Management from National Tsing Hua University in Taiwan.

Joanna S. Gorin is Vice President of Research at Educational Testing Service (ETS) in Princeton, NJ. As Vice President for Research, she is responsible for a comprehensive research agenda to support current and future educational assessments for K–12, higher education, global, and workforce settings. Her research has focused on the integration of cognitive theory and psychometric theory as applied to principled assessment design and analysis. Prior to joining ETS, Joanna was an Associate Professor at Arizona State University where her research focused on the application of cognitive theories and methods to the design and validation of tests of spatial reasoning, quantitative reasoning, and verbal reasoning. She received her PhD in Quantitative Psychology (minor: Cognitive Psychology) from the University of Kansas.

Qi Guo is a PhD student at the University of Alberta. His research interests include cognitive diagnostic assessment, test reliability, and structural equation modeling.

Jeffrey R. Harring is an Associate Professor of Measurement, Statistics and Evaluation in the Department of Human Development and Quantitative Methodology at the University of Maryland. Generally, his research focuses on the development and evaluation of statistical models and methods used in education, social and behavioral science research. His current research centers on methodological issues surrounding linear, generalized linear, and nonlinear latent variable models for longitudinal data. Other threads of his research focus on finite mixture models and nonlinear structural equation models.

Ari Houser is a doctoral candidate in the Measurement, Statistics and Evaluation program within the Department of Human Development and Quantitative

Methodology at the University of Maryland. He works concurrently as a Senior Methods Advisor in the AARP Public Policy Institute. His main research interests are on longitudinal models for discrete-valued latent variables.

Kristen Huff received her EdD in Measurement, Research and Evaluation Methods from the University of Massachusetts Amherst in 2003, and her MEd in Educational Research, Measurement, and Evaluation from the University of North Carolina at Greensboro in 1996. Her work focuses on ensuring the coherence of assessment design, interpretation, use, and policy to advance equity and high-quality education for all students. Currently, Kristen serves as Vice President, Research Strategy and Implementation at ACT.

G. Tanner Jackson is a Research Scientist at Educational Testing Service (ETS) in Princeton, NJ. His work focuses on innovative assessments and student process data, including the development and evaluation of conversation-based assessments (through ETS strategic initiatives) and game-based assessments (working in collaboration with academic and industry partners). Additionally, Tanner is interested in how users interact with complex systems and he leverages these environments to examine and interpret continuous and live data streams, including user interactions across time within educational environments.

Irvin R. Katz is Senior Director of the Cognitive, Accessibility, and Technology Sciences Center at Educational Testing Service (ETS) in Princeton, NJ. He received his PhD in Cognitive Psychology from Carnegie Mellon University. Throughout his 25 year career at ETS, he has conducted research at the intersection of cognitive psychology, psychometrics, and technology, such as developing methods for applying cognitive theory to the design of assessments, building cognitive models to guide interpretation of test-takers' performance, and investigating the cognitive and psychometric implications of highly interactive digital performance assessments. Irv is also a human-computer interaction practitioner with more than 30 years of experience in designing, building, and evaluating software for research, industry, and government.

Madeleine Keehner is a Managing Senior Research Scientist in the Cognitive, Accessibility, and Technology Sciences Center at Educational Testing Service (ETS) in Princeton, NJ. She received her PhD in experimental psychology from the University of Bristol and her BS degree (honors) in psychology from the University of London, Goldsmiths College. She also received a Certificate in Education from the University of Greenwich. Maddy has studied individual differences in spatial and general reasoning in medicine and the STEM disciplines. She is also interested in what we can infer from process data captured by new technologies such as interactive virtual models and simulations. Her current work focuses on understanding cognition in various domains within the NAEP program and is exploring cognitive processes related to interactive computer-based or tablet-based assessments.

Deirdre Kerr is an Associate Research Scientist in the Computational Psychometrics Research Center at Educational Testing Service (ETS) in Princeton, NJ. Her research focuses on determining methods of extracting information about student understanding and performance from low-level log data from educational video games and

simulations. Publications include Identifying Key Features of Student Performance in Educational Video Games and Simulations through Cluster Analysis, Identifying Learning Trajectories in an Educational Video Game, and Automatically Scoring Short Essays for Content.

Leanne R. Ketterlin-Geller is a Professor in Education Policy and Leadership at Southern Methodist University in Dallas, TX. Her research focuses on the development and validation of formative assessment systems in mathematics to support instructional decision making. She investigates the application of test accommodations and principles of universal design for improving accessibility of educational assessments for all students.

Jennifer L. Kobrin is Director of Institutional Research and Effectiveness at the Graduate Center, City University of New York. Her current research focuses on higher education assessment and institutional effectiveness. Her previous research focused on the promise of learning progressions for improving assessment, instruction, and teacher development. She holds a doctorate in Educational Statistics and Measurement from Rutgers University and a Masters in Educational Research, Measurement, and Evaluation from Boston College.

James Koepfler is a Senior Analytical Consultant at SAS. His areas of interest include operational diagnostic assessments, large-scale assessment implementation, IRT, vertical scaling, and applied statistics. He holds a PhD in Assessment and Measurement and a Masters in Psychological Sciences from James Madison University.

Patrick C. Kyllonen is Senior Research Director of the Center for Academic and Workforce Readiness and Success at Educational Testing Service (ETS) in Princeton, NJ. Center scientists conduct innovative research on (a) higher education assessment; (b) workforce readiness; (c) international large scale assessment (e.g., Program for International Student Assessment; PISA); and (d) twenty-first-century skills assessment, such as creativity, collaborative problem solving, and situational interviews. He received his BA from St. John's University and PhD from Stanford University and is author of *Generating Items for Cognitive Tests* (with S. Irvine, 2001); *Learning and Individual Differences* (with P. L. Ackerman & R. D. Roberts, 1999); *Extending Intelligence: Enhancement and New Constructs* (with R. Roberts and L. Stankov, 2008), and *Innovative Assessment of Collaboration* (with A. von Davier and M. Zhu, forthcoming). He is a fellow of the American Psychological Association and the American Educational Research Association, recipient of The Technical Cooperation Program Achievement Award for the "design, development, and evaluation of the Trait-Self Description (TSD) Personality Inventory," and was a coauthor of the National Academy of Sciences 2012 report, *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century.*

Emily Lai is Director of Formative Assessment and Feedback in the Efficacy and Research organization at Pearson. Emily's areas of interest include principled assessment design approaches, performance assessment, assessment for learning, and assessment of twenty-first-century competencies. Her most recent research includes co-developing

a learning progression and online performance assessments to teach and assess concepts related to geometric measurement of area. Emily holds a PhD in Educational Measurement & Statistics from the University of Iowa, a Masters in Library and Information Science from the University of Iowa, and a Masters in Political Science from Emory University.

Hollis Lai is Assistant Professor of Dentistry and the Director of Assessment for Undergraduate Medical Education program at the University of Alberta. His specialization is educational and psychological testing, with an emphasis on assessment designs in medical education, curriculum mapping, educational data mining, and item generation.

Roy Levy is an Associate Professor of Measurement and Statistical Analysis in the T. Denny Sanford School of Social & Family Dynamics at Arizona State University. His primary research interests include methodological developments and applications of psychometrics and statistical modeling in item response theory, Bayesian networks, and structural equation modeling, with applications in assessment, education, and the social sciences. He recently published *Bayesian Psychometric Modeling* (with Robert J. Mislevy).

Jacqueline P. Leighton is Professor and Chair of Educational Psychology and past Director of the Centre for Research in Applied Measurement and Evaluation (CRAME), a centre that is part of the Department she oversees at the University of Alberta. As a registered psychologist with the College of Alberta Psychologists, her research is focused on measuring the cognitive and socio-emotional processes underlying learning and assessment outcomes, including cognitive diagnostic assessment and feedback delivery and uptake. Funded by NSERC and SSHRC, she completed her graduate and postdoctoral studies at the University of Alberta, and Yale University, respectively. She has published in a variety of educational measurement journals, is past editor of *Educational Measurement: Issues and Practice*, and has published 3 books with Cambridge University Press.

Richard M. Luecht is a Professor of Educational Research Methodology at the University of North Carolina at Greensboro. His research interests include developing computer-based testing models and software, large-scale computerized assessment systems design, standard setting, innovative item design, item response theory parameter estimation, scaling linking and equating, automated test design algorithms and heuristics, and the application of design engineering principles to assessment.

Taylor Martin is an Associate Professor in Instructional Technology and Learning Sciences at Utah State University, where she is a principal investigator of the Active Learning Lab. Her work focuses on how learning, instruction, and practice come together in authentic contexts for Science, Technology, Engineering, and Mathematics education, focusing on topics ranging from how children learn fractions to how engineers refine their problem-solving skills. Her findings demonstrate that active learning strategies can improve motivation, encourage innovative thinking, and match

traditional strategies on developing core content knowledge. In addition, she employs data science methods to understand how these strategies impact important outcomes. She is currently on assignment at the National Science Foundation, focusing on a variety of efforts to understand how Big Data is impacting research in Education and across the STEM disciplines. Previously, she was at the Department of Curriculum and Instruction The University of Texas at Austin.

Robert J. Mislevy is the Frederic M. Lord Chair in Measurement and Statistics at Educational Testing Service (ETS) and Emeritus Professor at the University of Maryland. His research applies developments in technology, statistics, and cognitive science to practical problems in assessment. His work includes collaborating with Cisco Systems on simulation-based assessment of network engineering and developing an evidence-centered assessment design framework. Publications include Bayesian Networks in Educational Assessment, Bayesian Psychometric Modelling, and the Cognitive Psychology chapter in Educational Measurement.

Paul D. Nichols is a Senior Director in Research at ACT where Paul supports assessment and product design, the development of validity arguments and the use of qualitative methods. Paul's current research focuses on applying the theories and methods from the learning sciences to a broad range of activities in educational measurement. Paul holds a PhD and a Masters in educational psychologyfrom the University of Iowa.

Andreas Oranje is a Principal Research Director in the Research department of Educational Testing Service (ETS) in Princeton, NJ. He oversees various research centers focused on the development and validation of generalizable assessment capabilities including automated scoring evaluation, natural language and speech processing, dialogic and multimodal assessment, cognitive science, assessment and assistive technologies, and psychometric research related to group-score assessments. He serves as Project Director for Design, Analysis, and Reporting of the National Assessment of Educational Progress (NAEP 2013–2017). His research interests include designs for large scale (adaptive) assessments, psychometric research, and game- and scenario-based assessment.

William R. Penuel is a Professor of Learning Sciences and Human Development in the School of Education at the University of Colorado Boulder. His research focuses on the design, implementation, and evaluation of innovations in science and mathematics education. He has designed a number of innovations focused on improving classroom assessment in science and was a member of the committee that developed the consensus report, *Developing Assessments for the Next Generation Science Standards* (2014).

Hilary Persky is a Principal Assessment Designer in the Assessment Development division of Educational Testing Service (ETS) in Princeton, NJ. She has focused largely on forms of performance assessment in various subject areas, ranging from visual arts and theatre to science, writing, and most recently, reading tasks incorporating avatars. Her work is concerned with how to introduce meaningful innovation into large-scale, on-demand

assessment while retaining reliable measurement. She is also interested in ways of enriching assessment reporting with process data, in particular in the area of writing.

Amy Reilly is Director of Research Support with Pearson's Research and Innovation Network. Her previous work experience includes serving as the Pearson program manager for statewide assessment programs including Tennessee, Arkansas, and Utah and as a test development manager and lead content specialist in English/Language Arts. She also was a Texas public school teacher, specializing in reading interventions for special education students. Amy holds a BS in Interdisciplinary Studies from Texas A&M University and an MBA from St. Edwards University.

Lisa M. Rossi worked as a Research Analyst in the Educational Psychology Laboratory at Worcester Polytechnic Institute. She holds a Master's degree in Human-Computer Interaction from Georgia Institute of Technology and a Bachelor's degree in Psychological Science from Worcester Polytechnic Institute. Currently, she works as a UX Architect for State Farm in Atlanta, Georgia.

André A. Rupp is a Research Director at Educational Testing Service (ETS) in Princeton, NJ, where he works with teams that conduct comprehensive evaluation work for mature and emerging automated systems. His research has focused on applications of principled assessment design frameworks in innovative assessment contexts as well as translating the statistical complexities of diagnostic measurement models into practical guidelines for applied specialists. Through dissemination and professional development efforts he is deeply dedicated to help interdisciplinary teams navigate the complicated trade-offs between scientific, financial, educational, and political drivers of decision making in order to help shape best methodological practices.

Michael A. Sao Pedro gained his PhD under Janice Gobert's supervision while at Worcester Polytechnic Institute. He is a Co-Founder and the Chief Technology Officer of Apprendis. He specializes in the development of digital assessments for science using Educational Data Mining. Formerly, he was a Senior Software Engineer at BAE Systems (formerly ALPHATECH, Inc.). There, he led several artificial intelligence-inspired software efforts on several Phase I/II SBIR and DARPA projects.

Jason Schweid is a former classroom educator who received his EdD in Measurement, Research and Evaluation Methods in 2011 and his MEd in Counseling in 2008, both from the University of Massachusetts Amherst. His work focuses on assessment design, development, validation and education policy. Currently, Jason serves as a Fellow for Assessment at the USNY Regents Research Fund, where he advises the NY State Department of Education on assessment design and policy.

Lorrie A. Shepard is Dean and Distinguished Professor of Research and Evaluation Methodology in the School of Education at the University of Colorado Boulder. Her early research focused on test validity, contingent on the contexts of test use. Her current research focuses on classroom assessment. Drawing on cognitive research and

sociocultural theory, she examines ways that assessment can be used as an integral part of instruction to help students learn.

Valerie Shute is the Mack & Effie Campbell Tyner Endowed Professor in Education in the Department of Educational Psychology and Learning Systems at Florida State University. Her current research involves using games with stealth assessment to support learning – of cognitive and noncognitive knowledge, skills, and dispositions. Val's research has resulted in numerous grants, a patent, and publications (e.g., *Measuring and supporting learning in games: Stealth assessment*, with Matthew Ventura).

Lubin Wang is a doctoral candidate of Instructional Systems and Learning Technologies program at Florida State University. Her research interests include game-based learning and assessment. She is particularly interested in the assessment and improvement of problem-solving skills as well as identifying gaming-the-system behaviors during gameplay. She has participated in various funded research projects led by Dr. Shute, and coauthored or is coauthoring several papers and chapters with her.

Zachary Warner is a former high school math teacher who received his PhD in Educational Psychology from the University of Albany, SUNY in 2013. His research focuses on large-scale assessments and how results can best inform educational goals at school, district, and state levels. Zach has published research on using computer-based formative assessment tools and rubric-referenced student self-assessment. He currently serves as a state psychometrician for the New York State Education Department.

Mark Wilson is the Director of the Berkeley Evaluation and Assessment Research (BEAR) Centre, and a professor at the University of California, Berkeley, USA, and also is a professor in Education (Assessment) in the Assessment Research Centre at The University of Melbourne, Australia. He is an internationally recognized specialist in psychometrics and educational assessment, and is currently president of the National Council for Measurement in Education (NCME). His work spans assessment topics such as mathematics and science assessment, cognitive modeling, learning progressions, school-based assessment, and interactive online assessment of 21st century skills.

Diego Zapata-Rivera is a Senior Research Scientist at Educational Testing Service (ETS) at Princeton, NJ. His research focuses on innovations in score reporting and technology-enhanced assessment, including work on assessment-based learning environments and game-based assessments. He has published numerous articles and has been a committee member and organizer of conferences in his research areas. He is a member of the Editorial Board of the *User Modeling and User-Adapted Interaction* journal and an Associate Editor of *IEEE Transactions on Learning Technologies*.

Mo Zhang is a Research Scientist in the Research and Development Division at Educational Testing Service (ETS) in Princeton, NJ. Her research interests lie in the methodology of measurement and validation for automated and human scoring.

Foreword

"I don't give a hoot about cognitive psychology!"

This statement (but using much saltier language) was said to me by a senior colleague sometime in my first year working at Educational Testing Service (ETS). As possibly the first cognitive psychologist on staff at ETS, I expected some culture clash when I arrived in 1990, and I was not disappointed. My research training on detailed investigations of human problem solving in academic domains differed greatly in terms of methodologies, perspectives, and typical Ns (tens versus thousands) from the psychometric tradition. For example, in 1990, few of my psychometrician colleagues had heard of think-aloud studies in which students talk concurrently as they solved problems, let alone saw their value for real-world educational measurement. Sometimes I struggled to convince people that someone like me, with no formal measurement training, had something useful to contribute.

On the other hand, my assessment development colleagues showed a great interest in cognition. They wanted to know, for example, why their test questions weren't working as intended, such as being too difficult or too easy. As more than one assessment developer put it, "what were those test takers thinking?" Ah ha! Here was a clear use for cognitive psychology and think-aloud studies (sometimes called "cognitive labs"): provide insights to the people who write test items. I had found a niche, but felt dissatisfied that I had failed to bridge the gap between cognition and psychometrics.

Thankfully, in the past quarter century, psychometricians increasingly have been the bridge builders, taking up the challenge of accommodating measurement models and assessment practices to theories of cognition. No doubt growing measurement challenges pushed things along, such as the educational community's desire for assessments that reflect real-world situations, provide more detailed information than a scaled unidimensional score, and target knowledge and skills beyond those that had been traditionally investigated by educational measures.

With this *Handbook*, André and Jackie have brought together a truly multidisciplinary group: classically trained psychometricians who have developed methods to incorporate cognitive models into measurement models as well as cognitive psychologists

who have put their minds and theories to the problem of measurement. This is a special group of people, dedicated not only to rigorous science (you see their names regularly in major measurement journals), but also to bringing measurement science into real-world practice.

Given the multidisciplinary group of authors, it's no surprise that the *Handbook* should appeal to multiple audiences:

- *Psychometricians* who already use a psychometric modeling or assessment design framework approach involving cognitive models, but who want to learn about other methods and how they contrast with the ones with which they are familiar.
- *Educational measurement researchers* who are interested in techniques for assessing and scoring response data related to constructs beyond the academic disciplines that have been the traditional focus of measurement, or who have an interest in designing or utilizing assessments that gather new types of evidence, such as "clickstream" data or eye movements, of test taker knowledge and skills.
- *Cognitive psychologists* who seek to test theories of cognition at a large scale through assessment-like tasks, utilizing the advanced psychometric methods presented in the *Handbook*.
- *Assessment development practitioners* who are developing new types of assessments and want to use systematic assessment design methods or apply solid psychometric modeling techniques that can handle the complexity of evidence that is required to go beyond a unidimensional score.
- *Graduate students* who seek a comprehensive overview of the tough problems that interdisciplinary teams have to address in order to integrate models of cognition and principles for assessment design. There are many dissertation topics implied or suggested by this work!

The *Handbook of Cognition and Assessment* presents frameworks and methodologies, along with recent applications that, together, showcase how educational measurement benefits from an integration with cognition. It has been a long 25 years, but it seems that many psychometricians now *do* give a hoot about cognitive psychology. About friggin' time.

Irvin R. Katz
Educational Testing Service
April, 2016

Acknowledgements

André and Jackie would like to thank all of their authors for working so patiently and diligently with them through several revision cycles in order to ensure that the resulting *Handbook* that you are holding in your hands is as thematically coherent and consistent in style as possible – they really enjoyed learning from all of their colleagues! A huge thanks also goes out to Jennifer Petrino, who was responsible for a large number of internal organizational logistics as well as most external communication throughout the lifecycle of this project. While working under quite a bit of pressure she always remained courteous, professional, and supportive! A big thanks also has to go out to Denisha Sahadevan and Nivetha Udayakumar, who efficiently oversaw the production process at various stages, Carol Thomas, our rigorous copy-editor for the book, and Emily Corkhill, who was very supportive of a creative cover book design and listened to other product marketing input. Perhaps most importantly, they want to thank Jayne Fargnoli at Wiley Blackwell, who reached out to André several years ago to discuss the idea of a *Handbook*, encouraged both of them to make it a reality, and put her faith in their ability to deliver a high-quality product in a timely manner – André and Jackie sincerely hope that she likes the product as much as they do!

André would specifically like to thank his co-editor, Jackie Leighton, who was always been a professionally insightful, emotionally supportive, and just wonderfully fun person to do this with – he would do this again with her at any time! He would also like to express gratitude to his current and previous managers at various senior levels at ETS for supporting this project unconditionally, especially David Williamson, Andreas Oranje, Joanna Gorin, and Ida Lawrence. Furthermore, Jim Carlson and Kim Fryer at ETS were particularly respectful of the pressing timelines and helped make the internal review processes expedient and rigorous. He is also particularly appreciative of the two paintings from his very good friend Claire Sower, which she created just for this book and which were used to create the resulting cover image – he thinks the world of her and cannot wait to see where her artistic career takes her! Finally, he would also like to express his deep appreciation, gratitude, and love to his fabulous wife Brooke Sweet for being patient and supportive as always, and especially for listening to him talk repeatedly – at home, at restaurants, before shows, and in the car during road trips – about how fabulous a project this was!

Jackie extends her gratitude to partner-in-crime, André Rupp, first for the invitation to collaborate on a project that captures and extends the imagination of what is possible for assessment; and for André's vision, rigor, attention to detail, wit, sense of humor and constant thoughtfulness through a creative process that is both precise and scientific yet deeply philosophical. She has been enriched by the experience and welcomes future opportunities to learn from and work alongside André. Jackie also thanks so many of her departmental colleagues who quietly cheered her on as she co-edited this book while finishing her very (very) busy last year as department chair. Finally, Jackie thanks, as if words could even express, her love of almost 25 years, Greg Anderson, who continues to remain the only answer to her questions.

1

Introduction to Handbook

André A. Rupp and Jacqueline P. Leighton

Motivation for Handbook

The field of educational assessment is changing in several important ways at the time of this writing. Most notably, there has been a shift to embracing more complex ways of thinking about the relationship between core competencies, behaviors, and performances of learners at various developmental levels across the lifespan. These new ways of thinking have been fueled by new models of cognition that are increasingly more inclusive and accepting of correlates of basic knowledge and skill sets. In many educational assessment contexts, considerations of how cognitive, meta-cognitive, socio-cognitive, and noncognitive characteristics of individual learners affect their individual behaviors and performances – and those of teams that they are working in – are becoming increasingly common. Clearly, at a basic level, the mere conceptual consideration of such broader characteristics and their interrelationships is not intellectually new but the way in which they are nowadays explicitly articulated, operationalized, and used to drive instructional and assessment efforts is indeed something new.

Assessment of Twenty-First-Century Skills

In US policy, this trend is reflected in curricular movements such as the *Common Core* and its adoption by individual states as well as collections of states in consortia such as the *Partnership for Assessment of Readiness for College and Careers* and *Smarter Balanced*. While the degree of influence of these two particular consortia is likely to change over time, the foundational tenets and goals of the *Common Core* are less likely to vanish from our educational landscape. Importantly, *Common Core* standards articulate models of learning that are explicitly focused on the longitudinal development

The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, First Edition.
Edited by André A. Rupp and Jacqueline P. Leighton.

of learners over time across grades. Focal competencies include domain-specific knowledge, skills, and abilities as well as professional practices but also broader cross-domain competencies.

Such complex competencies are sometimes called "twenty-first-century skills" and include cognitive skills such as problem-solving, systems thinking, and argumentation skills, intrapersonal skills such as self-regulation, adaptability, and persistence, as well as interpersonal skills such as collaboration skills, leadership skills, and conflict resolution skills. Of note is the inclusion of information and communication technology skill sets, which are an integral part of the digitized life experiences of citizens in our times across the world. As a result, the kinds of intellectual and creative tasks that effective citizens need to be able to solve nowadays with digital tools are often qualitatively different in important ways from the tasks of the past. As a result, considerations of smart assessment design, delivery, scoring, and reporting have become much more complex.

On the one hand, more "traditional" assessments constructed predominantly with various selected response formats such as multiple-choice, true-false, or drag-and-drop are certainly here to stay in some form as their particular advantages in terms of efficiency of scoring, administration, and design are hard to overcome for many assessment purposes. This also implies the continued administration of such assessments in paper-and-pencil format rather than digital formats. While it clearly is possible to use tools such as tablets, smartphones, or personal computers for the delivery of innovative digital assessments, many areas of the world where education is critical do not yet have access to reliable state-of-the-art technological infrastructures at a large scale.

On the other hand, there are numerous persistent efforts all over the world to create "smarter" digital learning and assessment environments such as innovative educational games, simulations, and other forms of immersive learning and assessment experiences. Sometimes these environments do not proclaim their assessment goals up front and may perform assessment quietly "behind-the-scenes" so as to not disturb the immersive experience – an effort called "stealth assessment" by some. Since the tasks that we create for learners are lenses that allow us to learn particular things about them and tell evidence-based stories about them, we are nowadays confronted with the reality that these stories have become more complex rather than less complex. This is certainly a very healthy development since it forces assessment design teams to bring the same kinds of twenty-first-century skills to bear to the problem of assessment systems development that they want to measure and engender in the learners who eventually take such assessments.

Methodologies for Innovative Assessment

In the most innovative and immersive digital environments the nature of the data that are being collected for assessment purposes has also become much more complex. We now live in a world in which process and product data – the indicators from log files that capture response processes and the scores from work products that are submitted at certain points during activities – are often integrated or aligned to create more comprehensive narratives about learners. This has meant that specialists from the discipline

of psychometrics have to learn how to play together – in a common and integrated methodological sandbox – with specialists from disciplines such as computer science, data mining, and learning science.

Integrating disciplinary traditions. Clearly, professionals deeply trained in psychometrics have a lot to offer when it comes to measuring uncertainty or articulating evidentiary threads for validity arguments when data traces such as log files are well structured. Similarly, professionals deeply trained in more predominantly computational disciplines such as computer science or educational data mining have a lot to offer when it comes to thinking creatively through complex and less well-structured data traces. Put somewhat simplistically, while traditional psychometrics is often seen as more of a top-down architecture and confirmation enterprise, modern computational analytics is often seen as a more bottom-up architecture or exploration enterprise.

In the end, however, most assessment contexts require compromises for different kinds of design decisions and associated evidentiary argument components so that effective collaboration and cross-disciplinary fertilization is key to success for the future. This requires a lot of strategic collaboration and communication efforts since professionals trained in different fields often speak different methodological languages or, at least, different methodological dialects within the same language.

Paradoxically, we are now at a time when conceptual frameworks like assessment engineering or evidence-centered design – a framework that many authors in this *Handbook* make explicit reference to – will unfold their transformational power best, even though some of them have been around in the literature for over 20 years. None of these frameworks is a clear "how-to" recipe, however. Instead, they are conceptual tools that can be used to engender common ways of thinking about critical design decisions along with a common vocabulary that can support effective decision-making and a common perspective on how different types of evidence can be identified, accumulated, and aligned.

Integrating statistical modeling approaches. Not surprisingly perhaps, the statistical models that we nowadays have at our disposal have also changed in important ways. Arguably there has been a strong shift in the last decades toward unification of statistical models into coherent specification, estimation, and interpretation frameworks. Examples of such efforts are the work on generalized linear and nonlinear mixed models, explanatory item response theory models, and diagnostic measurement models, to name just a few. Under each of these frameworks, one can find long histories of publications that discuss individual models in terms of their relative novelties, advantages, and disadvantages. The unified frameworks that have emerged have collected all of these models under common umbrellas and thus have laid bare the deep-structure similarities across these seemingly loosely connected models.

This has significantly restructured thinking around these models and has helped tremendously to scale back unwarranted, and rather naïve, claims from earlier times about the educational impact that certain kinds of statistical models could have by themselves. Put differently, it has helped many quantitative methodologists to re-appreciate the fact that any model, no matter how elegantly it is specified or

estimated, is, in the end, just a technological tool. Like any tool, it can be used very thoughtfully as a "healthy connective tissue" for evidence or rather inappropriately leading to serious evidentiary "injuries."

Integrating assessment design and validity argumentation. From a validity perspective, which is foundational for all educational assessment arguments, the constellation of design choices within an assessment life cycle has to be based on sound scientific reasoning and has to rhetorically cohere to provide added value to key stakeholders. This typically means that the information that is provided from such assessments should provide real insight into learning, performance, and various factors that affect these.

As such, smart assessment design considers the system into which the assessment is embedded just as much as the tool itself. In fact, under views of the importance of measuring learning over time as articulated in the *Common Core*, for instance, it is impossible to think of the diagnostic process as a one-off event. Instead, assessment information needs to be interpreted, actions need to be taken, experiences need to be shaped, and new information needs to be collected in an ever-continuing cycle of learning, assessment, and development. In this new world of cognition and assessment such a longitudinal view will become more and more prevalent thus forcing many communities of practice to change the way they design, deliver, score, report, and use assessments.

This perspective critically affects the societal reverberations that assessments can have when serving underrepresented or disadvantaged groups in order to improve the life experiences of all learners across the societal spectrum and lifespan. It may certainly be too much to ask of measurement specialists – or at least it may be rather impractical for workflow considerations – to always keep the bigger philanthropic goals of assessments in mind as these do not always influence their work directly. For example, the optimal estimation of a complex latent variable model for nested data structures will not be directly affected by an understanding of whether this model is used in an assessment context where assessment scores are used to provide increased access to higher-education institutions for minorities or in an educational survey context where they are used for accountability judgments.

However, ensuring that assessment arguments are thoughtful, differentiated, and responsible in light of societal missions of assessment is important, especially in interdisciplinary teams that are charged with various critical design decisions throughout the assessment lifecycle. It will help these teams be more motivated to keep track of controversial design decisions, limitations of assessment inferences, and critical assumptions. In short, it will help them to make sure they know what evidence they already have and what evidence still needs to be collected in order to support responsible interpretation and decision making. As mentioned earlier, such a shared understanding, perspective, and communal responsibility can be fostered by frameworks such as assessment engineering or evidence-centered design.

Integrating professional development and practical workflows. These last points speak to an aspect of assessment work that is often overlooked – or at least not taken as seriously as it could – which is the professional development of specialists who have to work in interdisciplinary teams. There is still a notable gap in the way universities train graduate students with Master's or PhD degrees in the practices of assessment design, deployment, and use. Similarly, many assessment companies or start-ups are under immense

business pressures to produce many "smart" solutions with interdisciplinary teams under tight deadlines that take away critical reflection times.

In the world of *Common Core*, for example, short turnaround times for contracts from individual states or other stakeholders in which clients are sometimes asked to propose very complex design solutions in very short times can be problematic for these reflection processes. While short turnaround times would be feasible if the needed products and solutions truly fit a plug-and-play approach, the truth is that the new assessment foci on more complex, authentic, collaborative, and digitally delivered assessment tasks require rather creative mindsets. They also require new modes of working that go from a simple design-and-deploy approach, interspersed with one or two pilot studies and a field trial, to a much more consistent design-deploy-evaluate-revise lifecycle with shorter and more frequent bursts of activity, at least for formative assessments. These mindsets require time to cultivate and established processes require time to change, which is again why frameworks like assessment engineering and evidence-centered design can be so powerful for engendering best practices.

Handbook Structure

In the context of all of these developments it became clear to us that it would not be possible to create a single *Handbook* that would be able to cover all nuances of assessment and cognition, as conceived broadly, in a comprehensive manner. Instead, what we have strived to do is to provide a reasonably illustrative crosswalk of the overall landscape sketched in this brief introduction. We did so with an eye toward taking stock of some of the best practices of the current times while setting the stage for future-oriented ways of rethinking those best practices to remain cutting-edge. After some back-and-forth we eventually decided to divide this *Handbook* into three core parts even though readers will find a lot of cross-part references as many ideas are clearly interrelated. For simplicity of communication, we decided to label these three parts *Frameworks*, *Methodologies*, and *Applications*.

Frameworks

In the *Frameworks* section we invited authors to articulate broader ways of thinking around what models of cognition might offer in terms of the psychological infrastructure that sustain frameworks for assessment design, delivery, scoring, and decision making along with associated validation practices. This part, in many ways, is a conceptual cornerstone for any and all types of assessments that are primarily developed with the intention to support claims about the unobservable information processes, knowledge, and skills that accompany observed performance. The nine chapters in this part present distinct but overlapping perspectives on how models of cognition can inform – both conceptually and practically – the design and developments of assessments from start to finish.

In Chapter 2 on the role of theories of learning and cognition for assessment design and development, Nichols, Kobrin, Lai, and Koepfler present a framework and three criteria for evaluating how well theories of learning and cognition inform design and

decisions in principled assessment design, assessment engineering, and evidence-centered design. In Chapter 3 on cognition in score interpretation and use, Ferrara, Lai, Reilly, and Nichols further analyze the elements that define principled approaches to assessment design, development, and implementation before comparing and illustrating the use of different approaches. In Chapter 4 on methods and tools for developing and validating cognitive models in assessment, Keehner, Gorin, Feng, and Katz focus us on ways to characterize cognitive models, including the rationale for their development and the evidence required for validation so as to ensure their utility for meeting assessment goals. This includes clearly defined assessment targets, a statement of intended score interpretations and uses, models of cognition, aligned measurement models and reporting scales, and manipulation of assessment activities to align with assessment targets, all within a backdrop of ongoing accumulation and synthesis of evidence to support claims and validity arguments.

In Chapter 5 on an integrative framework for construct validity, Embretson illustrates how a cognitive psychological foundation for item design and development can not only influence reliability but also the five aspects of an integrated construct validity framework with special attention on how automatic item generators are supported within the context of the framework. Further expanding on this idea, in Chapter 6 on cognitive models in automatic item generation, Gierl and Lai similarly show us how cognitive item models can be operationalized to guide automatic item design and development to measure specific skills in the domains of science and medicine.

In Chapter 7 on social models of learning and assessment, Penuel and Shepard analyze ways in which research teams articulate the vertices of the "assessment triangle." This includes representations of how students become proficient in the domain, the kinds of activities used to prompt students to do or say things to demonstrate proficiency, and frameworks for making sense of students' contributions in these activities in ways that can inform teaching. In Chapter 8 on socio-emotional and self-management variables in assessment, Kyllonen explains the importance of noncognitive skills as predictors of cognitive skills development and as outcomes for which assessments should be developed for their own sake. In chapter 9 on the role of cognitively-grounded assessment practices in understanding and improving accessibility for special populations, Ketterlin-Geller outlines the ways in which educational assessments can be enhanced in their design and development to be accessible to students in special populations. Finally, in Chapter 10 on integrated perspectives of validation and automated scoring, Bejar, Mislevy, and Zhang discuss the various design decisions that have to made during the lifecycle of automated systems for scoring and feedback. They specifically discuss the history of certain key systems across a wide variety of domains with applications that span short and extended written responses, spoken responses, responses with multimodal outputs, and interactive response processes within virtual learning environments.

Methodologies

In the *Methodologies* section we asked authors to present statistical modeling approaches that illustrate how information about cognitive processes can be operationalized and utilized within the context of statistical models. One potential conceptual

dimension to draw between modeling approaches is that of parametric versus nonparametric modeling approaches. The former are generally characterized by explicit functional forms, which include parameters that can be interpreted, strong assumptions that are made about distributions of component variables for estimation, and a variety of computational approaches for obtaining parameter estimates given suitable data. These models allow for the power of formal statistical inference around these parameters so that interpretations about cognitive processes or behaviors in the population can be made with the sample data. This particular quantification of statistical uncertainty is unique to parametric models even though there are other ways of quantifying uncertainty in nonparametric approaches. Moreover, parametric models allow for an explicit assessment of model-data fit using the parameters in the model and can be used efficiently for applications that require modularity and component-based information such as computer-adaptive (diagnostic) assessment, automated item generation, automated form assembly, and the like.

Nonparametric approaches are generally characterized by weaker distributional assumptions and use either probabilistic or rule-based decision sequences to create data summaries. While the focus of inference may be similar as with parametric models, the kind of information obtained from these models and the way that one can reason with that information is thus structurally distinct. For example, diagnostic measurement models and clustering approaches can both be used to sort learners into unobserved groups. However, in the former parametric approach one obtains parameters that can be used explicitly to characterize the learners and the tasks that they were given. In the latter nonparametric approach, such characterizations have to be made through various secondary analyses without explicit model parameters as guideposts.

The formalism of parametric models is certainly important whenever assessments are administered at larger scales and when decisions take on a more summative nature, perhaps for state-wide, regional, national, or international accountability purposes. However, the power of parametric models can sometimes also be useful in more formative decision-making contexts such as digital learning and assessment environments that require certain kinds of automation of evidence identification and accumulation procedures. Consequently, the six chapters in Part II of the *Handbook* are skewed more toward the parametric space overall, which is arguably appropriate given how powerful and important this model space is for educational assessment.

In Chapter 11 on explanatory item response theory models, De Boeck, Cho, and Wilson discuss how to specify, estimate, and reason within a unified latent-variable modeling framework called explanatory item response theory. The general idea is that this framework subsumes simpler modeling approaches from item response theory, which are the current state-of-the-art for data modeling in large-scale assessment. However, they expand upon these foundations by allowing for the inclusion of additional variables – called covariates – for learners, tasks, or learner-task combinations that may help to "explain" observed performance differences. As with any statistical methodology, the degree to which such explanations are robust and defensible more broadly based on scientific grounds requires additional validation studies. In Chapter 12 on longitudinal latent-variable models for repeated measures data, Harring and Houser discuss how to specify, estimate, and reason within another unified latent-variable modeling framework that focuses on the modeling of data collected over time

or other conditions of replication. They describe how seemingly complicated design choices in mathematical structures of certain model components can be – and have to be – grounded in an understanding about cognitive processes in order to make interpretations defensible. As with explanatory item response theory models, this framework allows for the inclusion of various covariates at the learner, task, or occasion level with similar evidentiary requirements for thorough validation of interpretations.

In Chapter 13 on diagnostic classification models, Bradshaw discusses how to specify, estimate, and reason with yet another unified latent-variable modeling framework called the log-linear cognitive diagnosis model. The general idea here is that an a priori specification of how different tasks measure different skill sets can be used to create classifications of learners into different competency states that are describable through these skill sets. Just as in the other two chapters discussed previously, covariates at different levels can be included into these models for additional explanatory power. In Chapter 14 on Bayesian networks, González-Brenes, Behrens, Mislevy, Levy, and DiCerbo describe how to specify, estimate, and reason with a family of latent-variable models that share many similarities, but also display critical differences, with diagnostic classification models. Similar to the latter models, these models require an a priori specification of relationships between skill sets and tasks, which can be refined through model-data fit evaluations. However, in contrast to those models, all the variables in this approach are categorical, the specification of relationships between variables can accommodate a large number of dependencies relatively easily, and the estimation is very general and well aligned with conceptual understandings of how human beings reason more generally.

In Chapter 15 on the rule-space methodology and the attribute hierarchy method, Cui, Gierl, and Guo describe a predominantly nonparametric alternative to diagnostic classification models and Bayesian networks. Specifically, their two methods represent historical foundations for the parametric approaches and remain attractive alternatives in situations where the full power of parametric inference is not needed. Both methods are used predominantly for classifying learners, with less of an emphasis on obtaining detailed characterizations of tasks or explanatory narratives through additional covariates, at least not within a single estimation run. Finally, in Chapter 16 on educational data mining and learning analytics, Baker, Martin, and Rossi provide an overview of the utility of a variety of statistical analysis techniques in the service of performing cognitively grounded data mining work for assessment purposes. They illustrate this work through applications in innovative digital learning environments where a wide variety of behavior detectors have been used to characterize learner actions and to make inferences about underlying cognitive skill sets and meta-cognitive factors that affect performance. This last chapter serves as somewhat of a conceptual bridge between the *Methodologies* and the *Applications* parts of the *Handbook* as the latter part contains more such innovative applications along with slightly more traditional ones.

The six chapters in this section clearly do not cover the entire space of psychometric or computational techniques that could conceivably be brought to bear to model observable learner behavior and task performance in order to make inferences about certain cognitive correlates. Entire books have been written about each of the modeling approaches, both within disciplines and across disciplines, which make any claim to a truly comprehensive coverage prohibitive. For example, we could have included

chapters on structural equation models or traditional item response theory models as well as chapters on other nonparametric clustering techniques or multivariate analysis methods.

However, it was not our goal to develop yet another methodological *Handbook* that is oriented primarily toward specialists whose day-to-day job is to make smart decisions about data analysis. Instead, we wanted to create a meaningful cross-section of this broad methodological space in a way that gives explicit room for arguments about how to specify, estimate, and, most importantly, reason with these models. We made strong efforts to work with the authors to keep the chapters in a rather accessible language, structure, and level of detail so that specialists who do not think about statistical models on a daily basis would be able to learn a few meaningful and actionable pieces of information about these methodologies from the chapters. It is our firm belief that even a tentative understanding and an associated thirst to learn more about the strengths and limitations of different modeling approaches can go a long way toward fostering this shared methodological and evidentiary reasoning understanding that we have talked about at the outset.

Applications

In the *Applications* section we asked authors to traverse an equally diverse space of possible uses of models for cognition in the service of a broad range of assessment applications. For example, we decided to select a few very common assessment applications and encouraged the authors of the seven chapters in this part to describe both the broader contexts and frameworks within which their illustrations are embedded and to be forward-thinking in their description. That is, rather than asking them to merely describe the state of the world as it is now we explicitly wanted them to take some intellectual chances and speculate on what some key trends for their areas of work would be.

In Chapter 17 on large-scale standards-based summative assessments, Huff, Warner, and Schweid discuss how thinking about cognition influences the design and use of these kinds of assessments. They use three powerful examples across different use contexts to show surface-level differences and deep-structure similarities across these contexts using a recent framework for differentiating between cognitive models. Using these examples, they articulate how certain kinds of articulations and operationalizations of cognition are necessary to increase the inferential power of these assessments and how others can be quite harmful to this process as they are somewhat unrealistic – or poorly matched – in this context. In Chapter 18 on large-scale educational surveys, Oranje, Keehner, Persky, Cayton-Hodges, and Feng discuss the general aims of these kinds of assessments, which is accountability at state or country levels, and illustrate the current innovation horizon in this area through examples from an interactive national assessment in the United States. They demonstrate that historical notions of item type restrictions are only partly transferrable for the future of this line of work, and that more complex interactive assessment tasks are the generative framework that should be utilized to measure at least some twenty-first-century skill sets reliably at this level of assessment.

In Chapter 19 on professional certification and licensure examinations, Luecht provides practical examples to show why assessment engineering design components

and procedures, including task modeling, task design templates, and strong statistical quality control mechanisms, are an integral and important part of the many processes for developing cognitively based formal test specifications, building item banks, and assembling test forms that optimize professional knowledge assessment and/or skill mastery decisions. In Chapter 20 on the in-task assessment framework for in-task behavior, Kerr, Andrews, and Mislevy describe an articulation of the evidence-centered design framework within digital learning and assessment environments specifically. They describe a set of graphical tools and associated evidentiary reasoning processes that allow designers of such environments to make explicit the different steps for operationalizing construct definitions for complex skill sets. These tools then help to link observable behaviors captured in log files to different construct components to derive useful feedback and scores that are based on an explicit chain of evidence, a process that they illustrate with three examples from different domains.

In Chapter 21, on digital assessment environments for scientific inquiry skills, Gobert and Sao Pedro provide yet another application of cognitively inspired assessment – in this case, it is the design, data-collection, and data-analysis efforts for a student-based digital learning and assessment environment devoted to scientific inquiry and practices. In Chapter 22, on stealth assessment in educational video games, Shute and Wang look at how both commercial games and games designed or adapted for assessment purposes can be powerful levers for measuring twenty-first-century skills. They describe how evidence-centered design thinking coupled with systematic synthesis of the current cognitive literature on these skill sets are necessary prerequisites for instantiating best evidentiary reasoning practices through embedded assessment in these contexts. In Chapter 23 on conversation-based assessment, Jackson and Zapata-Rivera introduce us to the benefits of these kinds of assessment for collecting new types of explanatory evidence that potentially afford greater insight into test taker cognition and metacognition. They further propose a new framework to properly situate and compare conversation-based assessments with other kinds of assessment items and illustrate the power of conversation-based assessment through a prototype. Finally, the Handbook contains a glossary with definitions of key terms that are used across chapters. In each chapter, the first mention of any key term in the glossary is boldfaced for easy reference.

Closing Words

As this brief overview has underscored, the *Handbook* that you are holding in front of you is a complex labor of love that involved the participation of many wonderful members of scientific communities engaged in some type of educational assessment activity. These activities span the design of large-scale educational surveys, the development of formative learning systems, the evaluation of novel statistical methods that support inferences, and the conceptual articulation of frameworks that guide best practices, to name a few. We are infinitely grateful for all of our colleagues who have worked patiently with us to create our particular conceptual crosswalk of this landscape. We sincerely hope that the final product will be as much appealing to them as it is to us.

Most importantly, however, we sincerely hope that readers will find this *Handbook* powerful for changing the ways they think about the interplay of assessment and cognition. We hope that reading individual chapters, parts, or maybe even the entire

book will stimulate new ideas, new ways of thinking, a thirst for wanting to learn more from references that are cited, and a deep continued passion for improving the lives of learners across the world through thoughtful and innovative assessment design, development, deployment, and use. If we were to make even small but meaningful contributions to these efforts we would be eternally grateful.

Sincerely,
André A. Rupp and Jacqueline P. Leighton

Part I

Frameworks

2

The Role of Theories of Learning and Cognition in Assessment Design and Development

Paul D. Nichols, Jennifer L. Kobrin, Emily Lai, and James Koepfler

Assessment planning includes both design and development. Design emphasizes the formulation of a sequence of assessment development actions aimed at accomplishing specific goals (e.g., intended consequences of score use or desired levels of psychometric properties). Development emphasizes the execution of the planned course of action. Both assessment design and development involve numerous, interconnected decisions that should address the three elements described as the **assessment triangle** (Pellegrino, Chudowsky, & Glaser, 2001): a *theory* or set of beliefs about how students think and develop competence in a domain (*Cognition*), the *content* used to elicit evidence about those aspects of learning and cognition (*Observation*), and the *methods* used to analyze and make inferences from the evidence (*Interpretation*). The **targets of inference** for an assessment are the aspects of learning and cognition, typically a subset of a **theory of learning and cognition**, that are intended to be assessed.

Pellegrino et al. (2001) cautioned that the three elements comprising the assessment triangle must be explicitly connected and coordinated during assessment design and development or the **validity** of the inferences drawn from the assessment results will be compromised. We use the label "coherent" to refer to assessment design and development processes in which the three elements of the assessment triangle are connected and coordinated. Adapting the notion of *system coherence* described by the *National Research Council* (NRC, 2012), we distinguish between **horizontal coherence** and **developmental coherence**. Horizontal coherence is created when all the components of assessment design and development are connected and coordinated with the theories of learning and cognition in which the targets of inference are embedded. Developmental coherence is created when this coordination of assessment components with theories of learning and cognition is maintained across time as design and development activities unfold. Arguments for coherence are specific to a given interpretation of assessment performance. Coherence is argued for based on rationales and backing,

The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, First Edition.
Edited by André A. Rupp and Jacqueline P. Leighton.

supporting claims that the targets of inference, observation and interpretation are aligned horizontally and vertically. The development of this argument should commence with assessment design and continue through any modifications following assessment launch.

Maintaining coherence across the teams of professionals involved in different activities often unfolding simultaneously requires an assessment design and development approach that explicitly coordinates the targets of inference, observation, and interpretation. In this chapter, we offer **principled assessment design** (PAD) as an approach that fosters such coherence. PAD is a family of related approaches including **cognitive design systems** (Embretson, 1998), **evidence-centered design** (ECD) (Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003), **principled design for efficacy** (PDE) (Nichols, Ferrara, & Lai, 2014), and **assessment engineering** (AE) (Luecht, 2013). PAD fosters coherence by compelling assessment designers to justify, based on the target of inference definition, the chain of decisions relative to the other two elements of the assessment triangle (i.e., the content used to elicit evidence about those aspects of cognition and the methods used to analyze and interpret the evidence). Under PAD, the three elements of the assessment triangle are more likely to be explicitly connected and coordinated during assessment design and development.

Definitions of the targets of inference are often derived from and embedded within theories of learning and cognition. Researchers within specific domains have studied for many years the different types of **knowledge, skills, and abilities** (KSAs) that are often the targets of inference. Contemporary research in the learning sciences offers a number of different perspectives on how people learn knowledge and skills and use them in thinking and problem solving. These different perspectives emphasize different aspects of learning, thinking, and acting and have different implications for what should be assessed and how (Mislevy, 2006; Pellegrino et al., 2001).

In this chapter, we present and illustrate criteria for evaluating the extent to which theories of learning and cognition and the associated research support coherence among the three vertices of the assessment triangle when used within a PAD approach. The theories of learning and cognition found in this chapter are not the kind of broad, "exceptionless" generalizations from physics often represented as the ideal image. Theories from the social sciences tend to be exception-rich and highly contingent (see Mitchell, 2009). The theories referred to in this chapter describing learning and cognition in mathematics fit within the conceptual framework of **learning trajectories** (Daro, Mosher, & Corcoran, 2011). That being said, learning trajectories are certainly not the only conceptual frameworks available to inform assessment design and development.

We have divided this chapter into four sections. In the first section, we describe criteria recent writers have offered for evaluating the usefulness of theories of learning and cognition for informing assessment design and development decisions. In the second section, we summarize PAD and then use PAD as a lens through which to evaluate how well different theories of learning and cognition might support assessment design and development. In the third section, using PAD as a lens, we then illustrate the evaluation of a theory of learning and cognition, represented by a **learning progression** (LP) on "geometric measurement of area". Finally, in the fourth section, we summarize the implications of these decisions for constructing an argument for the validity of the interpretation and use of assessment results.

A Brief History of Evaluation Criteria for Theories of Learning and Cognition

A number of past writers have prescribed the use of theories of learning and cognition to inform assessment design and development. For example, Loevinger (1957) identified implications of assessment design and development decisions with respect to the targets of inference and validity. Glaser, Lesgold, and Lajoie (1987) called for a cognitive theory of measurement in which the measurement of achievement would be based on our knowledge of learning and the acquisition of competence. Lohman and Ippel (1993) described a cognitive diagnostic framework for creating assessments that took advantage of research by Gitomer and colleagues (Gitomer, Curtis, Glaser, & Lensky, 1987) on verbal analogies, research by Lewis and Mayer (1987) on mathematical problem solving, and research by others on identifying test-item features that could be manipulated to vary cognitive complexity and item difficulty (see Snow & Lohman, 1989, for a summary).

In step with these **construct-centered** assessment design and development approaches, recent writers have offered criteria for evaluating the usefulness of theories of learning and cognition used for informing assessment design and development decisions. In this section, we review the criteria that have been offered by Nichols (1994), Pellegrino et al. (2001), and Leighton and Gierl (2007, 2011) as a foundation from which to propose extended criteria later in the chapter.

According to Nichols (1994), theories of learning and cognition that are well suited to informing assessment design and development should include two elements that together constitute the **construct representation** for an assessment (Embretson, 1983). First, the theory should describe the KSAs related to the target of inference for the assessment, which should include how the KSAs develop and how more competent test takers differ from less competent test takers.

Second, the theory should identify the task or **item** features that are hypothesized to influence domain-specific cognition. As an example of such item features, Nichols (1994) cited mixed fraction subtraction problems in which the numerator of the first fraction must be less than the numerator of the second fraction and one must have a denominator not equal to 10 (Tatsuoka, 1990) (e.g., 2 1/5 – 2/5 = ?). Items with such features elicit evidence from seventh- and eighth-grade test takers of the common misconception that they must reduce the whole number by 1 and add 10 to the first numerator (e.g., 2 1/5 – 2/5 = 1).

A second set of criteria for identifying theories of learning and cognition that are likely to be useful for authoring assessments was offered by Pellegrino et al. (2001; also cited in Leighton & Gierl, 2011). Pellegrino et al. (2001) described the following five criteria for a theory of learning and cognition to effectively inform assessment design and development. The theory should:

- be based on empirical research in the assessment domain,
- identify performances that differentiate more from less accomplished learners in the domain,
- address differences in the way students learn and perform in a domain,
- address at least the targets of inference for an assessment; and

- support the aggregation of evidence to be useful for different assessment purposes, for example, a pre-unit formative assessment or an end-of-year summary assessment.

However, Leighton and Gierl (2007, 2011) argued that the criteria offered by Pellegrino et al. (2001) were too restrictive. They noted that few, if any, large-scale educational assessments were designed and developed using a theory of cognition and learning that met these five criteria. In response, Leighton and Gierl (2007, 2011) proposed the following three less restrictive and more general criteria that might offer more practical guidance in identifying theories of learning and cognition that may be useful for authoring assessments:

- The KSAs described in the theory should have the depth and breadth to support the design and development of an assessment for a given purpose. For example, the model may be broad but not deep to support the development of an end-of-year summary assessment, covering KSAs only at a coarse level.
- The theory must describe learning and cognition in a way that allows assessment designers to develop tasks or items to assess learning and cognition with the constraints of test administration. Currently, test developers depend on content experts' judgments of the how to manipulate test content to influence the test taker cognition elicited by tasks or items. However, Leighton and Gierl (2011) noted that little evidence exists linking items and tasks to the KSAs assumed to be elicited during testing.
- The KSAs described in the theory should be instructionally relevant and meaningful to a broad group of educational stakeholders (e.g., students, parents, teachers, and policymakers). Since assessments are part of a larger, complex system of instruction, assessment, and learning, the link between the theory and instruction can be established by having the theory address the KSAs described in the curriculum.

These three connected sets of past criteria for evaluating the usefulness of theories of learning and cognition for informing assessment design and development decisions by Nichols (1994), Pellegrino et al. (2001), and Leighton and Gierl (2007, 2011) have offered a starting point for linking these theories to assessment practices. At the same time, they included no commitment to a specific approach to assessment design and development. That is, the means through which a theory of learning and cognition is expressed in assessment design and development have been left to the reader.

In the next section, we choose PAD as our assessment design and development approach because, as we have argued above, PAD is more likely than conventional assessment design and development to support coherence. Given a PAD stance, we then describe criteria for evaluating the usefulness of theories of learning and cognition for informing assessment design and development decisions.

Principled Assessment Design as an Evaluative Lens for Theories of Learning and Cognition

PAD approaches provide frameworks for carrying out assessment design and development according to principles rooted in empirical research. The use of the term "principled" is not meant to imply that other approaches to assessment design and

development are "unprincipled" in comparison, but to emphasize that in PAD the principles take center stage in terms of the design process and the outcomes of that process. In this section, we offer a brief summary of three common characteristics of PAD approaches as they are distinguished from more conventional assessment design and development along with means by which these are accomplished in practice.

Characteristic 1: Construct-Centered Approach

For our purposes, the first – and perhaps most important – common characteristic across PAD approaches is the explicit construct-centered nature of all of these approaches (Messick, 1992). Under a construct-centered approach, assessment design begins with a careful and comprehensive examination of the constructs intended to be assessed – the targets of inference – and all subsequent design decisions cascade from that initial definition of the construct. The targets of inference are represented by the *Cognition* vertex of the assessment triangle and lay the foundation for the other two vertices, *Observation* and *Interpretation*. Assessment designers are compelled to justify, based on the definition of the target of inference, the chain of decisions necessary to implement an assessment program.

In contrast, conventional assessment design and development may be characterized as a kind of technology (Gordon, 2012) that routinizes decision making. Routinized design is characterized by the adoption of a fixed design solution. This design solution may be characterized as "best practice" and offers some efficiencies for the test developer but fails to consider the goals of the assessment program and the needs of the stakeholders. Research on design suggests that this kind of approach is based on *routinized thinking* (i.e., on the automatic use of *chunks*, which enable individuals to save mental effort; Laird, Newell, & Rosembloom, 1987; Newell, 1990), but, once implemented, the assessment designer is no longer searching for better design solutions and new ways of doing things.

Current routines used in making assessment design and development decisions evolved over decades to deal with conventional formats such as **multiple choice** items and essay writing **prompts** for large-scale projects. These guidelines and rules-of-thumb have in the past produced tests that tended to satisfy technical requirements. But the development of conventional routines could not anticipate the needs of current projects requiring efficient content creation for novel contexts that achieves predictable content difficulty and cognitive complexity targets on reduced schedules.

Characteristic 2: Engineering towards Intended Interpretations and Uses

A second characteristic common across PAD approaches is the intent to engineer intended interpretations and uses of assessment results through assessment design and development. For some, conventional item and task construction is viewed as an art (Millman & Greene, 1989) – an arcane process conducted by skilled item writers. In contrast, engineering applies scientific findings and mathematical tools to solve problems in a practical manner. PAD applies findings from the learning sciences along with measurement models under appropriate assumptions in attempting

to engineer the collection of evidence supporting probabilistic claims about the targets of inference with respect to the purpose for assessing.

Explicit manipulation of item or task features. A first means through which PAD attempts to engineer intended interpretations and uses is through the explicit manipulation of the *Observation* vertex of the assessment triangle and the features of content that have been identified as effectively eliciting evidence of status with regard to the targets of inference. Borrowing concepts from ECD (Mislevy & Haertel, 2006), these content features may be identified as either characteristic or variable. *Characteristic content features* are those that all content assessing the targets of inference should possess in some form because they are central to evoking or eliciting evidence about the targets of inference. *Variable content features* are features that can be manipulated to change the cognitive demand or complexity that the content elicits with respect to the targets of inference.

Information on the important content features for eliciting evidence of learners' status with respect to the targets of inference can be found in the studies associated with the theory of learning and cognition. These studies often provide rich descriptions of the items and tasks researchers have used and link features of these items and tasks to elicitation of evidence with respect to the targets of inference. The greater the breadth and depth of empirical studies that link features of these items and tasks to elicitation of evidence with respect to the targets of inference, the stronger the support that these content features are qualified to inform assessment design and development decisions.

Reliance on theories of learning and cognition. A second means through which PAD attempts to engineer intended interpretations and uses is through an analysis of the theory of learning and cognition, research associated with the theory, and the features of learners' performances that have been identified as evidence of status with regard to the targets of inference during that research. Again borrowing from ECD (Mislevy & Haertel, 2006), instructions for interpreting performance consist of three parts: **work product specifications**, **evidence rules**, and the **statistical model**. Specifically, work product specifications describe the structure and format of the performance that will be captured, evidence rules describe how to code the work product as in the use of a rubric, and the statistical model describes how the coding of the responses will be aggregated to make probabilistic inferences about what students know and can do. The psychometric methods commonly used to analyze and make inferences from the evidence provided by performance (e.g., **item response theory**, **structural equation models** and **cognitive diagnostic models**) are examples of statistical models (various chapters in Part II of this *Handbook* have relevant overviews).

As was the case for content features, information linking performance features to evidence for learners' status with respect to the targets of inference can be found in research studies associated with the theory of learning and cognition. The important features of performance that researchers have used as evidence for learners' status with respect to the targets of inference can be extracted from the rich descriptions often found in these studies.

Characteristic 3: Explicit Design Decisions and Rationales

The intent to engineer intended interpretations and uses of assessment results is accompanied by a concern with making all design decisions and the rationales for them explicit and transparent and collecting documentation to support them. These design decisions include a finer grained definition of the targets of inference in terms of **cognitive processes**, knowledge structures, **strategies** and mental models; the features of stimuli and items that tend to effectively elicit use of those targets of inference; the features of test-taker responses that are evidence of achievement with regard to the targets of inference and how those responses should be evaluated and aggregated to support those inferences. Theories of learning and cognition, along with relevant empirical evidence supporting those theories and models, inform those decisions and correspondingly offer support for the interpretation and use of assessment results.

A common way PAD approaches gain efficiencies and support engineering-intended interpretations and uses of assessment results is through reusable tools such as **design patterns** and **task models**. These reusable tools support both more controlled creation of assessment content as well as documentation of design decisions. As such, the reusable tools both enhance and document the **validity argument**.

Evidentiary Coherence to Enhance Validity

Given these characteristics common across PAD approaches, we argue that PAD is more likely than conventional assessment design and development approaches to foster coherence; we return to this point in more detail in the third section of this chapter. As shown in Figure 2.1, coherence is supported when all design and development decisions cascade from an initial definition of the construct, represented as targets of inference embedded within a theory of learning and cognition.

The ability to create a coherent assessment system rests heavily on the nature of the theory of learning and cognition. The theory guides choices relevant to the *Interpretation*

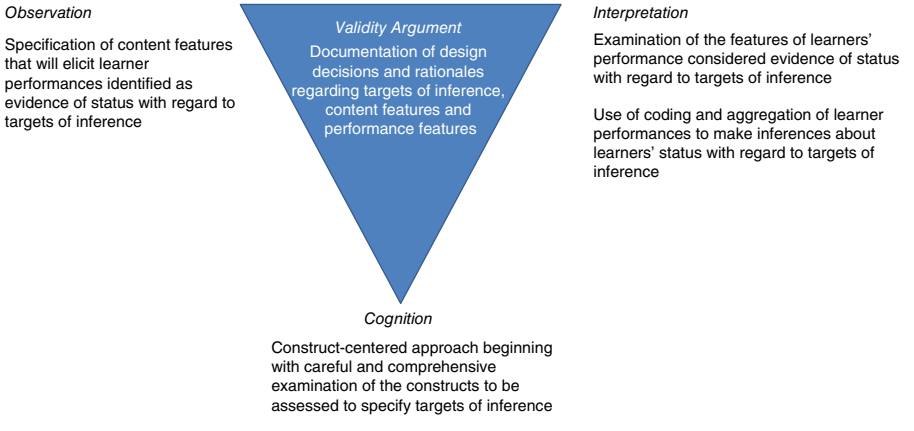

Figure 2.1 Using principled assessment design to foster coherence in the assessment triangle and support of a validity argument.

vertex of the assessment triangle, in terms of the work product specifications, evidence rules, and statistical models. The theory also guides choices relevant to the *Observation* vertex, in terms of characteristic and variable content features used to elicit performances that will serve as evidence of status with respect to the target of inference. Thus, criteria are needed to guide selection of a theory of learning to support such coherence.

Principled Assessment Design Evaluation Criteria for Theories of Learning and Cognition

The need for evidentiary coherence across the three elements in the assessment triangle (i.e., *Cognition Observation*, and *Interpretation*) has motivated us to assume a PAD perspective when creating criteria for evaluating the fitness of theories of learning and cognition. The lens of PAD influences our view on the nature of theories of learning and cognition that are likely to support decisions about assessment design and development and consequently encourage coherence among the vertices of the assessment triangle. In this section, we propose a set of three such criteria given the adoption of a PAD approach. First, we describe an LP for the "geometric measurement of area" to illustrate the application of these criteria. We then explicate each of the three criteria and illustrate their application using the LP.

Example of a Theory of Learning and Cognition

The conceptual framework of LPs has emerged from contemporary learning theories. LPs have been defined as "descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (NRC, 2007, p. 219). As suggested above, the mathematics field commonly uses the term learning trajectories to describe a similar concept.

LPs typically describe qualitatively different levels or stages that students go through in the course of their learning as their thinking becomes increasingly more sophisticated. As a perspective, most LPs assume that learners will use their knowledge at a particular level to reason about phenomena and/or solve problems in a variety of different contexts. While it is recognized that there may be some variability and that an individual may regress to a lower level of sophistication when confronted with a difficult or challenging problem, LPs largely assume that individuals' thinking is internally consistent and theory-like, and is applied somewhat consistently (Steedle & Shavelson, 2009). This assumption of consistency across contexts is necessary to diagnose a learner as being at a particular level of the LP and has strong implications for the ways in which we assess the learner and the methods used to make inferences from assessment results.

An example that we discuss in the following is the LP of "geometric measurement of area" (Lai et al., 2015) that was constructed to inform the iterative design and development of the *Insight Learning System*, which targets third grade students' understanding of ideas and concepts related to "geometric measurement of area". The system consists of a digital game, a set of online performance tasks, several instructional modules and classroom activities, and professional development experiences for teachers.

The target understanding in the LP is the conceptual understanding of the formula for area (i.e., area = length × width) and coordination of perimeter and area measurements. Lai, Kobrin, Holland, and Nichols (2015) used the findings from a number of separate studies that focus on different pieces of the progression to define a series of stages through which students might pass on their way to learning "geometric measurement of area"; for a summary see Barrett et al. (2011), Battista, Clements, Arnoff, Battista, and Borrow (1998), and Clements and Sarama (2009, 2013). They also relied on the work of the *Common Core Standards Writing Team* (CCSWT, 2012), which has produced several draft progressions that attempt to tie existing learning sciences research to specific *Common Core State Standards* in order to lay out a hypothetical progression of topics.

The LP is represented in the graphic shown in Figure 2.2. Although the focus of the LP is "geometric measurement of area," the LP includes three other topics that are related to students' learning and performance in the measurement of area: "length measurement," "figure composition and decomposition," and "geometric shapes." Research summarized in Clements and Sarama (2009, 2013) suggests that concepts and practices from these related topics are integrated with earlier concepts and practice in geometric measurement in forming later, more sophisticated concepts and practices in the "geometric measurement of area".

The LP begins with children's early understandings about area, which typically represents area as the amount of two-dimensional space enclosed by a boundary - the "attribute of area" (Baturo & Nason, 1996). Students initially can visually compare two objects or shapes directly by laying them side by side or by superimposing one on top of the other. At this "perceptual coordination of attributes" stage, some students are unable to compare a shape in two dimensions (Baturo & Nason, 1996). As students progress in their understanding, they can decompose a shape and rearrange its pieces so that it can fit inside the other shape. Through such experiences, students come to understand "conservation of area," or the idea that a shape can be rotated or decomposed and its pieces rearranged without changing the area of the shape (Baturo & Nelson, 1996; Kamii & Kysh, 2006; Kordaki, 2003).

Students eventually develop an understanding of the square as the unit of area, and learn how to quantify the amount of area in an object or shape. They initially do so by iterating ("area unit iteration") and counting equal-sized units ("equal area units") to determine the area of a shape (Baturo & Nason, 1996; Clements & Sarama, 2009; Zacharos, 2006). Students begin by counting individual unit squares to measure area ("using area units to measure") (Battista et al., 1998; Baturo & Nason, 1996; Zacharos, 2006) and are eventually able to make use of the row and column structure apparent in a rectangular array to compute area more efficiently. For example, once students learn to recognize rows and columns as collections of single units, they can use repeated addition or skip counting of the number of row or column units to compute area ("using area composites to measure") (Battista et al., 1998). Similarly, once students see an array as a collection of rows and columns, they can multiply the number of row units by the number of column units to compute area ("using multilevel area composites to measure") (Battista et al., 1998). The process of constructing arrays and understanding how and why they can represent area is crucial for the formula "area = length × width" to be understood conceptually (Battista et al., 1998).

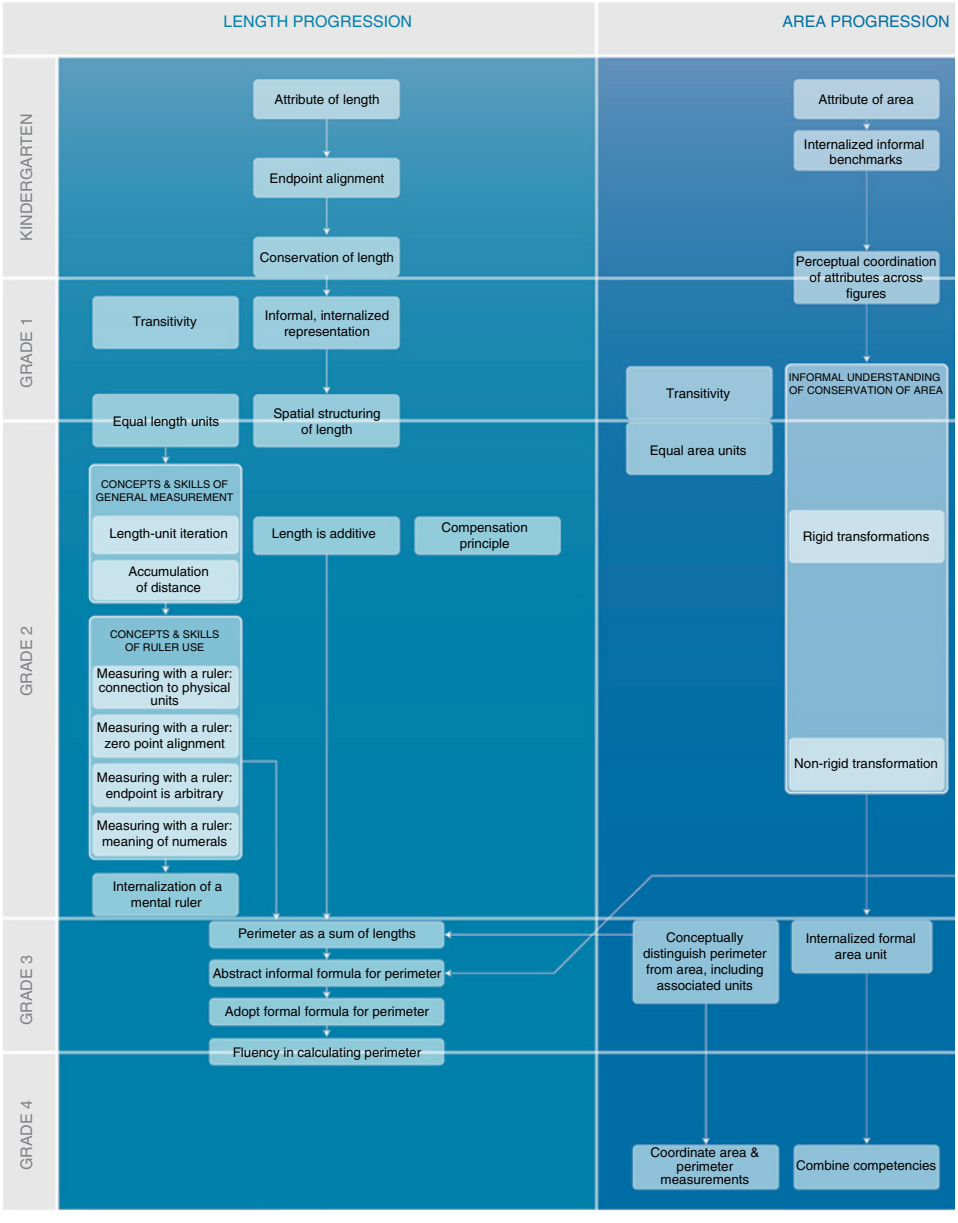

Figure 2.2 Learning progression for "geometric measurement of area."

Under the LP, students next develop the ability to estimate the area of objects with relative accuracy using standard square units ("internalized formal area unit"). Around the same time, they understand the idea that the area of a larger shape can be computed by adding together the area of smaller shapes that comprise it, as well as the idea that the area of a smaller shape can be computed by subtracting its area from a larger shape ("area is additive") (Zacharos, 2006). At some point, students no longer need to visualize the spatial structuring of shapes into rows, columns, and units. They understand the

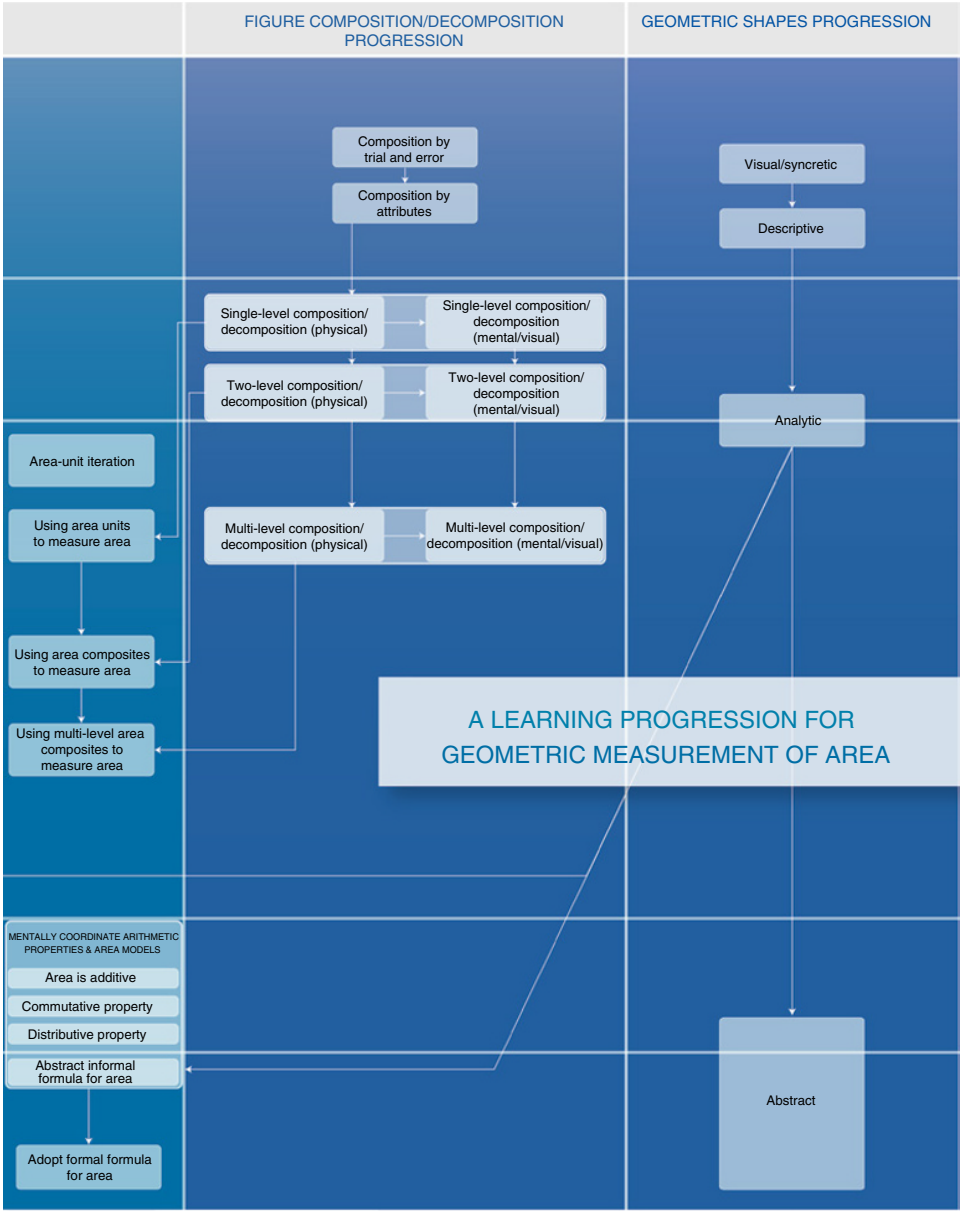

Figure 2.2 (Continued)

dimensions of a shape to represent the number of units per row and column and can multiply them to find the area ("abstract informal formula for area") (Clements & Sarama, 2009).

Finally, students must also be able to distinguish area from perimeter. Though perimeter measurement is a separate concept included in the length strand, students may struggle to differentiate the two, particularly their respective units (Baturo & Nason, 1996). In understanding area and perimeter in contrast to one another, students

reinforce their understanding of the two distinct concepts. Eventually, students should be able to coordinate area and perimeter measurements such that they realize two shapes can have the same area but different perimeters and vice versa (Baturo & Nason, 1996; Kordaki, 2003).

To make this LP useful for assessment design and development, Lai et al. (2015) had to go beyond descriptions of each stage. They did what Leighton and Gierl (2011; Kindle Locations 350) describe as "reading, sifting, and interpreting the research" to identify rich descriptions of the items and tasks researchers used to elicit the concepts and practices as well as the kinds of performances that served as evidence of different stages in understanding. For example, Lai et al. (2015) used the studies reported in Battista et al. (1998), Baturo and Nason (1996), De Bock, Verschaffel, and Janssens (1998), and Zacharos (2006) to identify the use of various types of shapes (regular rectangles, non-rectangular shapes such as T- or L-shapes, and irregular shapes such as blobs) that were used to elicit evidence of students' understanding of area measurement and the kinds of performances in the contexts of those shapes that served as evidence for stages in understanding of area measurement; more examples are provided in Table 2.1. However, Leighton and Gierl (2011) warn that translating research from the learning sciences in this way leaves the findings susceptible to error or bias in the process of synthesis.

Table 2.1 Content features of tasks that elicit evidence on students' understanding of shape composition/decomposition, length, and area measurement.

LP strand	*Content features*
Shape composition and decomposition	Tasks featuring geometric figures with varying levels of scaffolding for visualizing the spatial structure (full or partial grids, dot grids, and versions of shapes with units demarcated by hash marks; Battista et al., 1998; Battista, 2004). Tasks varying shape orientation (e.g., showing a shape on its side so that it looks like a diamond; Sarama and Clements, 2009).
Length	Tasks that ask students to compare the length of two or more objects using a variety of physically manipulable tools (length units, standard rulers, "broken rulers," straight edges, and nonstandard measuring units (e.g., a book; Barrett et al., 2006, 2011). Tasks involving perimeter or length "paths" that switch directions are more complex than those with simple one-dimensional length measurement (Barrett et al., 2006).
Area	Tasks including incomplete figures with varying levels of scaffolding (full and partial grids that support students in enumerating unit squares to varying degrees; Battista et al., 1998; Battista, 1994). Tasks featuring various types of shapes, i.e., regular rectangles, non-rectangular shapes (T- or L-shapes), and irregular shapes (i.e., blobs; Battista et al., 1998; Baturo and Nason, 1996; De Bock, Verschaffel, and Janssens, 1998; Zacharos, 2006). Tasks including manipulable shapes students can superimpose on top of one another or decompose into pieces (Kordaki, 2003).

Now that we have broadly described a LP for "geometric measurement of area" that we can use for illustration, we will present the following three criteria for evaluating the fitness of theories of learning and cognition to inform assessment design and development:

1. Clarification of the targets of inference,
2. Identification of the features of content; and
3. Identification of the features of learners' performance.

Criterion 1: Clarification of the Targets of Inference

Looking through the lens of PAD, the first criterion is the degree to which the theory clarifies the targets of inference so as to support assessment design and development. As discussed in the previous section, PAD is construct-centered meaning that design decisions are derived from the description of the targets of inference. We thus propose that a theory of learning and cognition that is adequately defined to support construct-centered assessment design and development would address the following four aspects: *aggregation*, *change*, *fairness*, and *backing*.

Aggregation. In terms of aggregation, following Pellegrino et al. (2001), theories of higher resolution that distinguish between different cognitive processes, knowledge structures, strategies and mental models should allow disaggregation to a lower resolution. The resolution level used in assessment design and development should be determined by the detail required to engineer assessments whose results are likely to support the intended user(s), use(s), and interpretation(s) of the assessment. A relatively low resolution LP may offer adequate detail to support the development of an assessment whose results are intended to be interpreted in terms of overall domain achievement. For example, assessment results may be intended to be used by policymakers to evaluate the success of current classroom practices in helping children learn geometric measurement. Given limits on testing time and the summary nature of the intended use, assessment designers may focus on low-resolution targets of inference such as the following (Clements & Sarama, 2009):

- Understanding length as a characteristic of an object found by quantifying how far it is between the endpoints of the object or between any two points,
- The capability to use a ruler correctly,
- Understanding area as an amount of two-dimensional surface that is contained within a boundary; and
- Understanding volume as spatial structuring in three dimensions.

Alternatively, assessment results may be intended to inform decisions made by teachers on the instructional experiences to arrange for students beginning to study area measurement. The relatively low-resolution LP described in the previous paragraph will likely offer little support for the development of an assessment whose results are intended to be interpreted in terms of the nature of students' understanding of area and

used to create or identify learning experiences that build on students' current understandings. For these teachers, assessment designers may focus on relatively higher-resolution targets of inference from the "geometric measurement of area" LP such as the following:

- Understanding the attribute of area as different from attributes of length or volume;
- Possessing intuitive, informal, internalized representations for amounts of area;
- Being capable of coordinating both length and height in estimating area; and,
- Being capable of visualizing two-dimensional shapes as collections of area composites.

Even when a lower-resolution theory offers adequate detail to support the intended interpretation of assessment results in terms of overall domain achievement, a higher-resolution theory of learning and cognition may be needed to inform assessment design and development decisions with respect to the features of test-taker performances that serve as evidence, the features of stimuli and items that tend to effectively elicit that performance, and how that performance will be evaluated and aggregated. For example, the relatively low-resolution target of inference of students understanding area as an amount of two-dimensional surface that is contained within a boundary may require items or tasks that sample from the four relatively higher-resolution targets of inference described in the last paragraph.

With respect to aggregation, the LP appears to offer natural levels of aggregation. For example, an assessment could report results for an overall understanding of area, for each of the four strands, or disaggregate understanding of area further into the concepts and practices within each of the four strands. However, while suggesting units into which assessment results could be aggregated or disaggregated, the LP lacks direction for the manner in which units should be aggregated or disaggregated. For example, the LP lacks any rationale for differentially or equally weighting results from each of the four strands when creating an overall indicator of understanding area.

Change. In terms of change, the theory of learning and cognition should include an explanation of the mechanisms that lead to learning in the domain and the pathways along which learning progresses. This explanation should describe how the mechanisms change the cognitive processes, knowledge structures, strategies, and mental models as a learner becomes more accomplished in a domain. Furthermore, the explanation of how learning occurs should describe the nature of the cognitive processes, knowledge structures, strategies and mental models at critical points given the purpose of assessment, as the learner progresses from less to more accomplished in the domain or transitions from a novice to an expert.

Both Suppes (2002) and Pellegrino et al. (2001) have championed the specification of the mechanism for learning. Suppes (2002) argued that what are essentially the evidence rules and statistical model under ECD should be identical or at least similar in structure to the phenomena being modeled. As already noted, Pellegrino et al. (2001) advocated that the methods used to elicit performances as evidence and analyze and make inferences from that evidence, represented by the *Observation* and *Interpretation* vertices of

the assessment triangle, respectively, should be explicitly coordinated with the targets of inference, represented by the *Cognition* vertex. The implications for assessment design of the view of how students learn was described by Pellegrino et al. (2001, p. 26):

> Current assessments are also derived from early theories that characterize learning as a step-by-step accumulation of facts, procedures, definitions, and other discrete bits of knowledge and skill. Thus, the assessments tend to include items of factual and procedural knowledge that are relatively circumscribed in content and format and can be responded to in a short amount of time. These test items are typically treated as independent, discrete entities sampled from a larger universe of equally good questions. It is further assumed that these independent items can be accumulated or aggregated in various ways to produce overall scores.

While the LP includes distinct stages as the learner progresses from less to more sophisticated understanding of length, area, shape composition and decomposition, and geometric shapes, it does not explicitly describe or explain the learning mechanisms that are hypothesized to produce more sophisticated concepts and practices. That is, the LP describes the nature of KSAs in different stages but omits a description of how the differences come about. Learners may move through stages by the step-by-step accumulation of discrete bits of knowledge and skill. If so, then assessments based on the geometric measurement LP would be developed and scored like conventional multiple choice tests. But if learners move through stages by reorganizing flawed mental models (Chi, 2008), then assessments based on the geometric measurement LP would be developed and scored to reflect an ordinal-level scale.

Fairness. In terms of fairness, the **model of learning** and the **model of cognition** should account for systematic differences in the way students from different backgrounds learn and think. Pellegrino et al. (2001) also noted that models should address differences in the way students learn and perform in a domain. Typically, these differences would be related to culture, but these differences may also be related to other learner variables such as age or learning style. With respect to fairness, Lai et al. (2015) failed to include potential differences in the way students from different backgrounds learn and perform in a domain.

As our relatively short discussion suggests, fairness is perhaps the least developed aspect of how well the LP clarifies "geometric measurement of area" as the target of inference. The LP provides the assessment designer no guidance with respect to designing an assessment that is fair to learners from different cultures or learning styles.

Backing. Finally, the theory of learning and cognition should be backed by empirical research. The assessment designer is making a claim that the theory of learning and cognition clarifies the domain-specific cognitive processes, knowledge structures, strategies, and mental models that explain the targets of inference. In general, the greater the breadth and depth of the set of empirical studies that back the theory the better the theory is qualified to inform assessment design and development decisions. However, the required strength of the backing may be related to the stakes associated with the assessment. For example, the design and development of a relatively low-stakes

assessment may be guided by a theory backed by less empirical support while the design and development of a relatively high-stakes assessment should be guided by a theory backed by extensive empirical support.

The determination that a theory of learning and cognition is qualified to inform assessment design and development decisions is complicated. We challenge the reader to find universally agreed-upon rules or criteria for judging the quality of a theory. Toulmin (1958; see also Jackson, 2011) asserts that how data (e.g., the findings from research studies) are interpreted as proof of a certain claim (e.g., that a theory of learning and cognition possesses structural validity) is highly field-dependent. Both the kind of warrants and the power of warrants that authorize the taking of data as proof of a claim grow from the transcendent, socially constructed, authority of a field. The conclusion on the quality of backing for a theory will depend on the conventions, practices, and values of the field to which the audience for that judgment belongs.

Toulmin is not alone in rejecting the use of universal criteria to evaluate backing for a theory. Drawing on discourse analysis, Hyland (2004, 2009) described a similar phenomenon that occurs across different academic fields, which have distinctive ways of asking questions, addressing a literature, criticizing ideas, and presenting arguments. These differences across fields even influence the verbs selected to describe findings from the literature (Wells, 1992): "It turns out, in fact, that engineers *show*, philosophers *argue*, biologists *find*, and linguists *suggest*" (Hyland, 2009, p. 11, italics in original). The use of field-dependent criteria to evaluate backing for a theory is a particular problem for evaluating the area LP for use in assessment design and development. For example, at least some researchers in the mathematics education field (Lesh & Sriraman, 2005), from which the research to construct the LP that we described previously was drawn, reject psychometric sources of backing as "perverse psychometric notions of 'scientific research'" (p. 500).

Some writers in the field of assessment have similarly acknowledged the field-dependent nature of judgments of backing. For example, Messick (1981, 1989) describes the ideologies of potential test users and argued that the different communities from which test users may be drawn bring different, but perhaps overlapping, conventions, practices, and values to the evaluation of validity and can reach radically different conclusions. Messick (1981, 1989; see also Kane, 2001; Hubley & Zumbo, 2011).

Critics may disagree and argue that, at some level, stakeholders from different fields will agree on criteria for determining that a theory is qualified to inform assessment design and development. We similarly acknowledge that evidence and rationales may be expressed in broad enough ways to be supported by stakeholders across most fields. But we nevertheless agree with Kuhn (1970) who argues that when expressed in that broad way, the evidence and rationales are powerless to settle difficult or contentious arguments. When expressed more precisely, such evidence and rationales diverge into field-dependent conventions, practices, and values. As an example, most – if not all – stakeholders would agree with the claim that assessments should be fair but we contend that the evaluation of arguments would quickly devolve into field-dependent conventions, practices, and values when pressed on the claims that a particular assessment is fair.

The establishment of some guidance for determining if a theory of learning and cognition used to inform assessment design and development has sufficient backing given the intended assessment use is important. Such a critical topic deserves more discussion than we can offer in this chapter. However, we urge the assessment community to avoid myopicism in proposing this guidance and remember that the audience for validity arguments, for which the support for the theory of learning and cognition that is used to inform assessment design and development is part of that argument, is likely drawn from diverse fields of practice including teachers and mathematics education.

Criterion 2: Identification of Relevant Content Features for Items or Tasks

As we discussed in the previous section, PAD approaches attempt to engineer intended interpretations and uses of assessment results through the explicit manipulation of the *Observation* vertex and the features of content that tend to effectively elicit targets of inference. Again, looking through the lens of PAD, a second criterion for the evaluation of theories of learning and cognition is the degree to which such theories give support for coordinating decisions with regard to the *Observation* vertex of the assessment triangle. From the perspective of PAD, theories of learning and cognition, along with the empirical research associated with the theories, should inform the identification of important content features for eliciting evidence of learners' status with respect to the targets of inference.

Typically, studies associated with theories of learning and cognition include rich descriptions of items and tasks employed in manipulating the use of cognitive processes, knowledge structures, strategies, and mental models; see, for example, Battista (1994), Battista et al. (1998), Baturo and Nason (1996), De Bock et al. (1998), and Zacharos (2006). Information on the important content features for eliciting evidence of learners' status with respect to the targets of inference can be found in these descriptions of materials. Assessment designers can review these studies and link features of these items and tasks to elicitation of evidence with respect to the targets of inference. The greater the breadth and depth of empirical studies that link features of these items and tasks to elicitation of evidence with respect to the targets of inference, the stronger is the support that these content features are qualified to inform assessment design and development decisions.

The research associated with the LP on "geometric measurement of area" that we described above can be used to demonstrate the evaluation of the degree to which theories of learning and cognition, along with relevant empirical studies supporting the model, inform the identification of important content features. As Table 2.1 shows specifically, the research literature identifies content features of tasks that elicit evidence with respect to students' understanding of shape composition/decomposition, length, and area measurement. For example, Battista et al. (1998) and Battista (1994) used incomplete figures with varying levels of **scaffolding** for students, such as full and partial grids, that support students in enumerating unit squares to varying degrees. Similarly, Zacharos (2006) provided a one-inch square tile for students to iterate. Another feature of the assessment tasks in the research literature is the types of shapes

presented to students, including both regular rectangles – where the area formula applies – as well as irregular rectangles (e.g., T- or L-shapes), non-rectangular shapes (e.g., ovals), and finally completely irregular shapes (e.g., blobs) (Battista et al., 1998; Baturo & Nason, 1996; De Bock et al., 1998; Zacharos, 2006).

Criterion 3: Identification of Relevant Features of Learners' Performances

Another means through which PAD attempts to engineer intended interpretations and uses is through the explicit manipulation of the *Interpretation* vertex of the assessment triangle and the associated features of performances that serve as evidence of learners' status with respect to the targets of inference. The theory of learning and cognition, along with relevant empirical studies supporting the theory, should inform the description of the performances collected as evidence of learners' status with respect to domain-specific cognitive processes, knowledge structures, strategies, and mental models. Furthermore, the theory should offer guidance on how those responses should be evaluated and aggregated to support inferences with regard to learners' status.

As was the case for content features in the previous sub-section, information on the important features of performances collected as evidence of learners' status with respect to the targets of inference can be found in the studies associated with the theory of learning and cognition. The same studies that provided descriptions of the items and tasks researchers have used also provide description of the kinds of performances that researchers have used as evidence of status with respect to the targets of inference. These studies can be reviewed and the features of performance extracted that are linked to status with respect to the targets of inference.

Again, the empirical studies associated with the "geometric measurement of area" LP can be used to demonstrate the evaluation of the degree to which the studies associated with theories of learning and cognition inform the identification of performances that can serve as evidence of learners' status with respect to the targets of inference. Student non-verbal behaviors are often accepted as evidence of the tenability of differentiating between six stages in the "area" strand of the LP from the research literature; these stages include (1) placing two shapes side by side to visually compare them along one dimension; (2) placing one shape on top of another; (3) decomposing, rearranging, and recomposing shapes from their constituent pieces; (4) using or creating units of equal sizes; (5) counting individual unit squares; skip counting unit squares; and (6) multiplying unit squares (Battista et al., 1998; Baturo & Nason, 1996; Kamii & Kysh, 2006; Kordaki, 2003; Zacharos, 2006).

For example, when iterating units in the "area-unit iteration" stage of the LP, students should cover the entire space without leaving gaps and without overlapping. Students may iterate with multiple copies of a single unit first and later move on to using a single unit and marking off their iterations with a pencil (Battista et al., 1998; Baturo & Nason, 1996; Zacharos, 2006). Whereas in the "two-level composition/decomposition" stage of the LP, students should be able to decompose a shape, such as an irregular rectilinear shape, into simpler shapes whose areas can be computed and added to find the area of the total shape.

Table 2.2 shows the nature of learners' performance that is typical in each of six stages from the "area" strand of the LP. The studies associated with the LP offer guidance on how those performances should be evaluated to support inferences with regard to

Table 2.2 Typical learner performance in each of six stages from the "area" strand of the LP.

Stage	*Performance*
Using area units	Learner counts individual unit squares to compute area
Area-unit iteration	Learner places area units end to end along the length and width of the object, leaving no gaps
Using area composites	Learner uses repeated addition to compute area
Using multilevel area composites	Learner multiplies the number of units in a row by the number of units in a column to compute area
Area is additive	Learner adds together the area of two smaller squares to compute area of the larger rectangle
Adopt formal formula for area	Learner multiplies length times width to compute area

learners' status on the LP; see, for example, the studies by Battista et al. (1998), Baturo and Nason (1996), Kamii and Kysh (2006), Kordaki (2003), and Zacharos (2006). The descriptions of performance in Table 2.2 suggest a qualitative, rather than a quantitative, approach to coding performance so that learner performance can be more easily associated with one or more stages in the LP, rather than being assigned a single quantitative score.

Furthermore, the LP should offer guidance on how coded responses should be aggregated to support inferences with regard to what students know and can do, which is done by a statistical model. The "geometric measurement of area" LP describes developing competence in this area in terms of the integration of concepts and practices from geometric composition, decomposition, and measurement of length with earlier concepts and practice in geometric measurement. The conceptual structure of the "geometric measurement" LP suggests that a multidimensional statistical model with four dimensions would more faithfully support inferences with respect to understanding of area than a lower-dimensional or unidimensional model. To this point, recall that Pellegrino et al. (2001) and Suppes (2002) underscore that the statistical model should be explicitly connected and coordinated with the model of learning and cognition during assessment design or the validity of the inferences drawn from the assessment results will be compromised.

Interface of Evaluation Criteria with Validity

Earlier in this chapter we noted that a common characteristic of PAD approaches is concerned with making explicit and transparent the assessment design and development decisions and the evidence and rationales supporting them. This concern leads naturally to the development of an argument for the validity of the interpretation and use of assessment results, which involves delineating and evaluating the plausibility of a set of claims (Kane, 2013).

Although several validation frameworks have been proposed over the years (e.g., Cronbach, 1988; Messick, 1989), Kane's **argument-based approach** (Kane, 2006, 2013) represents a popular current framework. Under this approach, theory-based interpretations involve constructs of interest that are not directly observable but are tied through the theory to observable indicators such as test-taker behaviors or

written or verbal responses to test items. According to Kane (2013), a theory-based interpretation rests on the claims that the theory is plausible and that the indicators provide a reasonable gauge of the state or level of the construct.

Such evidence can be analytical or empirical. For example, analytical evidence might include analyses of the relevance of each indicator to its construct, typically produced during test development. Empirical evidence would permit an evaluation of the fit of the observed student responses to predictions of the theory. Assessment design and development following a PAD approach would fall squarely under theory-based interpretations because all design decisions cascade from the conceptualization of the targets of inference.

The construction of an argument for the validity of a theory-based interpretation under the argument-based approach has two major components, which, although fused in practice, are helpful to distinguish conceptually: (a) the **interpretation and use argument**, which lays out "the network of inferences and assumptions inherent in the proposed interpretation and use," often represented as a series of claims (Kane, 2013, p. 2); and (b) the **validity argument**, which involves collecting and evaluating relevant evidence for each separate claim in the interpretation and use argument to support an overall judgment about validity.

An example of a theory-based interpretation is student status with respect to the LP of "geometric measurement of area." The assessment of status with respect to understanding the concepts and practices of the "geometric measurement of area" is not directly observable but is tied to specific kinds of performances – defined by performance features – occurring in the context of specific kinds of items and tasks – defined by content features. The validity of assessment results for "geometric measurement of area" rests on the claims that the LP is plausible and the content and performance features provide reasonable indicators of status.

The argument for the validity of the theory-based interpretation of results from the assessment of "geometric measurement of area" emerges naturally from the documentation of decisions about which kind of performances to identify as evidence when the performances occur in the context of what kinds of content. In identifying content features, the assessment designer is making a claim that this content elicits evidence of learners' status with respect to the targets of inference. The greater the number of studies in which these content features are found the stronger the empirical support for the claim that the features elicit evidence of learners' status with respect to the targets of inference. The strength of the empirical evidence that supports identification of the content features provides corresponding support for the validity of the inferences drawn from the assessment results.

In identifying performance features, the assessment designer is making a claim that these performances are evidence for learners' status with respect to the targets of inference. The greater the breadth and depth of studies found in which these performance features are used as evidence for the targets of inference, the stronger the empirical support for the claim that these performance features are indicators of status with respect to the targets of inference.

The interface of the evaluation criteria and validity is illustrated by the possible conceptual framework for the interpretation of assessment results with respect to understanding "geometric measurement of area" shown in Figure 2.3.

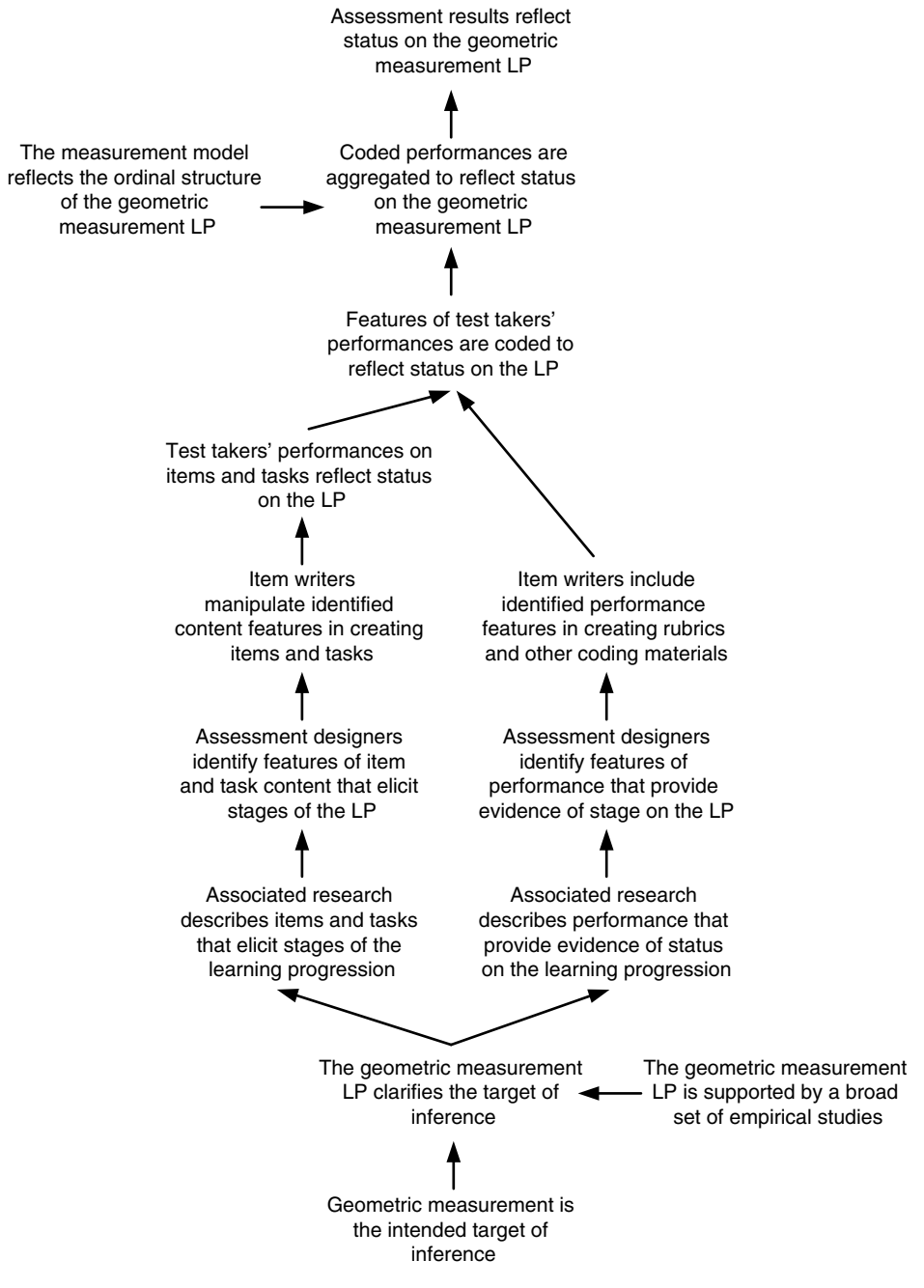

Figure 2.3 The validity argument for the interpretation of assessment results based on the "geometric measurement of area" learning progression.

Specifically, a possible interpretation and use argument for the assessment results is represented by the claims in text shown in Figure 2.3. The interface of the evaluation criteria for the LP and the interpretation and use argument with respect to

understanding "geometric measurement of area" is illustrated by the claims that are incorporated into the evaluation criteria for the LP.

Conclusion

In this chapter, we presented two common characteristics of PAD approaches and three criteria for evaluating how well models of learning and cognition inform assessment design and development decisions. These criteria were grounded in the view represented by PAD that assessment design and development is a systemic endeavor that coordinates activities representing all three of the vertices in the assessment triangle to create coherence; a lack of coherence threatens the validity of the interpretation of assessment results. The source of the coherence for this assessment system is the conscious and deliberate use of a theory of learning and cognition to guide and inform assessment design and development decisions throughout the process.

Our discussion of PAD has been narrowly focused on these three criteria. However, PAD, and the criteria grounded in PAD, might be broadened to embrace more of the concepts and practices involved in making judgments about the technical quality of assessments and in making judgments about the intended consequences of test use. One means of broadening PAD would be to integrate findings from studies of validity and **reliability** as a coherent component through the conscious and deliberate use of a theory of learning and cognition to guide and inform the design and interpretation of such studies.

A further means of broadening PAD is to consider the **theory of action** for an assessment system, which hypothesizes the cause–effect relationships among inputs, activities, and the intended outcomes or consequences. The theory of action has clear implications for how assessment items are designed to elicit evidence of student status with regard to the theory of learning and how student responses will be interpreted to make claims about student standing on the construct(s) targeted by the assessment. It prompts test designers as well as test users to be explicit about how the interpretive claims of the assessment will lead to the intended outcomes or effects, via action mechanisms, and should also anticipate potential unintended outcomes due to misinterpretation or misuse of test results.

The outcomes of a theory of action feed back into the assessment triangle by providing a check on the match between theory and outcomes. For example, an unintended assessment consequence is subgroup differences in test results (i.e., differential impact) which may lead to under- or over-selection of these groups for program placement, advancement, or other decisions. While subgroup differences may be the result of weaknesses in the measurement procedure (e.g., **differential item functioning** or differential test functioning), it may also suggest that the theory of learning is not generalizable to particular subgroups and may need to be modified. This may occur when the theory of learning is extended beyond the contexts in which it has been empirically tested (Kane, 2006). While the assessment triangle promotes coherence for assessment design and development, the iterative cycle between the three vertices is also impacted by test score use which occurs outside the cycle. A theory of action provides mechanisms for checks and balances between what happens within the assessment triangle and external uses of test results that lead to both intended and unintended outcomes.

Finally, we urge that everyone involved in assessment development insist that measurement professionals document evidence of and rationales for claims of coherence between the intended interpretations and uses of assessment results and the assessment design and development decisions. The evidence and rationales supporting claims of coherence should be expected as part of any report on assessment design and development. Documentation should also be expected that details how features of content and the features of performances support coherence with respect to their intended interpretation. By insisting on such documentation, they may disrupt established assessment design and development routines. However, these routines may incorporate outdated assumptions that are at odds with contemporary theories of learning and cognition and their embedded targets of inference.

References

Barrett, J. E., Clements, D. H., Klanderman, D., Pennisi, S. J., & Polaki, M. V. (2006). Students' coordination of geometric reasoning and measuring strategies on a fixed perimeter task: Developing mathematical understanding of linear measurement. *Journal for Research in Mathematics Education*, *37*(3), 187–221.

Barrett, J. E., Cullen, C., Sarama, J., Clements, D. H., Klanderman, D., Miller, A. L., & Rumsey, C. (2011). Children's unit concepts in measurement: A teaching experiment spanning grades 2 through 5. *ZDM*, *43*(5), 637–650.

Battista, M. T. (1994). On Greeno's environmental/model view of conceptual domains: A spatial/geo-metric perspective. *Journal for Research in Mathematics Education*, *25*, 86–94.

Battista, M. T. (2004). Applying cognition-based assessment to elementary school students' development of understanding of area and volume measurement. *Mathematical Thinking and Learning*, *6*(2), 185–204.

Battista, M. T., Clements, D. H., Arnoff, J., Battista, K., & Borrow, C. V. A. (1998). Students' spatial structuring and enumeration of 2D arrays of squares. *Journal for Research in Mathematics Education*, *29*, 503–532.

Baturo, A., & Nason, R. (1996). Student teachers' subject matter knowledge within the domain of area measurement. *Educational Studies in Mathematics*, *31*(3), 235–268.

Chi, M. T. H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 61–82). New York: Routledge.

Clements, D. H., & Sarama, J. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. New York City: Routledge.

Clements, D. H., & Sarama, J. (2013). Rethinking early mathematics: What is research-based curriculum for young children? In L. D. English & J. T. Mulligan (Eds.), *Reconceptualizing early mathematics learning* (pp. 121–147). Dordrecht, The Netherlands: Springer.

Common Core Standards Writing Team (CCSWT). (2012). *Progressions documents for the common core math standards, Draft K-5, Progression on geometry*. Retrieved from http://ime.math.arizona.edu/progressions/#products

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Daro, P., Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction* (Consortium for Policy Research in Education Report #RR-68). Philadelphia, PA: Consortium for Policy Research in Education.

De Bock, D., Verschaffel, L., & Janssens, D. (1998). The predominance of the linear model in secondary school students' solutions of word problems involving length and area of similar plane figures. *Educational Studies in Mathematics*, *35*(1), 65–83.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 300–396.

Gitomer, D. H., Curtis, M. E., Glaser, R., & Lensky, D. B. (1987). Processing differences as a function of item difficulty in verbal analogy performance. *Journal of Educational Psychology*, *79*, 212–219.

Glaser, R., Lesgold, A., & Lajoie, S. P. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J. C. Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology on testing* (Vol. *3*, pp. 41–85). Hillsdale, NJ: Erlbaum.

Gordon, E. W. (2012). To assess, to teach, to learn: A vision for the future of assessment in education. In The Gordon Commission on the Future of Assessment in Education. *The Gordon Commission final report* (pp. 142–162). Retreved from: http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*, 219–230. DOI 10.1007/s11205-011-9843-4.

Hyland, K. (2004). *Disciplinary discourses*. Ann Arbor, MI: University of Michigan Press.

Hyland, K. (2009). Writing in the disciplines: Research evidence for specificity. *Taiwan International ESP Journal*, *1*, 5–22.

Jackson, P. T. (2011). *The conduct of inquiry in international relations: Philosophy of science and its implications for the study of world politics*. London: Routledge.

Kamii, C., & Kysh, J. (2006). The difficulty of "length × width": Is a square the unit of measurement? *The Journal of Mathematical Behavior*, *25*(2), 105–115.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kordaki, M. (2003). The effect of tools of a computer microworld on students' strategies regarding the concept of conservation of area. *Educational Studies in Mathematics*, *52*(2), 177–209.

Kuhn, T. (1970). *The structure of scientific revolutions*. Chicago: The University of Chicago Press.

Lai, E. R., Kobrin, J. L., Holland, L., & Nichols, P. (2015). *Developing and evaluating learning progression-based assessments in mathematics*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 3–18). Cambridge: Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. Cambridge, UK: Cambridge University Press.

Lesh, R., & Sriraman, B. (2005). Mathematics education as a design science. *Zentralblatt für Didaktik der Mathematik*, *37*(6), 490–505.

Lewis, A. B., & Mayer, R. E. (1987). Students miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology*, *79*, 363–371.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Lohman, D. F., & Ippel, M. J. (1993). Cognitive diagnosis: From statistically-based assessment toward theory-based assessment. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 41–71). Hillsdale, NJ: Erlbaum.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*(1), 1–38.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, *10*(9), 9–20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York, NY: Macmillan.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed. pp. 257–306). American Council on Education and Praeger Publishers.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

Mitchell, S. (2009). Complexity and explanation in the social sciences. In C. Mantzavinos (Ed.), *Philosophy of the social sciences* (pp. 130–153). Cambridge, UK: Cambridge University Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

Newell, A. (1990) *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Nichols, P. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64*(4), 575–603.

Nichols, P., Ferrara, S., & Lai, E. (2014). Principled design for efficacy: Design and development for the next generation tests. In R. W. Lissitz (Ed.), *The next generation of testing: Common Core Standards, SMARTER-BALANCED, PARCC, and the nationwide testing movement* (pp. 228–245). Charlotte, NC: Information Age Publishing.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. New York, NY: Routledge.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, *46*(6), 699–715.

Suppes, P. (2002). *Representation and invariance of scientific structures*. Stanford, CA: CSLI Publications.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Wells, G. (1992). The centrality of talk in education. In K. Norman (Ed.), *Thinking voices: The work of the National Oracy Project*. London, UK: Hodder and Stoughton.

Zacharos, K. (2006). Prevailing educational practices for area measurement and students' failure in measuring areas. *The Journal of Mathematical Behavior*, *25*(3), 224–239.

3

Principled Approaches to Assessment Design, Development, and Implementation

Steve Ferrara, Emily Lai, Amy Reilly, and Paul D. Nichols

Long ago, John Bormuth referred to the process of item and test development as a "dark art," in which "construction of achievement test items [is] defined wholly in the private subjective life of the test writer" (Bormuth, 1970, pp. 2–3; also cited in Ferrara, 2006, p. 2). Much has changed – or, rather, much is in the process of changing. Some assessment programs now use principled approaches to assessment design, development, and implementation that shed light on the "dark art." Similarly, many assessment programs now use an argumentation approach to the validation of test score interpretations and uses (see Kane, 2006, 2013, 2016), though the matter of how to implement this approach in a consistent and rigorous manner is far from settled (see Borsboom & Markus, 2013; Lissitz & Samuelson, 2007).

In this chapter, we describe and develop five **foundation elements** and an **organizing element** that define principled approaches to assessment design, development, and implementation and the ongoing accumulation and synthesis of evidence to support claims and **validity arguments**. Specifically, the five foundation elements are (a) clearly defined **assessment targets**, (b) a statement of **intended score interpretations and uses**, (c) a **model of cognition, learning, or performance**, (d) aligned **measurement models** and reporting scales, and (e) **manipulation of assessment activities** to align with assessment targets. The overarching, organizing element is the ongoing accumulation of evidence to support validity arguments.

We illustrate five **principled assessment design** approaches currently in use that adapt and embed the foundation elements and discuss how the five approaches emphasize the five elements differently. The five approaches are:

1. **Evidence-centered design** (ECD),
2. **Cognitive design systems** (CDS),
3. **Assessment engineering** (AE),

The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, First Edition.
Edited by André A. Rupp and Jacqueline P. Leighton.