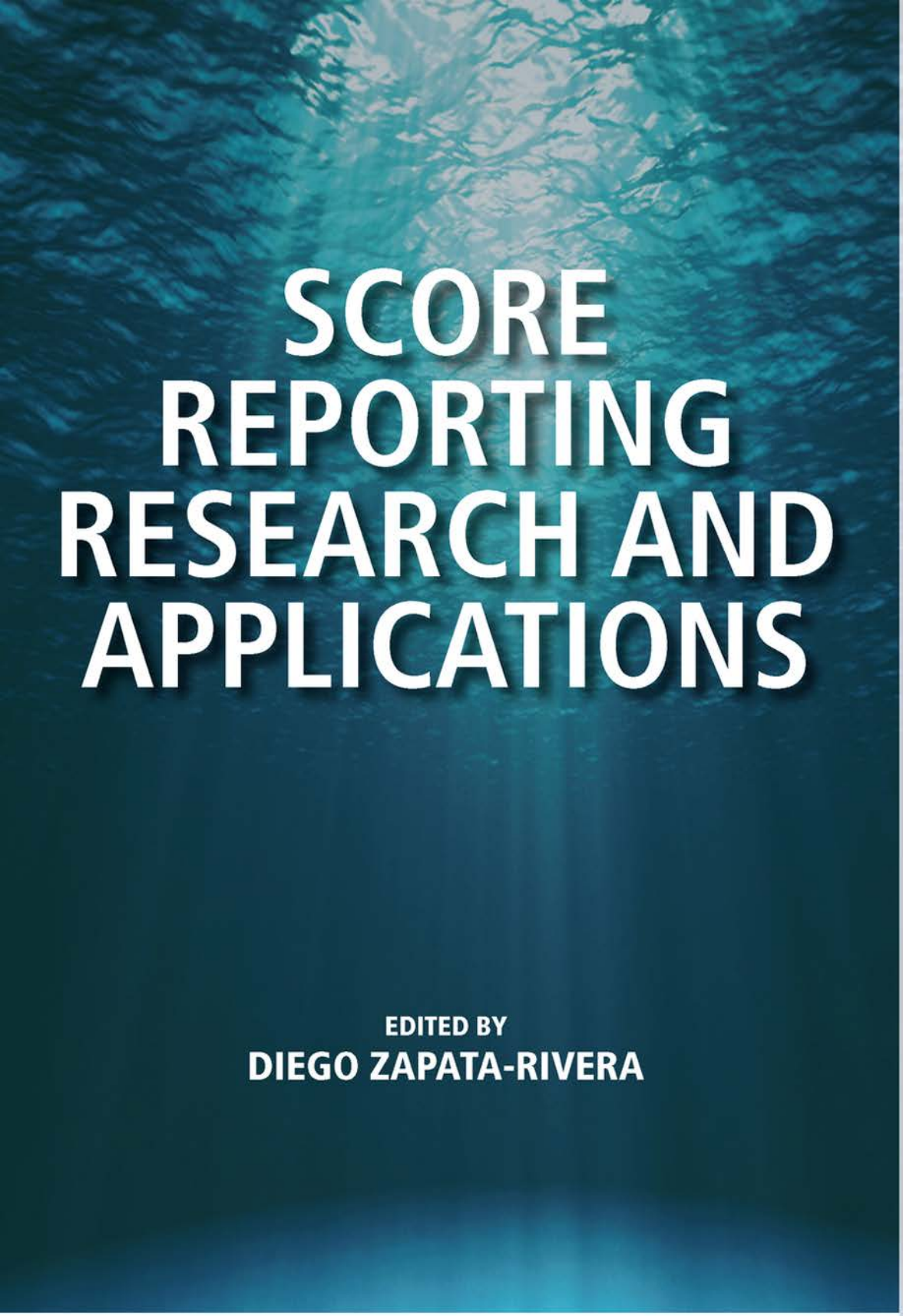


SCORE REPORTING RESEARCH AND APPLICATIONS

EDITED BY

DIEGO ZAPATA-RIVERA

APPLICATIONS OF EDUCATIONAL MEASUREMENT AND ASSESSMENT

NCME national council on measurement in education

ROUTLEDGE

Score Reporting Research and Applications

Score reporting research is no longer limited to the psychometric properties of scores and subscores. Today, it encompasses design and evaluation for particular audiences, appropriate use of assessment outcomes, the utility and cognitive affordances of graphical representations, interactive report systems, and more. By studying how audiences understand the intended messages conveyed by score reports, researchers and industry professionals can develop more effective mechanisms for interpreting and using assessment data.

Score Reporting Research and Applications brings together experts who design and evaluate score reports in both K-12 and higher education contexts and who conduct foundational research in related areas. The first section covers foundational validity issues in the use and interpretation of test scores; design principles drawn from related areas including cognitive science, human-computer interaction, and data visualization; and research on presenting specific types of assessment information to various audiences. The second section presents real-world applications of score report design and evaluation and of the presentation of assessment information. Across ten chapters, this volume offers a comprehensive overview of new techniques and possibilities in score reporting.

Diego Zapata-Rivera is Principal Research Scientist in the Cognitive and Technology Sciences Center at Educational Testing Service, USA. He is a member of the Editorial Board of User Modeling and User-Adapted Interaction and Associate Editor of *IEEE Transactions on Learning Technologies*.

The NCME Applications of Educational Measurement and Assessment Book Series
Editorial Board:

Technology and Testing: Improving Educational and Psychological Measurement
Edited by Fritz Drasgow

Meeting the Challenges to Measurement in an Era of Accountability
Edited by Henry Braun

Fairness in Educational Assessment and Measurement
Edited by Neil J. Dorans and Linda L. Cook

Testing in the Professions: Credentialing Policies and Practice
Edited by Susan Davis-Becker and Chad W. Buckendahl

Validation of Score Meaning for the Next Generation of Assessments: The Use of Response Processes
Edited by Kadriye Ercikan and James W. Pellegrino

Preparing Students for College and Careers: Theory, Measurement, and Educational Practice
Edited by Katie Larsen McClarty, Krista D. Mattern, and Matthew N. Gaertner

Score Reporting Research and Applications
Edited by Diego Zapata-Rivera

For more information about this series, please visit: www.routledge.com/series/NCME

Score Reporting Research and Applications

Edited by Diego Zapata-Rivera

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2019
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
Names: Zapata-Rivera, Diego, editor.
Title: Score reporting research and applications / edited by Diego Zapata-Rivera.
Description: New York, NY : Routledge, 2019. | Series: NCME Applications of educational measurement and assessment | Includes bibliographical references and index.
Identifiers: LCCN 2018017506 (print) | LCCN 2018034018 (ebook) | ISBN 9781351136501 (eBook) | ISBN 9780815353393 (hbk) | ISBN 9780815353409 (pbk) | ISBN 9781351126501 (ebk)
Subjects: LCSH: Educational tests and measurements—Evaluation. | Examinations—Validity | Grading and marking (Students)
Classification: LCC LB3051 (ebook) | LCC LB3051 .S37 2019 (print) | DDC 371.26/2—dc23
LC record available at https://lccn.loc.gov/2018017506

ISBN: 978-0-8153-5339-3 (hbk)
ISBN: 978-0-8153-5340-9 (pbk)
ISBN: 978-1-351-13650-1 (ebk)

Typeset in Minion
by Apex CoVantage, LLC

Contents

Contributors

Gavin T. L. Brown
Prof Gavin Brown is Associate Dean of Postgraduate Research and Director of the Quantitative Data Analysis and Research Unit in the Faculty of Education & Social Work at The University of Auckland. He is an Affiliated Professor in Applied Educational Science at Umea University, Sweden and an Honorary Professor in Curriculum & Instruction at the Education University of Hong Kong. His Ph.D. (Auckland) was on teacher conceptions of assessment. His research focuses on the formative qualities of educational assessment and the psychology of teachers and students around assessment. He is the author of Assessment of Student Achievement (Routledge, 2018) and the Chief Section Editor for Assessment, Testing, and Applied Measurement in Frontiers in Education.

Linda Corrin
Dr. Linda Corrin is a senior lecturer in Higher Education in the Williams Centre for Learning Advancement at the University of Melbourne. In this role, she provides support for curriculum development, delivery, and assessment to staff in the Faculty of Business and Economics. Linda holds bachelor's degrees in Law and Information and Communication Technology (University of Wollongong, Australia), a Postgraduate Certificate in Learning and Teaching in Higher Education (University of Roehampton, London), and a Ph.D. in Education (University of Wollongong). Linda's research interests include students' engagement with technology, learning analytics, feedback, and learning design. Currently, she is working on several large research projects that explore ways that learning analytics can be used to provide meaningful feedback to academics and students. Linda is chair of her faculty's Learning Analytics Research Community and co-founder/coordinator of the Victorian and Tasmanian Learning Analytics Network.

Mingyu Feng
Mingyu Feng, Ph.D., is a Senior Research Associate at WestEd. She has a doctorate in Computer Science from Worcester Polytechnic Institute, focusing on research and development of intelligent tutoring systems. Dr. Feng specialized in using educational data mining methods to analyze data from computer-supported learning systems and detecting behavioral patterns in user learning process. She serves on the board of the International Society on Educational Data Mining (EDM) and was program co-chair for EDM 2016. Dr. Feng has received several grants from the Institute of Educational Sciences and National Science Foundation to conduct research to examine efficacy and implementation of advanced educational technologies and their impact on instruction and learning. She has published on extensively on formative assessment, data mining and learning analytics, and computer-supported learning and instruction.

Shuchi Grover
A computer scientist and learning scientist by training, Dr. Grover's work in computer science (CS) and STEM education since 2000 has spanned both formal and informal settings in the US, Europe, and Asia. Her current research centers on computational thinking (CT), CS education, and STEM+CT integration mainly in formal K-12 settings. Dr. Grover is a recipient of several grants from the National Science Foundation to conduct research on curriculum and assessments in STEM learning and CT in varied PK-12 contexts. She also works at the intersection of learning, assessment, and big data analytics to shape future environments for deeper learning. She has authored close to 100 scholarly and mainstream articles. She is a member of the ACM Education Council and the Computer Science Teachers Association's task force on Computational Thinking, advisor to the K-12 CS Framework and to K-12 school districts on CS implementation/integration, and on the editorial board of ACM Transactions on Computing Education.

Shelby J. Haberman
Shelby J. Haberman is a Distinguished Research Scientist at Edusoft. He previously was a Distinguished Presidential Appointee at Educational Testing Service. He has a Ph.D. in Statistics from the University of Chicago, and is a Fellow of the American Association for the Advancement of Science, the American Statistical Association, and the Institute of Mathematical Statistics. He received the NCME Award for Outstanding Technical or Scientific Contributions to the Field of Educational Measurement twice, once each in 2009 and 2015. He is the author of the books *The Analysis of Frequency Data*, *Analysis of Qualitative Data*, and *Advanced Statistics*. He is the author or co-author of more than 100 publications concerning statistics and psychometrics. His research interests include Bayesian statistics, detection of test fraud, missing data analysis, model checking and model selection methods, and reporting of diagnostic scores. His primary research interests include statistics and psychometrics.

Ronald K. Hambleton
Ronald K. Hambleton holds the titles of Distinguished University Professor and Executive Director of the Center for Educational Assessment at the University of Massachusetts Amherst in the US. He earned his Ph.D. in 1969 from the University of Toronto with specialties in psychometric methods and statistics. He is a Fellow of Divisions 5 and 15 of the American Psychological Association, a Fellow of the American Educational Research Association, a Fellow of the International Association of Applied Psychology, and a member of the National Council on Measurement in Education and the International Test Commission. He is the recipient of several national and international awards from ATP, NCME, APA, and AERA for his measurement research. He is a co-author of several textbooks including "Fundamentals of Item Response Theory." He is currently conducting research on three topics: Computer-based testing (e.g., detecting item exposure), methods and guidelines for adapting tests from one language and culture to another, and design and field-testing of new approaches for reporting test scores.

John A. C. Hattie
John Hattie is Laureate Professor at the Melbourne Educational Research Institute at the University of Melbourne, Chair of the Australian Institute of Teaching and School Leaders, and co-director of the Science of Learning Research Centre. His areas of interest are measurement models and their applications to educational problems, and models of teaching and learning. He has published and presented over 1000 papers and supervised 200 theses students and 31 books.

Mary Hegarty
Mary Hegarty is a professor in the Department of Psychological & Brain Sciences at the University of California, Santa Barbara. She received her Ph.D. in Psychology from Carnegie Mellon University. Her research is concerned with spatial cognition, broadly defined and includes research on small-scale spatial abilities (e.g., mental rotation and perspective taking), large-scale spatial abilities involved in navigation, comprehension of graphics, and the role of spatial cognition in STEM learning. She served as chair of the governing board of the Cognitive Science Society and is associate editor of *Topics in Cognitive Science* and past associate editor of *Journal of Experimental Psychology: Applied*.

Priya Kannan
Priya Kannan is a Research Scientist in the Cognitive and Technology Sciences Center at Educational Testing Service (ETS). She received her Ph.D. in Psychometric Research Methodology from the University of Pittsburgh in 2011, and her M.A. in Industrial/Organizational Psychology from Minnesota State University in 2003. Her current research interests focus primarily on score reporting; she has conducted seminal work on how parents from diverse subgroups understand information presented in score reports. She has disseminated her research through a collection of articles in various peer-reviewed journals. She has served on various NCME committees, including as chair of the diversity issues in testing committee in 2016–17 and as chair of the Bradley Hansen award committee in 2017–18. She also serves on the "Score Report Design Team"—a cross-divisional team responsible for the design of all score reports (both for ETS testing programs and for external K-12/TLC contracts) designed at ETS.

Irvin R. Katz
Irvin R. Katz, Ph.D., is Senior Director of the Cognitive and Technology Sciences Center at Educational Testing Service (ETS) in Princeton, New Jersey. He received his Ph.D. in Cognitive Psychology from Carnegie Mellon University. Throughout his 28-year career at ETS, he has conducted research at the intersection of cognitive psychology, psychometrics, and technology. His research has involved 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. Dr. Katz is also a human-computer interaction practitioner with more than 35 years of experience in designing, building, and evaluating software for research, industry, and government.

Lisa A. Keller
Dr. Lisa A. Keller is Associate Professor of Education at the University of Massachusetts, Amherst and is committed to research in educational assessment that focuses on fairness issues. Fairness in educational assessment takes many forms, and is a complex topic. Her research interests have spanned the range of more technical aspects of psychometrics, to more policy-based research. Her current focus is on the development of image-based assessment to use cross-culturally, for use in many contexts. While interested in psychometric issues, the goals of assessment, and the consequences of the uses of assessments, is of primary interest to help, and protect vulnerable populations.

Andrew Krumm
Andrew Krumm, Ph.D., is Director of Learning Analytics Research at Digital Promise, where he specializes in developing data-intensive research-practice partnerships with educational organizations of all types. As a learning scientist working at the intersection of school improvement

and data-intensive research, Andrew has developed multiple tools and strategies that partnerships can use to engage in collaborative data-intensive improvement projects. His recently published book *Learning Analytics Goes to School: A Collaborative Approach to Improving Education* presents a framework for engaging in education research and improving education practice through the use of newly available data sources and analytical approaches.

Samuel A. Livingston
Samuel A. Livingston is a psychometrician at Educational Testing Service, where he has worked since 1974. His work at ETS has included the development of performance tests, research on methods of equating test scores, and the planning and coordination of statistical operations for several tests. His current duties include reviewing drafts of ETS test manuals and research reports, giving classes in psychometrics for new ETS staff, and serving on the Score Report Design Team. He has written journal articles, book chapters, and research reports about test score equating, item analysis, reliability estimation, and standard setting. He has also written booklets and articles for non-technical readers, including *The "Plain English" Pocket Guide to Testing and Psychometric Terms*. He has reviewed manuscripts for several professional journals and served on the board of advisory editors for the *Journal of Educational Measurement*, *Applied Measurement in Education*, and the *Journal of Experimental Education*.

Francis O'Donnell
Francis O'Donnell is a doctoral candidate in the Research in Educational Measurement and Psychometrics (REMP) program at the University of Massachusetts, Amherst. She is also a senior research assistant at the Center for Educational Assessment at the University. She has completed summer internships at the National Board of Medical Examiners, where she researched approaches for improving subscore interpretation, and Educational Testing Service, where she evaluated validity evidence for a learning outcomes assessment. Prior to entering the REMP program, Francis worked as an assessment analyst at a multi-campus community college in Massachusetts. Her research interests include validation, score reporting, and assessment issues in higher education. In her dissertation, she is investigating differences in the connotation of achievement level labels used in score reports.

Timothy M. O'Leary
Dr. Timothy O'Leary is a recent graduate from the Melbourne Graduate School of Education at The University of Melbourne. The focus of his Ph.D. was on effective score reporting. In particular, his research considered score reporting from a validity perspective.

Yooyoung Park
Yooyoung Park is in her third year of the Research, Educational Measurement, and Psychometrics program at the University of Massachusetts, Amherst. Her research interests include validity studies, multistage-adaptive testing, and detecting item exposure with application of measurement theories in educational testing. Prior to pursuing a Ph.D., she was an elementary school teacher in South Korea after earning her B.A. in Elementary Education from Seoul National University of Education in Korea and her M.A. in Educational Assessment from the Institute of Education, University College London

Gautam Puhan
Gautam Puhan is Principal Psychometrician at Educational Testing Service, Princeton, New Jersey; gpuhan@ets.org. He earned his Ph.D. in educational psychology from the University of Alberta, Canada, in 2003. He has conducted research in the area of test score equating,

differential item functioning, and reporting subscores. This research has been published in important journals such as *Educational Measurement: Issues and Practice*, *International Journal of Testing*, and *Journal of Educational Measurement*. His research in the area of subscore reporting was recognized by the NCME and he was awarded the NCME Award for Technical or Scientific Contributions to the Field of Educational Measurement in April 2009. His work as a lead psychometrician for assigned testing programs has been recognized by Educational testing Service, which awarded him the 2011 Educational Testing Service Presidential Award, one of the most prestigious awards given at Educational Testing Service.

Marc Silver
Marc Silver is Executive Director of User Experience & Digital Strategy at Educational Testing Service, where he is responsible for the usable design of ETS score reports, products, applications, and websites. He has designed and overseen the usable design of hundreds of products and applications in a career spanning more than 30 years. Mr. Silver currently serves on the Industrial Advisory Board for the Rutgers University Master of Business and Science program, where he advises the program and its students in the field of user experience design. He is author of *Exploring Interface Design* (Cengage Learning), a college textbook covering user experience design practices and methods that won adoptions worldwide. Mr. Silver is the inventor of mobile games that were produced and sold worldwide by UK-based Astraware. One such game, branded as *Rubik's RoundUp*, was featured for nearly 10 years on the official Rubik's Cube website.

Sandip Sinharay
Sandip Sinharay is a Principal Research Scientist at Educational Testing Service in Princeton, New Jersey. He received his doctoral degree from the Department of Statistics at Iowa State University in 2001. He received the Brad Hanson Award, the Award for Outstanding Technical or Scientific Contribution to the Field of Educational measurement, and the Jason Millman Promising Measurement Scholar Award from the National Council on Measurement in Education in 2018, 2015, and 2006, respectively. Dr. Sinharay is the joint editor of two books including Volume 26 on Psychometrics of the Handbook of Statistics series. He is the author of more than 100 research articles in peer-reviewed journals in educational measurement and statistics. He was an editor of the Journal of Educational and Behavioral Statistics between 2010 and 2014. His research interests include Bayesian statistics, detection of test fraud, missing data analysis, model checking and model selection methods, and reporting of diagnostic scores.

Stephen G. Sireci
Stephen G. Sireci, Ph.D. is Distinguished University Professor and Director, Center for Educational Assessment in the College of Education at the University of Massachusetts, Amherst. He earned his Ph.D. in psychometrics from Fordham University, and his master's and bachelor's degrees in psychology from Loyola College Maryland. He is a Fellow of AERA and of Division 5 of the American Psychological Association. He is a former Board Member for the National Council on Measurement in Education (and forthcoming President), former President of the Northeastern Educational Research Association, and a Council Member for the International Test Commission. He has received awards from UMass (College of Education's Outstanding Teacher Award, the Chancellor's Medal, Conti Faculty Fellowship), from NERA (Distinguished Mentoring, Outstanding Service), and from ETS and the International Language Testing Association (Samuel Messick Memorial Lecture Award). He is on the editorial board for several journals and is founder of Sireci Psychometric Services, Inc.

Sharon Slater
Sharon Slater is a psychometrician at Educational Testing Service. She currently leads the Score Report Design Team, a cross-departmental group of staff responsible for assisting various testing programs in creating or revising their score reports. In this role, she has overseen the development of over 30 score reports since the team was formed in 2014. Her other responsibilities at ETS include providing psychometric support for various ETS testing programs and K12 assessments, reviewing ETS test manuals and research reports, writing grant proposal text and providing costing and budgeting estimates for psychometric work. She has written on the topics of score reporting, computer adaptive testing, standard setting, equating, and mode comparability. She enjoys working with clients and striving to make assessment information understandable to various score report users.

Richard J. Tannenbaum
Richard J. Tannenbaum is a Principal Research Director in the Research and Development Division of Educational Testing Service (ETS). In this role, Richard has strategic oversight for multiple Centers of Research that include more than 100 directors, scientists, and research associates. These Centers address foundational and applied research in the areas of academic-to-career readiness; English language learning and assessment (both domestic and international); and K-12 student assessment and teacher credentialing. Prior to this position, Richard was the Senior Research Director for the Center for Validity Research at ETS. Richard holds a Ph.D. in Industrial/Organizational Psychology from Old Dominion University. He has published numerous articles, book chapters, and technical papers. His areas of expertise include licensure and certification, standard setting and alignment, validation, and assessment development.

Diego Zapata-Rivera
Diego Zapata-Rivera is a Principal Research Scientist in the Cognitive and Technology Sciences Center at Educational Testing Service in Princeton, New Jersey. He earned a Ph.D. in computer science (with a focus on artificial intelligence in education) from the University of Saskatchewan in 2003. His research at ETS has focused on the areas of innovations in score reporting and technology-enhanced assessment including work on adaptive learning environments and game-based assessments. Dr. Zapata-Rivera has produced over 100 publications including journal articles, book chapters, and technical papers. His work has been published in journals such as *Educational Assessment, Assessment in Education: Principles, Policy & Practice, International Journal of Artificial Intelligence in Education*, and *International Journal of Human Computer Studies*. Dr. Zapata-Rivera is a member of the Editorial Board of *User Modeling and User-Adapted Interaction* and an associate editor of *IEEE Transactions on Learning Technologies Journal*.

April L. Zenisky
Dr. April L. Zenisky is Research Associate Professor in the Department of Educational Policy, Research, and Administration in the College of Education at the University of Massachusetts, Amherst, and Director of Computer-Based Testing Initiatives in UMass' Center for Educational Assessment (CEA). Her primary responsibilities at the CEA involve project management and operational psychometrics for computer-based tests for adult basic education programs in Massachusetts. Her main research interests include score reporting, technology-based item types, and computerized test designs, and her collaborative work on score reporting with Ronald K. Hambleton has advanced best practices for report development relative to both individual and group reporting, with a focus on strategies for online reporting efforts.

Rebecca Zwick
Rebecca Zwick is a Distinguished Presidential Appointee in the Psychometrics, Statistics, and Data Sciences area at Educational Testing Service and Professor Emerita at the University of California, Santa Barbara. She has a doctorate in Quantitative Methods in Education from the University of California, Berkeley and an M.S. in Statistics from Rutgers University. Dr. Zwick is a fellow of the American Educational Research Association and the American Statistical Association and is the President of the National Council on Measurement in Education. She is the author of more than 100 publications in educational measurement and statistics and education policy. Her books include *Fair Game? The Use of Standardized Admissions Tests in Higher Education* and *Who Gets In? Strategies for Fair and Effective College Admissions*. Her recent research has focused on college admissions, test validity and fairness, and ways to improve score reporting.

Foreword

Score reports have received short shrift in the measurement research literature and in measurement practice, their design often perceived as secondary to traditional measurement concerns, such as reliability and validity. In practice, score report design sometimes seemed like an afterthought, attended to only after assessment development, score modeling, or other "more pressing" issues. This attitude has always seemed shortsighted to me.

- If validity is a property of the inferences that score users make about tests and test results, then should not the lens through which score users perceive the test results—the score report and other test information—be a central concern of everyone in the assessment field?
- Score reports are the public face of an assessment. When the public seeks information about an assessment, score reports are the most visible, and they can influence public opinion. With so much negative news about assessments and the testing industry in the news of late, should not the academic, government, and corporate assessment world be more careful about the design of score reports?

Because you have picked up this book, you likely agree that score reports are important. Perhaps you create assessments, or otherwise work in the assessment field, so you feel a responsibility to ensure that the person reading the report understands the information to be conveyed, especially the lack of precision. After all, you are a testing professional and the people who are reading the score report likely are not. They do not understand statistics, measurement, or what goes into creating an assessment. Thus, clarity of statistical presentation and explanation of the limits of drawing inferences based on point estimates (scores) should be most critical, right?

Wrong. Well, sort of wrong. Those aspects are important, but a score report should be more than just a score, a comparison with others scores, and fine-print warnings for the reader (albeit an unfair characterization of some score reports, but a fair description of many others). A score report serves a purpose for the reader: to make a decision or to draw an inference about a test taker or group of test takers. Design of score reports must be driven by that purpose and, therefore, driven by the needs of the score user: the student who wants to learn, the parent who wants to help their child learn, the teacher who wants to guide his or her class, the administrator who wants to argue for more resources.

The best score reports are tailored to the needs of the score user, whether the intended audience is students, parents, teachers, administrators, or other stakeholders. However, how does one do that tailoring? As someone who has worked in the testing industry for almost 30 years, I know that such design is tough. When I worked on a new assessment, people wanted to know

"what am I going to get from this assessment?" They were sometimes interested in seeing sample tasks, hearing about the underlying construct, and learning about our validity research, but they always insisted on seeing the score reports. We had two reports—one tailored to the test taker and two for the institutions—that were created using some of the techniques (e.g., prospective score report, audience analysis, iterative design) discussed in the book you now hold.

I found this book to be an excellent exploration of the design challenge of tailoring score reports. Especially useful is the advice on design processes and on understanding the test user, which may be found in many chapters. Several chapters cover deliberate, systematic approaches to the score report design process that puts the score user—the intended audience—front-and-center. To design for a particular audience requires some knowledge of how people comprehend and process information. Thus, some chapters cover how people perceive score reports, whether accounting for how people comprehend and process visual inputs, or what inferences people draw from different types of displays (and what incorrect inferences they are likely to draw). Designing useful score reports for particular audiences can be tough, but the experts that editor Diego Zapata-Rivera recruited to contribute should help anyone who wants to create better score reports.

The contributors to this book represent some of the multi-disciplinarity that should go into score report creation. Certainly, there are psychometricians, particularly measurement experts. There are also cognitive psychologists, experts in validity theory, computer scientists, education researchers, data scientists, and statisticians. It takes combinations of these disciplines to understand how test users make meaning from a score report, how they draw inferences, and how they make decisions. Building on this understanding, these different disciplines have much to say about how to design tailored reports that ensure that the inferences and decisions test users make are appropriate; that they reflect the strengths and limits of the information presented in the score report.

Whether you are a researcher, practitioner, or both, *Score Reporting Research and Applications* will provide you with much to think about regarding, and concrete recommendations for, the design of score reports that help score users make appropriate, evidence-based decisions.

Irvin R. Katz
Princeton, NJ
March 28, 2018

Acknowledgments

I would like to thank the authors who agreed to participate in this project. In addition, I would like to express my gratitude to those who offered me their support and encouragement during the preparation of this book, especially to Rebecca Zwick and Irvin R. Katz. I also want to thank the members of the NCME book series editorial board for giving the authors and me the opportunity to make this project a reality. My gratitude goes to the members of the senior leadership at ETS including Ida Lawrence, Joanna Gorin, Randy Bennett, and Andreas Oranje for supporting this project and our research in the area of score reporting throughout the years. Furthermore, I would like to recognize the work of ETS staff who reviewed chapters authored by ETS authors and Andrew Chin who assisted me in the later stages of the project. Finally, I would like to thank my wife Lucia and my sons Christian and Lucas for their unconditional support.

Introduction

Why Is Score Reporting Relevant?

Diego Zapata-Rivera

Research in the field of score reporting has increased considerably in the last decade. This area of research is no longer limited to investigating the psychometric properties of scores and subscores but instead its scope has been broadened to include aspects such as: designing and evaluating of score reports taking into account the needs and other characteristics of particular audiences; exploring appropriate use of assessment information; investigating how particular graphical representations are understood by score report users; designing support materials to facilitate understanding and appropriate use of score report information; designing and evaluating interactive report systems; and foundational research on the cognitive affordances of particular graphical representations.

Validity is a property of the interpretation and use of test results (AERA, APA, & NCME, 2014; Hattie, 2009; Kane, 2006). By studying how particular audiences understand the intended messages conveyed by score report information, we can devise mechanisms for supporting valid interpretation and use of score report information. With this book, we bring together a group of researchers working in various areas relating to the field of score reporting including researchers designing and evaluating score reports in the K-12 and higher education contexts and those doing foundational research on related areas that may inform how to communicate assessment information to various audiences.

Papers in this volume build on a growing body of literature in the score reporting field that includes work on frameworks for designing and evaluating score reports (e.g., Hambleton & Zenisky, 2013; Zapata-Rivera & VanWinkle, 2010), approaches for tailoring score reports to particular audiences (e.g., Jaeger, 2003; Zenisky & Hambleton, 2012; Wainer, 2014; Zapata-Rivera & Katz, 2014), and evaluating score reports for teachers (Rankin, 2016; Zapata-Rivera, Zwick, & Vezzu, 2016; Zwick, Sklar, Wakefield, Hamilton, Norman, & Folsom, 2008, Zwick, Zapata-Rivera, & Hegarty, 2014), parents (Kannan, Zapata-Rivera, & Leibowitz, 2016; Zapata-Rivera et al., 2014), students (Goodman & Hambleton, 2004; Vezzu, VanWinkle, & Zapata-Rivera, 2011), and policy makers (Hambleton & Slater, 1997; Underwood, Zapata-Rivera, & VanWinkle, 2010; Wainer, Hambleton, & Meara, 1999).

A Balance of Research and Practice

This volume is divided into two sections, providing a balance of research and practice. The first section includes foundational work on validity issues related the use and interpretation of test scores, design principles drawn from related areas including cognitive science, human-computer interaction, and information visualization and research on presenting particular types of assessment information to various audiences (e.g., subscores, growth, and measurement error information). The second section provides a select compilation of practical applications of designing and evaluating score reports in real settings. In aggregate, the papers describe current work on what assessment information to present and how to present it to particular audiences for various purposes. Altogether, this volume provides interested readers with a unique source of current research and applications in the area of score reporting.

Part 1. Foundational Work

This section includes five chapters. In Chapter 1, Tannenbaum focuses on validity aspects of score reporting. The author makes an argument for clearly communicating score report information, since the ability of the stakeholders to make appropriate decisions is dependent, in part, on the relevance and accuracy of the assessment results, and the ability of the stakeholders to understand the reported information in the way intended. This chapter highlights key concepts and practices that are intended to support the validity of score reports. The author elaborates on sources of validity evidence and their implications for score reporting, describes strategies to build alignment between tests and score reports, and provides guidelines on practices for developing score reports.

In chapter 2, Hegarty reviews current cognitive science research on how people understand visualizations of quantitative information. She described cognitive models of visualization comprehension focusing on the roles of perception, attention, working memory, and prior knowledge. Principles of data-visualization design that take into account the properties of the displays and the individuals are discussed. Finally, the author discusses research on comprehension of test scores and the implications for the design of test score representations for different stakeholders.

In chapter 3, Sinharay, Puhan, Haberman, and Hambleton focus on the quality of subscores. The authors discuss current methods to evaluate when subscores satisfy the professional standards on the reliability, validity, and distinctness of subscores (AERA, APA, & NCME, 2014). Finally, the authors discuss alternative approaches to subscores for the case in which reporting of subscores is not warranted and provide general recommendations on communicating subscores.

Chapter 4 by Zenisky, Keller, and Park addresses the issue of reporting student growth. The authors describe current work in this area and elaborate on the role that this type of information is playing recently in informing high-stakes educational decisions. The authors present results from a small-scale study to evaluate understanding of several common growth reporting display strategies. Finally, the authors elaborate on implications of the result for reporting student growth and the need for additional research in this area.

In chapter 5, Zapata-Rivera, Kannan, and Zwick review work on measurement error and on communicating measurement error information to teachers and parents. They focus on analyzing the processes of designing and evaluating score reports taking into account to knowledge, needs, and attitudes of the target audience. They elaborate on the need to consider teachers and parents as different audiences and examine the potential for using similar research methods and materials when conducting research with these audiences. Finally, they provide

recommendations for this area that include the need for targeted research on (a) effectively communicating information in score reports, (b) investigating both user preferences and user comprehension, and (c) evaluating instructional tools to support understanding and appropriate use of assessment information with particular audiences.

Part 2. Practical Applications

This section covers the last five chapters. These chapters include score reporting work in various contexts: large-scale assessment programs in K-12, credentialing and admissions tests in higher education, using reports to support formative assessment in K-12, applying learning analytics to provide teachers with class- and individual-level performance, and evaluating students' interpretation of dashboard data.

In chapter 6, O'Donnell and Sireci review current research and practices in score reporting in assessments for credentialing and admissions purposes. The authors review frameworks for the development and evaluation of score reports, offer perspectives on validity issues regarding the appropriate communication of assessment results so the intended purposes of the tests can be fulfilled, and any potential negative consequences can be minimized. The authors conclude by discussing the importance of having the stakeholders' needs in mind when designing score reports and interpretive materials for admissions and credentialing programs.

Chapter 7 by Slater, Livingston, and Silver focuses on score reporting issues for large-scale testing programs. The authors discuss the score report design process, which starts by assembling an interdisciplinary team of experts. The steps of an iterative score report design process are described. These steps include: gathering information about the test and the scores to be reported; creating a schedule; creating report design prototypes or "mockups"; getting the client's reactions to the mockups and revising them accordingly; conducting usability testing or focus groups to get reactions to the mockups from potential users of the report; revising the mockups based on feedback from users; and making a final choice and getting approval from the client. The authors emphasize the need for frequent communication between the client and the design team. Finally, lessons learned on applying this process are provided.

In chapter 8, Brown, O'Leary, and Hattie describe various principles of effective report design derived from decades of empirical and theoretical research. These principles emphasize the utility of reports in terms of having a clear purpose and explicit guidance for interpretation and subsequent action, and the clarity of design, guidance, displays, and language used in the reports. An on-line teaching and learning system, the Assessment Tools for Teaching and Learning system (asTTle), that has been deployed in New Zealand's schools is used to illustrate these design principles. The system allows teachers to create tests for their students and interact with a suite of reports including group and individual-level reports. This system was designed to support effective formative assessment practices by teachers.

In chapter 9, Feng, Krumm, and Grover explore the use of learning analytics to make sense of learner performance data collected in various contexts and provide stakeholders with the information they need to support their instructional decisions. Four case studies are presented. The case studies vary in context, subjects, focal constructs, analytical approaches, format of data collected, and student learning tasks. This chapter examines how the needs of practitioners shaped the work, the processes undertaken to develop data products, and the ways in which data products were ultimately used by stakeholders. The authors recognize the need to provide teachers with training to help them understand and make a good use of the information provided by these systems. The authors suggest working directly with practitioners to reduce the complexity of the process and better understanding their needs. The authors recognize the potential of learning analytics to support teachers in doing formative assessment.

Finally, in chapter 10, Corrin reviews the design and evaluation of dashboards. Dashboards are being used to share assessment information and provide feedback to a variety of stakeholders. The author examines how data are presented to students and the intended uses of interactive dashboards. The author reviews the literature on score reporting and identifies insights that can be used to inform the design and evaluation of dashboards. Two case studies of dashboard use in higher education are presented, each profiling important design elements of educational dashboards and the methods of evaluation adopted. The chapter concludes with a discussion of issues for future research in this area.

Final Remarks

These chapters provide a good account of the issues researchers in the area are currently exploring. Frameworks for designing and evaluating score reports, guidelines, lessons learned, and insights for future research are discussed. Both foundational research and practical applications are covered. The following questions were provided to authors and served as general guidance for writing their chapters:

- How are target audience characteristics such as knowledge, needs, and attitudes taken into account when designing their score reports/report systems?
- How are their score reports/report systems intended to be used?
- What kinds of claims about student knowledge, skills, and abilities are made and what data are used to support those claims?
- What kinds of support mechanisms are implemented in order to ensure appropriate use of score report information by educational stakeholders?

The chapters highlight the importance of clearly communicating assessment results to the intended audience to support appropriate decisions based on the original purposes of the assessment. Support may take the form of a clean and simple graphical design that clearly answers the main concerns of the audience, interpretive materials, on-line help, video tutorials, interactive learning materials, and professional development. As more technology-rich, highly interactive assessment systems become available, the more important it is to keep in mind that the information provided by these systems should support appropriate decision making by a variety of stakeholders. Many opportunities for research and development involving the participation of interdisciplinary groups of researchers and practitioners lie ahead in this exciting field.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*, 145–220.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Hambleton, R., & Zenisky, A. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (pp. 479–494). Washington, DC: American Psychological Association.

Hattie, J. (2009, April). *Visibly learning from reports: The validity of score reports*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA.

Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

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

Kannan, P., Zapata-Rivera, D., & Leibowitz, E. (2016, April) *Evaluating validity of score reports with diverse subgroups of parents*. Paper presented at the meeting of the National Council on Measurement in Education, Washington, DC.

Rankin, J. G. (2016). *Standards for reporting data to educators: What educational leaders should know and demand*. New York, NY: Routledge.

Underwood, J. S., Zapata-Rivera, D., & VanWinkle, W. (2010). *An evidence-centered approach to using assessment data for policymakers* (Research Report 10–03). Princeton, NJ: Educational Testing Service.

Vezzu, M., VanWinkle, W., & Zapata-Rivera, D. (2011, April). *Designing and evaluating an interactive score report for students*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Wainer, H. (2014). Visual revelations: On the crucial role of empathy in the design of communications: Genetic testing as an example. *Chance*, *27*, 45–50.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36*, 301–335. doi:10.1111/j.1745-3984.1999.tb00559.x

Zapata-Rivera, D., & Katz, I. (2014). Keeping your audience in mind: Applying audience analysis to the design of score reports. *Assessment in Education: Principles, Policy & Practice*, *21*, 442–463.

Zapata-Rivera, D., & VanWinkle, W. (2010). *A research-based approach to designing and evaluating score reports for teachers* (Research Memorandum 10–01). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., Vezzu, M., Nabors Olah, L., Leusner, D., Biggers, K., & Bertling, M. (2014, April). *Designing and evaluating score reports for parents who are English language learners*. Paper presented at the meeting of the American Educational Research Association, Philadelphia, PA.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*(3), 215–229. doi:10.1080/10627197.2016.1202110

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26. doi:10.1111/j.1745-3992.2012.00231.x

Zwick, R., Sklar, J., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice*, *27*(2), 14–27.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*(2), 116–138.

Part I
Foundational Work

1
Validity Aspects of Score Reporting

Richard J. Tannenbaum

Educational tests are intended to capture and provide information about students' content knowledge, skills, competencies, thought processes and response strategies, personalities, interests, and so on. Of course, no one test captures all such valued information. In any case, some form of a score report is provided to the information-users (e.g., teachers, counselors, school-level and state-level administrators, parents, and students) so that they may act on the reported scores and any accompanying information. The score report is the bridge between the information captured by the test and the decisions or actions of the information-users (Zapata-Rivera & Katz, 2014). A score report that is not well-aligned with the test is of little value; similarly, a score report that is well-aligned, but not communicated to users in a way understandable to them is of little value. Stakeholders cannot make reasonable decisions or take reasonable actions from information that they do not satisfactorily understand, no matter how accurate that information may be in reality.

Proper understanding and use of information is central to the concept of validity: "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (*Standards for Educational and Psychological Testing*, 2014, AERA, APA, NCME, p. 11). Validity is about the extent to which inferences from test scores are appropriate, meaningful, and useful (Hubley & Zumbo, 2011).

This chapter reinforces the growing recognition that score reports and the interpretation of their meaning are part of the overall argument supporting test validity (Maclver, Anderson, Costa, & Evers, 2014; O'Leary, Hattie, & Griffin, 2017). The criticality of score reports in this regard was nicely summarized by Hambleton and Zenisky (2013): "Quite simply, reporting scores in clear and meaningful ways to users is critical, and when score reporting is not handled well, all of the other extensive efforts to ensure score reliability and validity are diminished" (p. 479).

Subsequent sections of the chapter will introduce models and approaches that support opportunities for better alignment between what the test is intended to measure and the score report, how to build meaningful score reports, and validity-related questions and validation methods to evaluate the effectiveness of score reports. Sources of validity evidence for test scores and how these sources relate to score reports and their interpretation are highlighted first.

The *Standards for Educational and Psychological Testing* [*Standards*] (2014) discuss five sources of validity evidence. Each source reflects a different focus, and depending on the intended test score use, one or two sources may take priority. It is not the case that each and every source must be applied to each and every score use. The source of validity evidence should be aligned with the purpose of the test and offer backing for the intended use of the test scores.

Evidence Based on Test Content

This source addresses the extent to which the content of the test reflects the content domain the test is intended to represent. Essentially, the notion is that a test is really a sample of a much larger domain of knowledge, skills, abilities, judgments, and so on that is of interest. A one- or two-hour test likely cannot cover everything of interest. Even a test battery as extensive as that required to become a certified accountant, which includes four tests each covering a specific domain, with each test being four-hours long, still only represents a sample of what accountants are expected to know and be able to do (www.aicpa.org/BecomeACPA/CPAExam/Pages/default.aspx). It therefore becomes important that there is evidence that a test is a reasonable reflection of the larger domain or domains of interest. The transfer to score reports is that there should be evidence that the reported information is aligned with the test content, and presented in a way that is understandable to the stakeholders. The score report should not introduce test-irrelevant information, but should be a faithful reflection of what the test measures and how the test taker(s) performed on the test, which may include areas to improve upon and the identification of resources to assist in that regard.

Evidence Based on Response Processes

This source of evidence becomes more important when the test scores are intended to reflect the strategies, approaches, or cognitive processes test takers are using to address the test items or tasks. Evidence supporting intended response processes may come, for example, from students verbally reporting out either how they are solving the task or how they solved the task (cognitive laboratories or think-alouds, e.g., Leighton, 2017); from maintaining keystroke logs, for example, to understand writing strategies (e.g., Leijten & Van Waes, 2013); and or from eye-tracking data to document and evaluate where on a task students are paying more and less attention (e.g., Keehner, Gorin, Feng, & Katz, 2016). When applied to score reporting, evidence should support that the score-report users are attending to the more relevant or salient features of the report, and interpreting that information as intended. How easily users locate reported information and focus on that information is closely related to how well the score report is designed and organized (Hambleton & Zenisky, 2013).

Evidence Based on Internal Structure

One focus here is on evidence that substantiates the intended structure of the test. Structure in this context means the number of dimensions or constructs the test was designed to address. For example, if a test was intended to measure reading comprehension and reading fluency, there should be evidence supporting this. Evidence of internal structure may also relate to assuring that individual items on a test do not function differently by subgroups of test takers (e.g., boys and girls; African American test takers and White test takers, or that the test scores, overall, have the same meaning for subgroups of test takers (Sireci & Sukin, 2013)). When applied to score reporting, evidence should confirm that stakeholders recognize the intended relationship among the information reported; for example, how a reported range of scores corresponds to

the width of a band on a figure in the report. Further, evidence should support that subgroups of stakeholders understand the same reported information in the way intended. If, for example, the same score report is intended to be shared with native and non-native speakers of English, the reported information should be accessible to both groups. The report should not include language that is unnecessarily complex and therefore less understandable by non-native speakers.

Evidence Based on Relationships to Other Variables

This source of evidence is most applicable when test scores are expected to be related to another measure or an outcome. One example of this is when a test is designed to predict the likelihood that high school students may be successful in their first year of college (as measured by their grade-point average). In the context of score reporting, evidence could take the form of comparing how closely the level of students' competency expressed on the score report is to teachers' evaluations of those students' competencies. A convergence between the report and teacher evaluation would be confirmatory evidence. A disparity between the two sources would be more difficult to interpret, as it could reflect inconsistency between how the test has operationalized the content and how the teacher defines and values that same content.

Evidence for Validity and Consequences of Testing

When a test score is used to inform a decision or to guide an action there will be consequences for one or more stakeholders (Geisinger, 2011; Lane, Parke, & Stone, 1998). Some of the consequences are intended and desired. A test, for example, that is intended to support student learning and development, and in fact, does just that, is a positive intended consequence. However, sometimes tests may have unintended, negative consequences (*Standards*, 2014). One well-known example is in the context of K-12 state-wide accountability testing, where teachers may teach only to the material covered by the test and not to the full curriculum (Koretz & Hamilton, 2006; Lane & Stone, 2002). When applied to score reporting, evidence should indicate that stakeholders are acting on the reported information in ways consistent with reasonable expectations, and not making inaccurate interpretations leading to inappropriate decisions. Such evidence could be collected from surveys of or interviews with stakeholders about how they actually used the reported information, as well as what aspects of the reported information were most and least useful.

A Brief Recap

It may be helpful at this point to recast the information above. Score reports are intended to provide stakeholders with the information they need, in a way that they understand, so that they may reasonably act on that information. There should be evidence that supports that the reported information is interpreted and understood, as intended, and that the decisions and actions based on that interpretation and understanding foster positive, intended consequences and very few, if any, negative consequences. These are important validity-related goals of score reporting. The next section describes general strategies for helping to assure these goals are met, beginning with assuring the alignment between tests and score reports.

A Strategy to Build Alignment Between Tests and Score Reports

Yogi Berra, hall-of-fame catcher and philosopher, once remarked: "If you don't know where you are going, you might wind up someplace else" (http://nypost.com/2015/09/23/35-of-yogi-berras-most-memorable-quotes)—words of wisdom when thinking about score reports. What

do we want to conclude about test takers from their performance on the test? Who will use this information? In addition, what are they expected to do with that information? In other words, where do we want to wind up? If we cannot answers these types of questions, it is not likely that the test, score report, and the actions taken based upon the reported scores will be of much value. There should be a direct and transparent alignment among the test purpose (what it is supposed to accomplish), the test content and format, how responses to the test items or tasks are scored, and how summaries of those scored responses are reported to stakeholders. It is through accurate alignment that there is a much greater likelihood that the stakeholders will act on the reported information in a way that is consistent with what the test was supposed to accomplish. This point is reinforced by the International Test Commission (2014), which emphasizes that

> the test development, scoring, and analysis stages should all take into consideration the final product—the reported interpretation of the scores. In this sense, the underlying aim, or tacit first step of the whole process of test development, is ensuring that the reported score will be properly understood.
>
> (p. 205)

Stated somewhat differently, work backward from what is desired to be communicated in the score report to assure that the test ultimately provides the desired information. This is, in fact, the value of developing what is known as a Prospective Score Report (PSR, Zapata-Rivera, Hansen, Shute, Underwood, & Bauer, 2007; Zieky, 2014).

Prospective Score Reports

A PSR is a mock-up of what the final score report should include and look like. It lays out all the specifications that are to be communicated. "The PSR should be used at the beginning of assessment design, with the reporting measures continually being refined, as necessary, through the remainder of the assessment design process" (Zapata-Rivera et al., 2007, p. 275). In other words, the test and the score report should be developed more or less simultaneously. Once the claims of the test and needed evidence are clearly formulated, the PSR should be developed; and as the test development process moves ahead, implications for the score report need to be documented and represented in changes to the PSR.

However, more often than may be desirable, score reports tend to be considered near the end of the test-development process, perhaps because they occur at the end of the process. That is neither effective nor efficient. There are spatial organizations to score reports as well as technical parameters and constraints. Deciding at the end of the test-development process, for example, to have the score report provide exemplars of students' responses at different score levels may not be feasible, if the related information-technology system has not made prior accommodations for that detailed reporting. That kind of specificity should have been thought about at the beginning of the test-development process, or sometime during the process, so that that expectation could have been factored into the reporting system in a timely manner.

A failure to construct PSRs—waiting until the end to develop score reports—also runs the risk of holding the score report "accountable" for reporting information that was not satisfactorily collected by the test. This is all too common. One instance of this is where a test was not designed to provide sufficient evidence at a subscore level (a score about a specific test section or group of test items), but an influential stakeholder later expects meaningful subscores to be reported. The expectations of the stakeholder should have been included at the very beginning of the test-design stage, when claims were being formed about what to measure and what was to

be concluded about test takers. This way, the test could have been designed to include enough items or tasks to generate the evidence needed to provide meaningful subscores. It may also be the case that constraints on testing time and test fees would not permit this to occur. This would be important to know much sooner than later, so the expectations of stakeholders could be moderated.

Guidelines and Practices for Developing Score Reports

The prior section noted the importance of working backward, and the value of Prospective Score Reports in that regard. This section discusses specific steps that one may follow to construct score reports. First, it is important to note that there are some questions that must be answered from the start, as they have implications for the test and score report; as such, they may be thought of as design claims.

Questions to Consider

One question is will the reported information be used for formative purposes or for summative purposes? A formative use supports learning and development opportunities (Black & Wiliam, 2009). Where are students in need of more focused instruction? A formative-focused reporting of scores, therefore, would include feedback to let the teacher and the students know what content or skills are to be bolstered. Brown, O'Leary, and Hattie (this volume) provide further discussion of score reports used for formative purposes. A summative use is more often associated with outcomes: What have students learned, for example, at the end of the semester? Which students on the state-mandated test are classified as Proficient? In these cases, the score report tends not to include the kinds and level of feedback sufficient to support the current cohort of students learning.

A second question is will the score report be static or dynamic? A static report is more or less the traditional form of a report; there is one "view" of the information offered to stakeholders, and the report may be paper-based or digital. However, there is no opportunity to customize the report to meet varying needs. This does not mean that the information is necessary lacking or that the report is of low quality (Zenisky & Hambleton, 2016). If a static report will meet stakeholder needs, then that is fine. On the other hand, dynamic reports allow stakeholders to interact with the reporting system (to varying degrees) to customize the reported information. Sometimes, the reporting system may include pre-determined views of test-score summaries that can be selected (e.g., scores by different subgroups, line graphs versus pie charts), so the system is partially interactive (Tannenbaum, Kannan, Leibowitz, Choi, & Papageorgiou, 2016). In other instances, the stakeholders may be able to "request" new or additional analyses for inclusion in the report. The most well-known system for reporting group-level scores is the NAEP Data Explorer (www.nces.ed.gov/nationsreportcard/naepdata). Highly interactive systems are not very common, as they are costly and somewhat difficult to develop. For more details about dynamic score reports see Feng, Krumm, Grover, and D'Angelo (this volume).

A third question is will the score report include subscores? As several researchers have noted (e.g., Haberman, 2008; Sinharay, 2010; Zenisky & Hambleton, 2016), increasingly stakeholders are expecting to receive subscore information. This is mainly due to the belief that with subscore information teachers and students will know where additional attention is needed—subscores serve a formative purpose. This is both logical and reasonable. That said, to have confidence in the meaning of a subscore, there needs to be sufficient numbers of items addressing that area. If there are too few, the subscore is not likely to be an accurate reflection of a student's competency in that area; Sinharay (2010) suggests that 20 items may be needed. Further, he notes

that the information in one subscore needs to be sufficiently distinct from other subscores to be meaningful. If different subscores are essentially providing redundant information (are highly correlated with one another), there is little justification for reporting them. Sinharay, Puhan, Hambleton, and Haberman (this volume) provide more information about reporting subscores.

The last question to consider is how to communicate to stakeholders the imprecision associated with test scores? Those of us in the field of measurement readily accept that test scores are not perfectly reliable; and we can communicate with each other in terms of standard errors of measurement, conditional standard errors of measurement, and confidence intervals (Zenisky & Hambleton, 2016; Zwick, Zapata-Rivera, & Hegarty, 2014). However, communicating uncertainty in scores to other stakeholders (e.g., teachers, students, parents, administrators, politicians) in a way that they will understand is not easy. Yet, our professional guidelines (*Standards*, 2014, Standard 6.10) expect that score reports will include explicit information about the measurement error (imprecision) associated with reported test scores. Zwick et al. (2014) illustrate the challenges with presenting this information to teachers and college students. Zapata-Rivera, Zwick, & Kannan (this volume) provide detailed discussion of communicating measurement error to parents as well as teachers.

Steps to Follow to Develop Score Reports

Specific guidelines for developing and evaluating score reports are offered by Hambleton and Zenisky (2013), the *Standards* (2014; e.g., Standards 6.10, 8.7. 8.8, and 9.8), Zapata-Rivera, VanWinkle, and Zwick (2012), and Zenisky and Hambleton (2012). These sources, as may be expected, share much in common; the Hambleton and Zenisky (2013) model serves as the basis for the following discussion.

The first three steps in their seven-step model are likely best implemented concurrently, as they collectively relate to defining the relevant stakeholders and their needs, and gathering examples of existing score reports that may be useful to consider. These three "framing" steps, should occur early in the test conceptualization and design process to better assure that alignment to the test purpose is engineered.

Step 1

This includes the explicit delineation of the purpose of the score report—what needs are the reported information expected to meet? What is the reported information intended to provide? A primary focus of this step is to assure that the score report content accurately reflects what the test set out to accomplish. For example, if the test was intended to be formative, the score report needs to reflect the qualities of feedback important to fulfill that use.

Step 2

This step, as noted, is closely aligned with the first step, in that it focuses on identifying the relevant stakeholders who will rely on the score report to make decisions or to take actions. In fact, Hambleton and Zenisky (2013) note that key stakeholders should be consulted during Step 1 to flesh out the specific needs to be met by the score report. It is during these first two steps that the differing needs of stakeholders will emerge, not only in terms of desired content, but also in terms of delivery mode or level of interactivity expected.

Zapata-Rivera and Katz (2014) expand upon the need to consider stakeholders through their recommendation of conducting an audience analysis. One aspect of the analysis is clarifying audience needs, which is closely related to what was already discussed: identifying score-user

goals and what they intended to do with the reported scores. The second aspect addresses the audience's test-relevant knowledge and general test literacy. What does the audience know about the test, about the test-taking population, and about general measurement principles? Less test-savvy audiences, for example, may need more explanations and supporting information to make proper score-based interpretations. As noted earlier, while measurement experts may have little difficulty understanding standard errors, this statistic may be confusing to less test-savvy audiences. The last aspect of audience analysis focuses on audience attitudes or preconceived notions or biases about testing (consider the recent opt-out movement regarding K-12 accountability testing; Bennett, 2016). This includes what the audience expects of test takers, and about how much emphasis they will likely place on the reported information or perceived time they have to consider and act upon the information.

Step 3

The focus here is on reviewing existing samples of score reports (conducting a review of published or otherwise retrievable examples) to see if some of the needs and functionalities revealed during the first two steps have already been addressed. The goal here is to take advantage of representations, or portions thereof, that may be useful and to avoid repeating others' mistakes.

Step 4

This step involves pulling together the information collected from the prior steps to develop one or more Prospective Score Reports (PSRs, Zieky, 2014). It is during this stage that care must be taken not only to build prototypes that appear to meet stakeholder data needs and the objectives of the test, but also to consider issues of design and presentation clarity, scaffolding to support proper interpretation, and, accessibility, more generally. As Hambleton and Zenisky (2013) note, this stage may require the involvement of many different expert groups, such as content specialists, measurement experts, user interface experts, information technology specialists, graphic designers, cognitive scientists, and others, depending on the intended form and functionality of the reports. Given the role of score reports in the overall validity argument for a test, the investments made here are worthwhile.

Steps 5 and 6

Once PSRs are available, it is important to gather data about the extent to which the score reports are communicating the information to stakeholders as intended (Step 5) and to revise accordingly (Step 6). These steps should be considered iterative, as more than one round of evaluation and revision will likely be needed. During the Step 5 evaluation, data should be collected about stakeholders' reactions to the how the information is displayed—is the information visually accessible? This would include questions, for example, about readability, preferences for different data presentations, and aspects of the score report that may not be attended to as expected. Data should also be collected about stakeholders' understanding of the information. Do stakeholders interpret the information as intended? Are they able, for example, to describe accurately what the scores mean from the information displayed, or do they misinterpret certain pieces of information?

Hambleton and Zenisky (2013) produced more than 30 questions that may be considered as part of the development and evaluation of prospective score reports (as such, these questions may inform Steps 4 through 6). The questions address eight key areas: Needs Assessment (i.e., Does the score report reflect the expectations of stakeholders?); Content—Report and

Introduction and Description (e.g., Is the purpose of the test described?); Content—Scores and Performance Levels (e.g., Is information about proper and improper use of score classifications provided?); Content—Other Performance Indicators (e.g., If subscores are reported, is information about imprecision also included?); Content—Other (e.g., Does the score report provide contact information if questions arise?); Language (e.g., Is the report free of technical jargon that may be confusing?); Design (e.g., Is the report clearly and logically laid out to facilitate readability?); and Interpretive Guides and Ancillary Materials (e.g., If an interpretive guide is provided, what evidence is there that it is understood by stakeholders)?

The International Test Commission (2001, 2014) also provides guidance regarding the development and evaluation of score reports. This includes, for example, using a reporting structure and format that is aligned with the test purpose; assuring that the technical and linguistic levels of the reported information are appropriate for the score-report users; and providing sufficient scaffolding to support proper interpretation. Hattie (2009) and Van der Kleij, Eggen, and Engelen (2014) discuss other considerations when designing score reports.

Different methods may be used to collect the needed feedback (validity evidence). Frequently used approaches include interviews (including think-aloud approaches), eye tracking, focus groups, surveys, and more formal experimental designs (International Test Commission, 2014; Hambleton & Zenisky, 2013; Zenisky & Hambleton, 2016; *Standards*, 2014). Tannenbaum et al. (2016), for example, provided different versions of mock score reports about English learners' English proficiency to a committee of teachers of English as a Second Language. During the focus-group meeting the teachers were asked to consider what aspects of the mock reports were more or less useful to inform possible instruction, and what new or additional information would be helpful in that regard. Their recommended modifications were used to revise the mock ups.

Kannan, Zapata-Rivera, and Leibowitz (2016) successfully utilized one-on-one think-aloud protocols with parents (of varying levels of English proficiency and educational backgrounds) to gather evidence of their understanding of mock student-focused score reports. This approach enabled Kannan et al. to better understand those aspects of the reported information (e.g., errors) that were especially problematic for parents who are themselves English learners or who have comparatively lower levels of education (e.g., high school or some post-secondary education).

Zwick et al. (2014) conducted two experimental studies whereby teachers and then college students were randomly assigned to one of four conditions, with each condition depicting different ways of reporting score imprecision. In each study, the participants completed comprehension and display-preference questionnaires, as well as a background questionnaire to document, in part, participants' prior knowledge of measurement and statistics. The experimental design structure enabled Zwick et al. to uncover misconceptions about interpreting error and confidence bands, as well as to understand the relationship between familiarity with measurement and statistics and the ability to comprehend more technically sophisticated score reports.

Step 7

This final step addresses the importance of monitoring the usefulness of the score report once the test becomes operational. Questions to consider here include, for example: What decisions are stakeholders actually making from the test scores? How are the test scores being used in conjunction with other information? What reported score information is most or least helpful? What information might be important to add to the report in a future version? It is also desirable at this stage to try and gather evidence of any unintended, negative consequences resulting from the interpretation and use of the reported scores. Unintended consequences may arise

from misunderstanding the meaning of the scores, including attributing more meaning to the scores than is justified. The results from this step may suggest the need to make revisions to the report structure, content, or supporting materials to foster proper interpretation and use.

Conclusions

No matter how well a test is conceptualized, designed, and implemented, if the scores reported are not readily understandable to stakeholders, all the prior hard work and effort may have been in vain. A correct understanding and interpretation of score reports is a prerequisite for stakeholders to make reasonable decisions (Van der Kleij et al., 2014). "The interpretability of score reports . . . is of the utmost significance and is fundamental in claims about validity" (O'Leary et al., 2017, p. 21).

Evidence of validity is one of the central tenets of quality testing. However, as Hattie (2014) noted, too little attention has traditionally been devoted to including score reports in the overall validity argument of testing; although that is changing, as this chapter, in part, strived to highlight. There are two key messages from this chapter: First, score reports are, in fact, integral to assuring the validity and utility of test score interpretation and use; and second, score reports must be conceptualized and developed as early in the test-development process as possible to better assure alignment to the test purpose and testing objectives, to better assure we end up where we had intended, and not someplace else.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bennett, R. E. (2016). *Opt out: An examination of issues* (Research Report 16–13). Princeton, NJ: Educational Testing Service.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31.

Geisinger, K. F. (2011). The future of high-stakes testing in education. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in k-12 settings* (pp. 231–248). Washington, DC: American Psychological Association.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol 3. Testing and assessment in school psychology and education* (pp. 479–494). Washington, DC: American Psychological Association.

Hattie, J. (2009, April). *Visibly learning from reports: The validity of score reports*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Hattie, J. (2014). The last of the 20th-century test standards. *Educational Measurement*, *33*, 34–35.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*, 219–230.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*, 93–114.

International Test Commission. (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*, 195–217.

Kannan, P., Zapata-Rivera, D., & Leibowitz, E. A. (2016, April). *Evaluating the validity of score reports with diverse subgroups of parents*. Paper presented at the annual meeting of the National Council for Measurement in Education, Washington, DC.

Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2016). Developing and validating cognitive models in assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 75–101). New York, NY: Wiley Blackwell Press.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education and Praeger.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement*: *Issues and Practice*, *17*, 24–27.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment for accountability programs. *Educational Measurement*: *Issues and Practice*, *21*, 23–30.

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. New York, NY: Oxford University Press.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, *30*, 358–392.

Maclver, R., Anderson, N., Costa, A. C., & Evers, A. (2014). Validity of interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment*, *22*, 149–164.

O'Leary, T. M., Hattie, J. A., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice*, *36*, 16–23.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174.

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger (Editor-in-chief), *APA handbook of testing and assessment in psychology* (Vol. 1, pp. 61–84). Washington, DC: American Psychological Association.

Tannenbaum, R. J., Kannan, P., Leibowitz, E., Choi, I., & Papageorgiou, S. (2016, April). *Interactive score reports: A strategic and systematic approach to development*. Paper presented at the meeting of the National Council for Measurement in Education, Washington, DC.

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the computer program LOVS: A redesign study. *Studies in Educational Evaluation*, *43*, 24–39.

Zapata-Rivera, D., Hansen, E., Shute, V. J., Underwood, J. S., & Bauer, M. (2007). Evidence-based approach to interacting with open student models. *International Journal of Artificial Intelligence in Education*, *17*, 273–303.

Zapata-Rivera, D., & Katz, I. (2014). Keeping your audience in mind: Applying audience analysis to the design of score reports. *Assessment in Education: Principles, Policy & Practice*, *21*, 442–463.

Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL™ teachers* (Research Memorandum 10–01). Princeton, NJ: Educational Testing Service.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*, 21–26.

Zenisky, A. L., & Hambleton, R. K. (2016). A model and good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 585–602). New York, NY: Routledge.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*, 79–87.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*, 116–138.

2

Advances in Cognitive Science and Information Visualization

Mary Hegarty

Information visualizations are important forms of human communication. They are used to convey many types of data, including the results of scientific experiments (Belia, Fidler, Williams & Cumming, 2005), risks of disease (Garcia-Retamero & Cokely, 2013), weather forecasts (Lowe, 1996), and test reports (Goodman & Hambleton, 2004) to a variety of stakeholders including domain experts, policy analysts, and the general public. This chapter provides an overview of current cognitive science research on how people understand visualizations of quantitative information. In this chapter I first review theories of how visualizations "augment cognition" and describe cognitive models of comprehension of data graphs, including the contribution of perception, attention, working memory and prior knowledge to comprehension. Next, I discuss how different visualizations of the same data sometimes convey different messages and how individuals with various levels of expertise and prior knowledge sometimes interpret displays differently. This research will be used to argue for principles for the design of effective visualizations that have emerged from both theory and empirical studies of comprehension of information visualizations, taking into account the fact that different displays may be more and less effective for different uses and for different individuals. Finally, I review recent research on comprehension of test scores in the light of cognitive science theories and empirical research, and derive implications for the design of test scores for different stakeholders such as parents, teachers, and educational administrators.

The broad category of "information visualizations" includes diagrams, maps, and graphs, that is, external visuo-spatial displays that can represent objects, events, numerical data, or more abstract information. Visualizations can be categorized based on the relation between the representation and what it represents, and how space is used to convey meaning. One category of visual displays consists of *iconic* displays such as pictures, maps, and drawings. In iconic displays, space on the page represents space in the world and the properties displayed (shape, color, etc.) are also visible properties of what is represented (e.g., the curve of a road on a road map, the color of blood in a diagram of human heart). Displays in second category, *relational displays*

Author Note: Preparation of this paper was supported in part by grant 1212577 from the National Science Foundation

are metaphorical in that they represent entities that do not have spatial extent or visible properties (e.g., when an organization chart shows the hierarchy of positions in a business, or a graph shows the price of stocks over time). In these displays, visual and spatial properties represent entities and properties that are not necessarily visible or distributed over space. Visual-spatial variables, such as color, shape, and location can be used to represent any category or quantity. The term information visualization is usually used to refer to this type of display (Card, Mackinlay, & Schneiderman, 1999). In contrast to iconic displays, which can be traced back to ancient cave drawings, these types of displays are a relatively recent invention. Specifically the invention of the data graph is attributed to Playfair in the 18th century (Wainer, 2005).

Why Visualize? The Advantages of Visualizing Data

Information visualizations are often said to enhance or "augment" cognition (Card et al., 1999; Larkin & Simon, 1987; Scaife & Rogers, 1996). Cognitive scientists and other specialists have proposed a number of ways in which presenting data graphically can enable people to understand or reason about the data. First, information visualizations store information externally, freeing up working memory for other aspects of thinking (Card et al., 1999; Scaife & Rogers, 1996). Second, they can organize information by indexing it spatially, reducing search and facilitating the integration of related information (Larkin & Simon, 1987; Wickens & Carswell, 1995). Graphs organize entities by placing them in a space defined by the x and y axes. As a result, similar entities are visualized as close together. For example, in Figure 2.1, which shows a scatter plot relating a (fictional) sample of children's test scores to their parent's income, the dots representing children with similar levels of test scores and parental income are located close together in the display. Graphs can also allow the offloading of cognitive processes onto

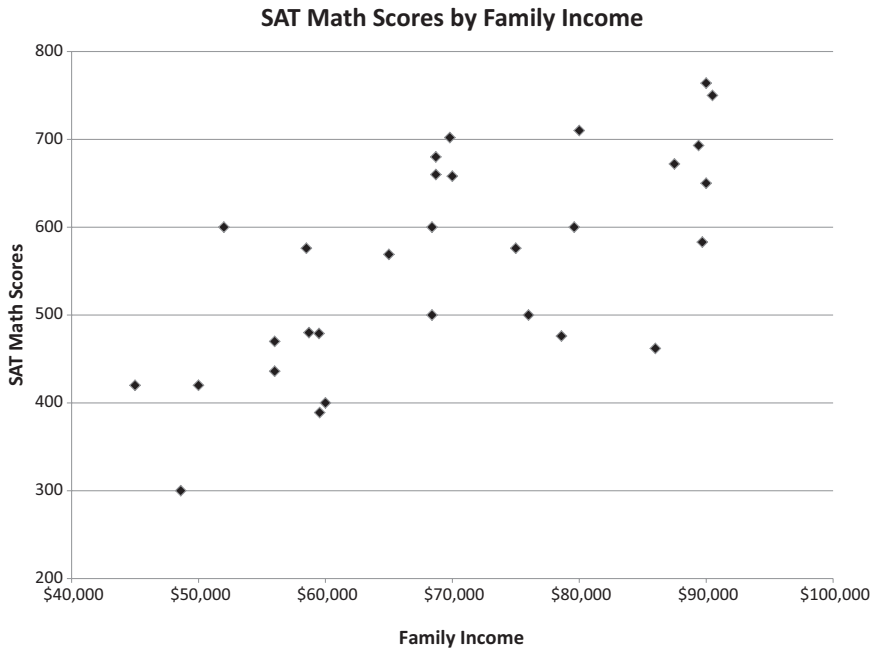


Figure 2.1 Scatter plot showing the relationship of SAT Math scores to family income for a fictitious class of 30 students.

automatic perceptual processes (Scaife & Rogers, 1996). When non-visual data are mapped onto visual variables, patterns often emerge that were not explicitly built in, but which are easily picked up by the visual system, for example, when a line in a graph reveals a linear relationship between variables (Shah, Freedman, & Vekiri, 2005). They can enable complex computations to be replaced by simple pattern recognition processes.

Cognitive Processes in Using Visual-Spatial Displays

Although visual-spatial displays can enhance thinking in many ways, this does not mean that their use is necessarily easy or transparent. Figure 2.2 presents a model of comprehension of information visualizations, which is adapted from other cognitive models of how people understand visuo-spatial displays such as graphs (Carpenter & Shah, 1998; Hegarty, 2011; Kriz & Hegarty, 2007; Pinker, 1990). According to these models, graph comprehension involves a complex interplay between display driven (bottom-up) and knowledge driven (top-down) processes. First, the visual system senses the basic visual features of the display, such as color and shape and encodes these features to construct an internal representation of the visualization itself. In complex visualizations, not all features are necessarily encoded, and which features are encoded depends on attention, which might be directed by the viewer's goals and expectations (top-down processing) or what is salient in the display (bottom-up processing). For example, one difficulty in display comprehension might arise if the viewer is distracted by highly salient but task-irrelevant information such as a picture in the background of a graph, so that the viewer fails to encode the critical information, although it is presented.

In addition to basic perceptual, attentional, and encoding processes, which construct a visual representation of the display, the user of an information visualization typically has to apply knowledge to construct a meaningful representation of the information presented in the display. This can include knowledge of the conventions of the display, for example, that the independent variable in an experiment is typically represented on the x axes and the dependent variable is represented on the y axis of a data graph (Gattis & Holyoak, 1996), or the meaning of error bars in a graph (Cumming & Finch, 2005). This type of knowledge is often referred to as a graph schema (Pinker, 1990; Ratwani & Trafton, 2008). Understanding a graphic can also include making further inferences based on domain knowledge (for example understanding whether a student's test score is in the normal range for his or her grade level) or visual-spatial processes (e.g., detection of a linear increase in test scores over time) so that the resulting internal representation comes to contain information that is not presented explicitly in the external display.

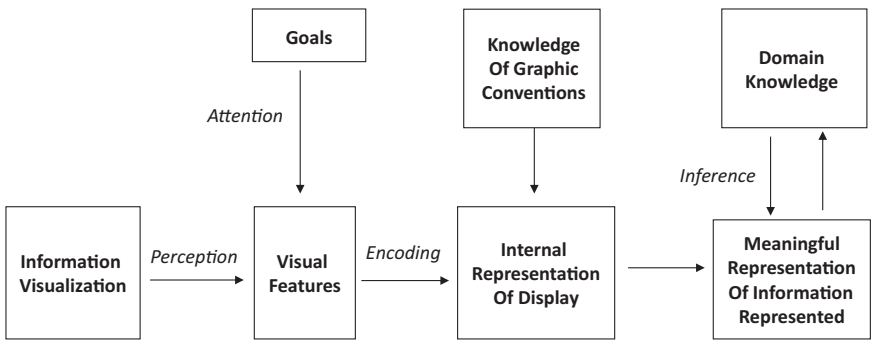


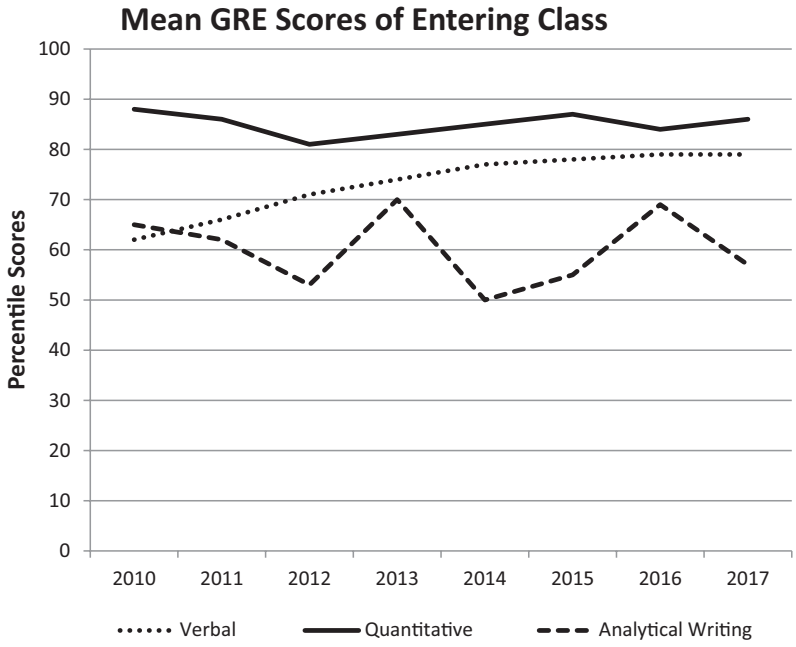
Figure 2.2 Schematic overview of the different representations (indicated by boxes) and processes (indicated by arrows) involved in understanding an information visualization.

All Visualizations Are Not Equal

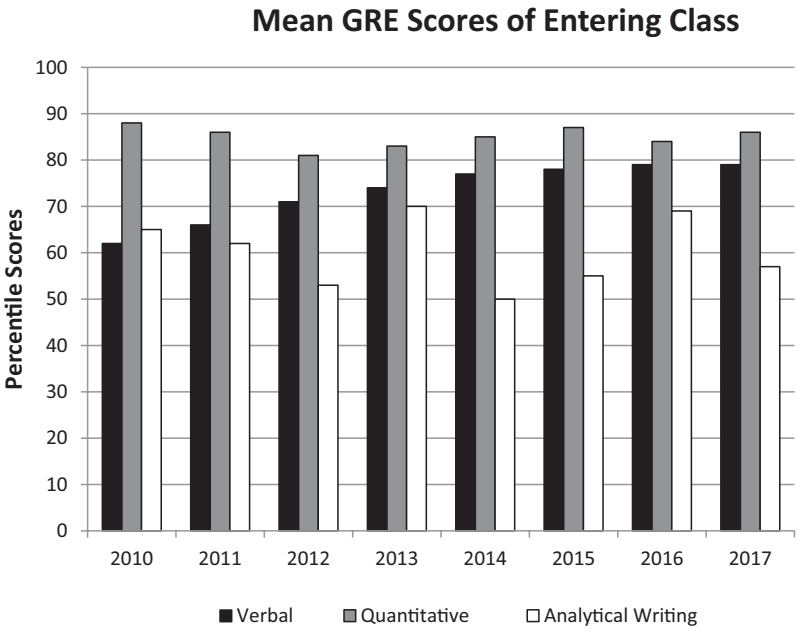
A prominent conclusion from research on comprehension of graphs and other visual displays is that there is no such thing as a "best" visualization of a given data set, independent of the task to be carried out with this display or independent of the user of the display. First, visual displays are used for many different purposes such as recording and storing information, serving as computational aids, data exploration, and conveying information to various stakeholders. Visualizations that are effective for one purpose (e.g., data exploration) might not be effective for another (e.g., communication to the general public). Second, the way in which information is displayed graphically can have powerful effects on how it is interpreted and processed, providing evidence for bottom-up influences of display design. People interpret the same data differently, depending on whether they are presented in pie charts or bar graphs (Cleveland & McGill, 1984; Simkin & Hastie, 1986), bar graphs or line graphs (Shah, Mayer, & Hegarty, 1999), and which variables are assigned to the x and y axes (Gattis & Holyoak, 1996; Peebles & Cheng, 2003; Shah & Carpenter, 1995). Some of these effects can be traced to the Gestalt principles of perceptual organization, which determine which elements of displays are grouped, and can be compatible or incompatible with the tasks to be carried out with a display. For example, line graphs facilitate comparisons for the variable plotted on the x axis (time in Figure 2.3a) because the lines group data points as a function of this variable, reflecting the Gestalt principle of good continuation (Shah & Freedman, 2011). In contrast, bar graphs facilitate comparisons between the variables shown in the legend (GRE sub-score in Figure 2.3b), because the bars comparing data points with respect to this variable are closer, reflecting the Gestalt principle of proximity. Thus, in Figure 2.3a it is relatively easy to see that the verbal sub-score increased over the period shown or that the writing sub-score was more variable over time. In contrast Figure 2.3b seems to emphasize the differences between the three GRE scores, for example that for this department, the quantitative scores are highest, and the writing scores are lowest.

Displays that are effective for one task may be ineffective for another. For example, tables are better than graphs for communicating specific values whereas graphs are better than tables for conveying trends in data (Gillan, Wickens, Hollands, & Carswell, 1998). Pie charts are an interesting case in point. During the 20th century, statisticians developed a strong bias against the pie chart preferring divided bars to display proportions. However, careful experiments indicated that for some tasks, pie charts are as effective as divided bar charts, and for other tasks they are actually more effective (Simkin & Hastie, 1987; Spence & Lewandowsky, 1991). Simple judgments (e.g., comparing the proportions of two entities) were slightly more effective with bar graphs, but complex comparisons (e.g., comparing combinations of entities) were more efficient with pie charts (Spence & Lewandowsky, 1991). For example, in Figure 2.4a it is easier to see that about half of the class got either A or B grades whereas in Figure 2.4b it is easier to see that the proportion of B and C grades was approximately equal.

An important current issue in information visualization is how to best represent data uncertainty in graphical displays (Kinkeldey, MacEachren, & Schiewe, 2017; Spiegelhalter, Pearson & Short, 2011). A common method of presenting information about uncertainty is to use error bars to show confidence intervals, but error bars are often misunderstood, even by researchers who use them in interpreting their own data (Belia, Fidler, Williams, & Cumming, 2005). More novice participants also have misconceptions about error bars, such as the assumption that the estimate is equally likely to be anywhere within the error bars and not at all likely to be outside the error bars (e.g., Correll & Gleicher, 2014; Zwick, Zapata-Rivera, & Hegarty, 2014). Because of misconceptions in interpreting error bars, researchers have advocated alternative forms of uncertainty visualizations, including violin plots and faded representations that show graded probability of estimates with more distance from the mean (Correll & Gleicher, 2014; Cumming, 2007).

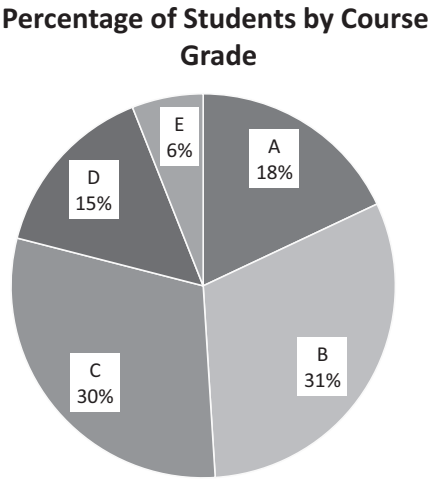


(a) Line Graph

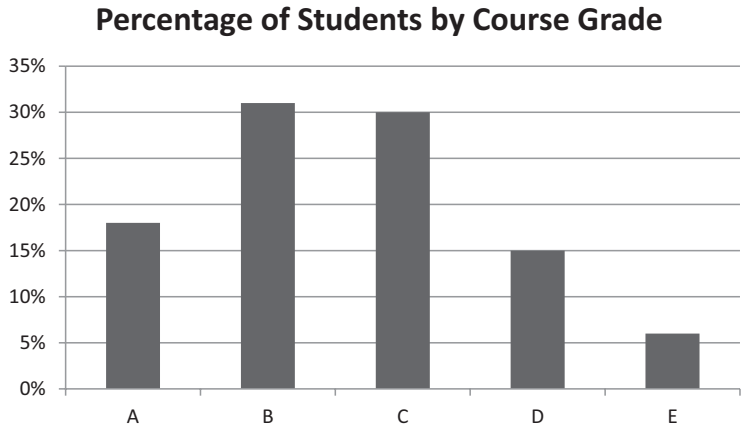


(b) Bar Graph

Figure 2.3 Line and bar graphs showing mean percentile scores of students admitted to a fictitious science department over an eight-year period. The data displayed in the two graphs are identical and show mean percentile scores over time for the verbal, mathematics, and writing subtests of the GRE.



(a) Pie Chart



(b) Bar Graph

Figure 2.4 Pie and bar charts showing distribution of grades in a fictitious college course.

All Consumers of Visualizations Are Not Equal

Comprehension of graphical displays is also influenced by knowledge. Experts and novices attend to different aspects of visual displays and extract different information from these displays. These top-down effects of knowledge on graphics comprehension can be separated into effects of knowledge or facility with graphic conventions, knowledge of mathematics and statistics (numeracy) and knowledge of the domain or topic of the graph data

Knowledge of Specific Graphic Conventions

First, understanding the graphic conventions of a display involves understanding *how* the display conveys information, for example the meaning of the axes and which visual variables (color, shading fill patterns etc.) represent each aspect of the data. While some of this information is often included in a figure caption or legend accompanying a graph, more basic information is often assumed. For example, Kozhevnikov and colleagues (Kozhevnikov, Hegarty, & Mayer, 2002) showed simple graphs of motion to undergraduate students who had not taken any physics classes and had been classified as "high-spatial visualizers" and "low-spatial visualizers." The high-spatial visualizers correctly interpreted the graphs. For example, when shown the graph in Figure 2.5, one student described the display as follows: "*At the first interval of time the position is the same: it cannot move. It has a constant velocity at the second interval. It is moving constantly at a constant speed*." In contrast some low-spatial visualizers erroneously interpreted the displays, for example one student interpreted the same graph as follows: "*The car goes constantly and then goes downhill. . . . It does not change its direction. It goes downhill. This is a hill*." In this example, low-spatial visualizers, were subject to the "graph-as-picture" misconception (McDermott, Rosenquist, & van Zee, 1987). Because they did not have a schema for this type of graph, they erroneously interpreted it as a picture.

Another case in point is the cone of uncertainty used to display hurricane forecasts (see example in Figure 2.6). The cone of uncertainty is a forecast display produced by the National Hurricane Center that indicates the current location of a hurricane storm, the storm's projected path (track) over the next three days, and a cone shape surrounding the track line. A basic convention of this graphic is that the width of the cone at any point in time represents the amount of uncertainty in the forecasted location of the storm at that time point (specifically the 67% confidence interval). Expert meteorologists and emergency management personnel know these conventions. However, the uncertainty cone is often presented in news media without an explanation of how it is created, or what is represented by the cone of uncertainty. Broad, Leiserowtiz, Weinkle, and Steketee (2007) reported that people in hurricane-affected regions of the US hold misconceptions about what is represented by this visualization. One misconception was that the cone shows the hurricane getting larger over time. Another was that the hurricane was unlikely to travel outside the region depicted by the cone. Evidence of these misconceptions was also found in a recent laboratory study in which students had to respond to true-false statements about the meaning of the display (Ruginski et al., 2016). For example, 69% who viewed the visualization in Figure 2.6 endorsed a statement indicating that the display shows the hurricane getting larger over time, and 49% endorsed a statement that the damage was not likely to extend beyond the cone. A more recent study indicated that people were less likely to endorse these misconceptions if they first read a description of the display conventions (Boone, Gunalp, & Hegarty, in press).

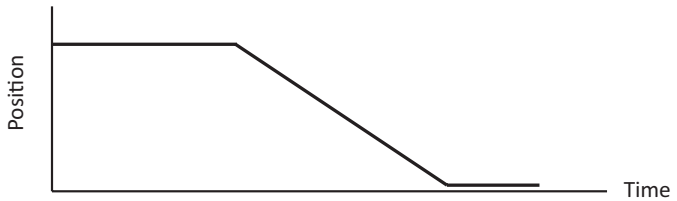


Figure 2.5 Graph of position as a function of time, used by Kozhevnikov et al., 2002.

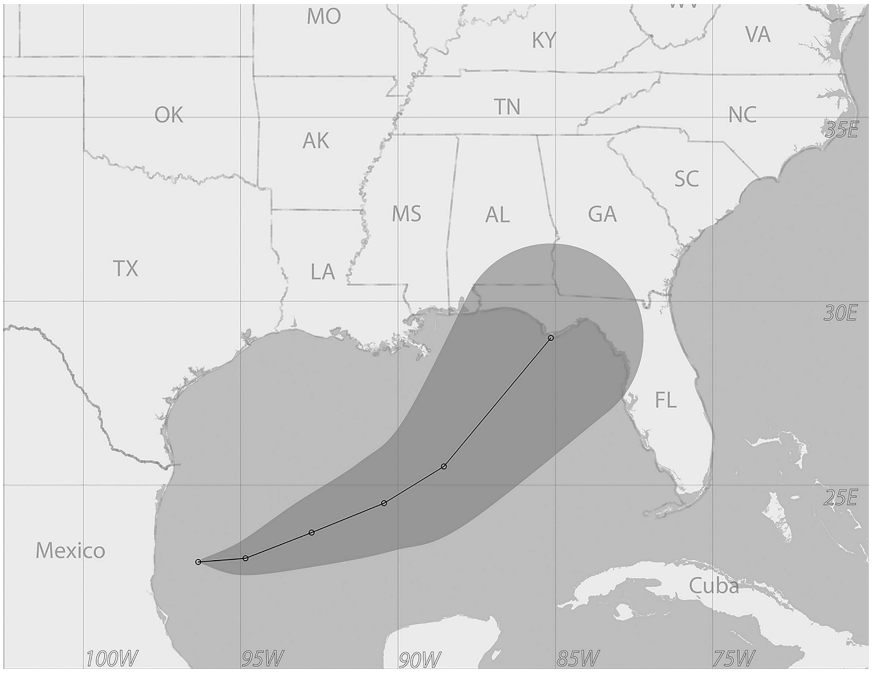


Figure 2.6 Example of a hurricane forecast showing the cone of uncertainty.

Mathematical Knowledge

Comprehension of information visualizations can also depend on numeracy, that is, quantitative or mathematical literacy. An important recent application of information visualization is in communicating medical risks to the general public (Ancker, Senathirajah, Kukafka, & Starren, 2006; Garcia-Retamero & Cokely, 2013). To make medical decisions, patients and their doctors often have to understand the risks of disease, unhealthy behaviors (e.g., smoking), and both the benefits and possible side-effect risks of various medical treatments. However members of the general public have poor understanding of basic numerical and probabilistic concepts necessary to understand risks (Peters, 2012). Researchers in medical decision making have had good success in designing visual aids for communicating medical risks to the general public. For example, they have found that bar charts are effective for comparing magnitude of risk for different groups (e.g., nationalities), line graphs are effective for showing trends over time such as survival curves, and icon displays are good for reducing denominator neglect, that is, the tendency to focus only on the number of people affected by a disease (numerator) and ignore the number of people who could potentially have been affected (denominator) (Lipkus, 2007). However, these visual aids are not equally effective for all individuals. Researchers in medical decision making have developed measures of basic *numeracy* related to medical risks (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) and *graphicacy* or the ability to interpret common graph formats such as bar graphs and icon displays (Galesic & Garcia-Retamero, 2011). They found that people with high numeracy often have good understanding of medical risks, regardless of whether the data are presented numerically or graphically while graphic displays are more effective than numerical descriptions for individuals with low numeracy. However graphic displays were not effective for all low-numerate individuals. Specifically, individuals with poor numeracy but relatively good graphicacy had good comprehension of

visual aids for medical decision making whereas the visual aids were ineffective for individuals with poor graphicacy and numeracy (Galesic & Garcia-Retamero, 2011).

Domain Knowledge

Content or domain knowledge about the topic of a graphic can also affect its interpretation. Lowe (1996) conducted a series of studies in which expert and novice meteorologists had to interpret weather maps. Although both groups were familiar with the graphical conventions, they differed in their interpretations. For example, the experts related features of the maps that were causally associated whereas the novice related features that were visually similar or close together. Moreover, the experts often made inferences including predictions of how the weather would change in the future, whereas the novices focused on describing the current weather represented by the map.

Finally, top-down effects of knowledge can interact with bottom-up effects of display design and these interactions can affect both where people look on the maps and how they interpret the maps. In experimental studies, college students were given the task of interpreting weather maps to predict wind direction in a region of the map, given information about pressure (Canham & Hegarty, 2010; Fabrikant, Rebich-Hespanha, & Hegarty, 2010; Hegarty, Canham, & Fabrikant, 2010). Bottom-up effects of display design were investigated by manipulating the number of displayed variables on the maps, or the visual salience of task-relevant vs. task-irrelevant information. Top-down effects of domain knowledge were investigated by examining performance and eye fixations before and after participants learned relevant meteorological principles. Map design and knowledge interacted such that salience had no effect on performance before participants learned the meteorological principles, but after learning, participants were more accurate if they viewed maps that made task-relevant information more visually salient.

Principles of Effective Visualization

Based on both information processing theories and empirical research on comprehension of visualizations, Hegarty (2011) summarized a set of principles for the design of effective visualizations. These are best considered as heuristics for the design of displays. Many of them have been documented by Kosslyn (1989, 1994, 2006) in a series of articles and books about graph design and by Gillan, Wickens, Hollands, and Carswell (1998). As noted earlier, a general meta-principle that there is no such thing as a "best" visualization, independent of the task to be carried out with this display and the user of the display. In addition, Hegarty summarized the following sets of principles:

Principles Related to the Expressiveness of Displays

One set of principles refers to how much information should be included in a visualization. A general principle, referred to as the *relevance* principle by Kosslyn (2006) states that visualizations should present no more or no less information than is needed by the user. Presenting all of the relevant information in the display relieves the user of the need to maintain a detailed representation of this information in working memory, whereas presenting too much information in the display leads to visual clutter or distraction by irrelevant information (Wickens & Carswell, 1995). A related principle, the *principle of capacity limitations* points out that graphics should be designed to take account of limitations in working memory and attention. The relevance principle is related to the idea of data-ink ratio, proposed by statistician and information design pioneer Edward Tufte (2001). Tufte advocated deleting all non-data ink and all redundant data ink

"within reason" (p. 96), including deleting background pictures that are often included in newspaper graphics, deleting the lines and tick marks on the axes of graphs and deleting the bars and filler patterns in bar graphs (which he referred to as redundant coding). Gillan and Richman (1994) provided empirical evidence that increasing the data ink ratio in graphs improved accuracy and decreased response time, but their research also indicated that Tufte's principle was too simplistic. For example, background pictures were generally disruptive, including the x and y axes was generally beneficial, and the effects of redundant coding in displaying the data (e.g. fill patterns for bars in bar graphs) were inconsistent and depended on the task and type of graph.

Principles Related to the Perception of Displays

To be effective, a visual display has to be accurately perceived. Tversky, Morrison, and Betrancourt (2002) refer to this as the apprehension principle of visual displays. For example, Kosslyn (2006) draws on basic research in psychophysics to make the point that visual forms indicating a difference between two variables need to differ by a large enough amount to be perceived as different. This is related to the general principle that one should use visual dimensions that are accurately perceived (Cleveland & McGill, 1984) and avoid using visual variables that lead to biased judgments (Wickens & Hollands, 2000). For example, comparing the position of two entities along a common scale is easier than comparing the position of entities on identical but non-aligned scales, and perception of length is more accurate than perception of area (Cleveland & McGill, 1984).

Finally another important principle is what Kosslyn (2006) refers to as the *principle of perceptual organization*, that people automatically group elements of displays into units. This principle is based on the Gestalt principles of perceptual organization, which determine which elements of displays are grouped. These groupings can be compatible or incompatible with the tasks to be carried out with a display. For example as discussed earlier, line graphs facilitate comparisons between the units plotted on the x axis (see Figure 2.3a) because the lines group data points as a function of this variable, reflecting the Gestalt principle of good continuation (Shah & Freedman, 2011).

Principles Related to the Semantics of Displays

A third set of principles refers to the semantics of visual displays. A visualization is easier to understand if its form is consistent with its meaning (the compatibility principle, Kosslyn, 2006). For example, the use of visual variables to convey meaning needs to be consistent with common spatial metaphors in our culture, such as up is good, down is bad and larger graphical elements represent more of something. Other common assumptions include that lines indicate connections, circles indicate cyclic processes, and the horizontal dimension is naturally mapped to time (Tversky, 2011).

An important principle emphasized by several theorists (Bertin, 1983; Zhang, 1996; Mackinlay, 1986) is matching the dimensions of the visual variables with the underlying variables that they represent in terms of scales of measurement. Both representing and represented dimensions can vary in scale from categorical to interval, ordinal, or ratio (Stevens, 1946). For example, shape is a categorical variable, shading is an ordinal dimension, orientation is an interval dimension and length is a ratio dimension. Zhang (1996) proposed that representations are most *accurate* and *efficient* when the scale of the representing variable corresponds to the scale of the represented variable. *Efficiency* refers to the fact that the relevant information can be perceived because it is represented in the external representation. For example, the use of error bars to display measurement error violates this principle because error bars are discrete (categorical)

representations of a continuous function (Cumming & Finch, 2005). The use of error bars therefore can drive the interpretation that values of a variable can only fall within the error bars and are equally likely anywhere within the error bars. The match, in terms of scales of measurement, between the representing and represented variables is a central principle coded by Mackinlay (1986) in a system that automated the design of relational graphics.

Principles Related to Pragmatics and Usability

A final set of principles relates to the pragmatics and usability of visualizations. Pragmatics refers to the broader context in which visual displays communicate and their rhetorical function. One general pragmatic principle, the principle of salience, states that displays should be designed to make the most important thematic information salient (Bertin, 1983; Dent, 1999; Kosslyn, 2006). A related principle, the principle of informative changes (Kosslyn, 2006) is that people expect changes across properties of a display to carry information. More broadly, the ways in which information is visually displayed can subtly communicate information. For example, people have more confidence in data that are presented in realistic displays, although this in not necessarily warranted (see Smallman & St. John, 2005; Wainer, Hambleton & Meara, 1999). Moreover, the proportion of space taken up by graphs in journal articles varies across the sciences, with more space devoted to graphs in disciplines rated as "hard sciences" (Cleveland, 1984; Smith, Best, Stubbs, Archibald, & Roberson-Nay, 2002). How data are displayed might therefore give an impression of their reliability or scientific nature.

Usability of a visualization refers to ensuring that the viewer has the necessary knowledge to extract and interpret the information in the display. Visual displays are based on graphic conventions and users need to know the conventions of a particular graphic form in order to comprehend it. Kosslyn (2006) refers to this as the *principle of appropriate knowledge*. This knowledge is often thought of as being part of the graph schema (Pinker, 1990; Ratwani & Trafton, 2008). The conventions of a visual display are often provided in a legend and in cartography a legend is considered to be an obligatory component of every map (Dent, 1999). While data graphs often include legends, for example stating which colors, shading etc. refer to which variables in a bar or pie chart, their comprehension often depends on more basic assumptions that the user is expected to have. For example, an understanding of measurement error is necessary to interpret error bars (Zwick, Zapata-Rivera, & Hegarty, 2014) and knowledge of meteorology is necessary to make predictions from weather maps (Lowe, 1996). Thus, providing a legend or graph schema is not sufficient for understanding a visual display.

Applications to the Design of Score Reports

Educational testing is becoming increasingly important, at least in the United States. Consequently, there is increasing need for the production and comprehension of score reports. Score reports come in different forms and are designed for different purposes (see also Tannenbaum, this volume). One important distinction is between reports of student performance at the aggregate level (e.g., comparing nations on international assessments such as PISA and TIMMS) and test reports for individual students; another important distinction is between reports of formative and summative assessments. With the advent of Cognitive Diagnostic Assessments, score reports might include information on specific skills that relate to a model of achievement (Roberts & Gierl, 2010). Test reports are also used by a variety of stakeholders including students, parents, teachers, administrators, policy makers, and researchers. Not surprisingly there are a number of existing papers that examine test reports with respect to principles of graph design (e.g., Allalouf, 2007; Goodman & Hambleton, 2004; Zenisky & Hambleton, 2012;

Wainer, Hambelton, & Meara, 1999). One general conclusion from this research is that there is not much standardization in test reports while another, echoing the theme of this review, is that what makes an effective score report depends on the type of data and the consumer of the report. In this final section of this paper, I raise issues regarding the design of score reports with respect to the types of graphical principles discussed earlier.

First, we can review score reports with respect to principles related to the *expressiveness* of displays, and specifically, the principle of appropriate knowledge. The design of score reports needs to be responsive to the needs of the user. This means that one might not necessarily design the same score report for teachers and parents, depending on what teachers and parents need to know. It is important to give various stakeholders the right amount of information. Giving them too much information, or extraneous information (such as superimposing a graph on a picture) might just create a cluttered display in which the most relevant information to their needs is not salient.

Turning to principles related to the *perception* of displays, it is important to ensure that text and graphic elements are large enough to be accurately perceived and that the graph uses visual elements that are accurately perceived. For example, scales showing different test scores should be aligned if it is important to compare a student's relative performance on different subtests (Cleveland & McGill, 1984). It is also important to consider the perceptual organization of the display and how different types of graphs more readily communicate different aspects of the data. For example, as noted by Figure 2.3, it is easier to see trends over time in line graphs, whereas bar graphs might be more appropriate if the most important information is about the comparison of different groups.

In terms of *semantic* principles, it is important that the visual variables used to display various quantities are consistent with natural semantic mappings, such as using higher values on a graph to display larger scores and using the horizontal dimension (x axis) to show trends over time. Finally, it is important to match the scale of measurement of a numerical variable to the visual variable used to depict that variable. For example, color might be a good choice of visual variable to depict average percent correct on a test for different schools (because school is a categorical variable in this instance), whereas height of a bar on a bar graph might be more appropriate for visualizing the average percent scores themselves, because both the scores and length of a bar are ratio variables.

Finally, score reports need to be designed to follow principles of *pragmatics* and *usability* of information displays. The principle of appropriate knowledge is critical here. People need to have the necessary knowledge to understand a graph. This can often be provided in a legend, or interpretive guide. However a legend or guide explaining graphic conventions might not be sufficient if the person interpreting the score report does not have basic mathematical or statistical knowledge necessary to understand what is depicted, or if the guide uses statistical jargon that the reader is not able to understand (Goodman & Hambleton, 2004). For example, a graph of scaled scores will not be meaningful to a parent who is not familiar with the scale used. The same point can be made about error bars showing the confidence interval for an observed score. For example Zwick et al. (2014) examined teachers' comprehension of different score reports that showed confidence intervals either as error bars or as more graded violin plots that showed variable width confidence bands (see Figure 2.7). When asked which displays they preferred and why, some teachers expressed misconceptions about the nature of confidence intervals. For example, one teacher expressed a preference for the fixed width confidence bands stating, "the bar of equal width gives you a better picture of another score being equally likely to occur in that range" (p. 135). In sum, a legend or guide to a score report can only do so much. While researchers have made some inroads to designing displays taking account of the knowledge of consumers (Zapata-Rivera & Katz, 2014), we need more basic research on what different stakeholders know about both measurement concepts and graphic conventions.

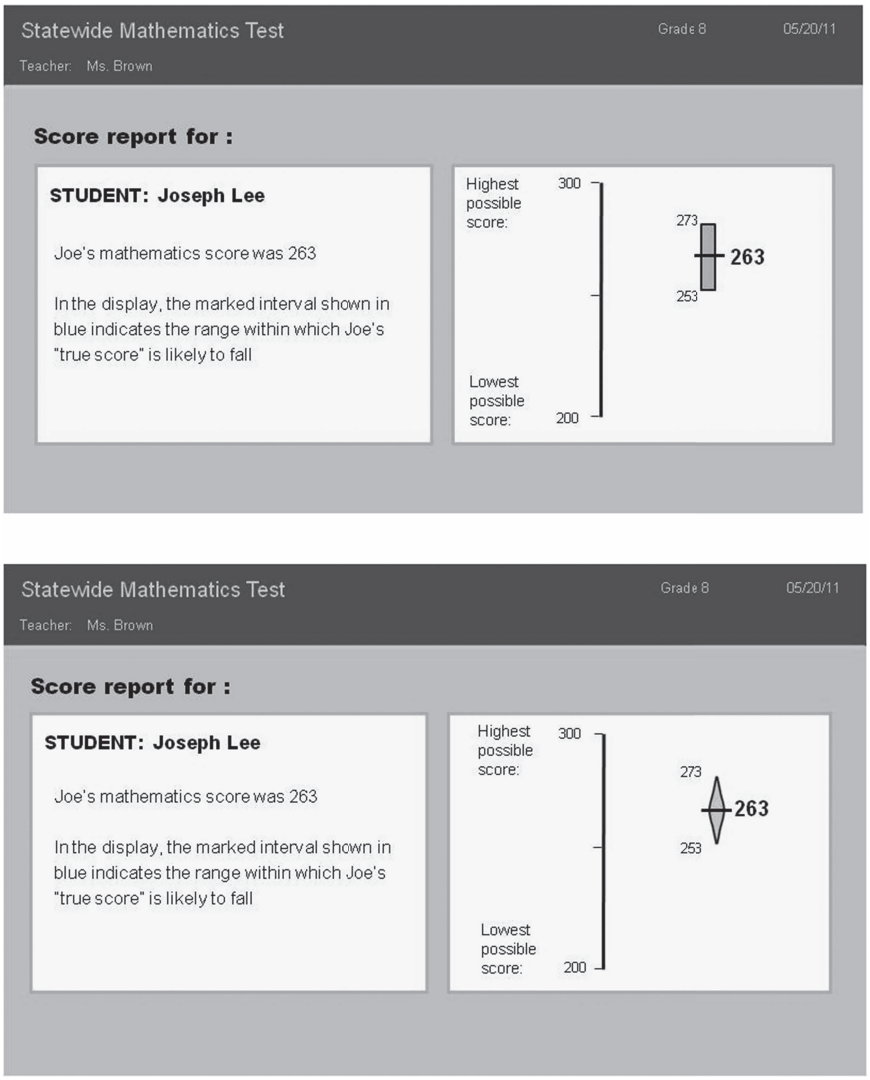


Figure 2.7 Examples of score reports compared by Zwick et al. (2014).

More generally, the development of effective score reports has to be a process of iterative design and evaluation and any score reports that are designed need to be evaluated with the actual stakeholders (teachers, parents, policy makers etc.) for which they are intended. A number of researchers (Zapata-Rivera, Vanwinkle, & Zwick, 2012; Zenisky & Hambleton, 2012) have proposed frameworks for designing and evaluating score reports that follow this general principle). This evaluation needs to go beyond considering what types of score reports people prefer, because preference for a display is often dissociated from ability to understand it (Smallman & St. John, 2005; Wainer et al., 1999). Another issue is that there seems to be very little standardization of similar types of test reports across different contexts, for example test reports in different states (Goodman & Hambleton, 2004). Development of standard displays and graphic conventions for similar types of score reports would ensure that consumers have to

master fewer basic formats for these reports and are able to transfer their understanding of these reports more easily across contexts.

Finally, given the increased prevalence of test scores and other information visualizations in our lives, it is worth considering whether we need better means of educating people about how interpret graphic displays and about the measurement concepts that are necessary to interpret these displays. Recent studies have has some success in educating both student teachers and working teachers about measurement concepts and score reports (Zapata, Zwick, & Vezzu, 2016; Zwick et al., 2008). Having students interpret their own test score reports might be a "teaching moment" for educating them about both graphical conventions and measurement concepts.

In conclusion, cognitive scientists have made significant progress in understanding how people understand information visualizations. Insights from cognitive science have suggested general principles that can be used to design more effective visualizations of score reports. At the same time, cognitive science studies suggest that not all current problems in interpretation of score reports can be solved by display design alone. These studies also suggest that we also need to educate stakeholders to be more knowledgeable about the nature of educational measurement and graphical conventions, so that they can be better consumers of score reports.

References

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, *26*(1), 36–46.

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association*, *13*, 608–618.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396.

Bertin, J. (1983). *Semiology of graphics: Diagrams networks maps* (W. Berg, Trans.). Madison: University of Wisconsin Press.

Boone, A. P., Gunalp, P., & Hegarty, M. (in press). Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. *Journal of Experimental Psychology: Applied.*

Broad, K., Leiserowitz, A., Weinkle, J., & Steketee, M. (2007). Misinterpretations of the 'cone of uncertainty' in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society*, *88*(5), 651–667.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, *20*(2), 155–166.

Card, S. K., Mackinlay, J. D., & Schneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco: Morgan Kaufmann Publishers.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*, 75–100.

Cleveland, W. S. (1984). Graphs in scientific publications. *American Statistician*, *38*, 261–269.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*, 531–554.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*, 25–47.

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142–2151.

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics*, *29*(3), 89–93.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170.

Dent, B. D. (1999). *Cartography: Thematic map design*. Boston, MA: McGraw-Hill.

Fabrikant, S. I., Rebich-Hespanha, S., & Hegarty, M. (2010). Cognitively inspired and perceptually salient graphic displays for efficient inference making. *Annals of the Association of American Geographers*, *100*, 13–29.

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical Decision Making*, *31*(3), 444–457.

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, *22*(5), 392–399.

Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *22*, 231–239.

Gillan, D. J., & Richman, E. H. (1994). Minimalism and the syntax of graphs. *Human Factors*, *36*, 619–644.

Gillan, D. J., Wickens, C. D., Hollands, J. G., & Carswell, C. M. (1998). Guidelines for presenting quantitative data in HFES publications. *Human Factors*, *40*, 28–41.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145–220.

Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in Cognitive Science*, *3*(3), 446–474.

Hegarty, M., Canham, M., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *36*, 37–53.

Kinkeldey, C., MacEachren, A. M., Riveiro, M., & Schiewe, J. (2017). Evaluating the effect of visually represented geodata uncertainty on decision-making: Systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science*, *44*, 1–21.

Kosslyn, S. M. (1989). Understanding charts & graphs. *Applied Cognitive Psychology*, *3*, 185–226.

Kosslyn, S. M. (1994). *Elements of graph design*. New York, NY: W. H. Freeman.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. New York, NY: Oxford University Press.

Kozhevnikov, M., Hegarty, M., & Mayer, R. E. (2002). Revising the visualizer-verbalizer dimension: Evidence for two types of visualizers. *Cognition and Instruction*, *20*(1), 47–77.

Kriz, S., & Hegarty, M. (2007). Top-down and bottom-up influences on learning from animations. *International Journal of Human-Computer Studies*, *65*, 911–930.

Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–99.

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, *27*(5), 696–713.

Lowe, R. K. (1996). Background knowledge and the construction of a situational representation from a diagram. *European Journal of Psychology of Education*, *11*, 377–397.

Mackinlay, J. D. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, *5*, 110–141.

McDermott, L. C., Rosenquist, M. L., & van Zee, E. H. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, *55*, 503–513.

Peebles, D., & Cheng, P. C-H. (2003). Modeling the effect of task and graphical representation on response latency in a graph-reading task. *Human Factors*, *45*, 28–46.

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31–35.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Erlbaum.

Ratwani, R. M., & Trafton, J. G. (2008). Shedding light on the graph schema: Perceptual features versus invariant structure. *Psychonomic Bulletin & Review*, *15*, 757–762.

Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, *29*, 25–38.

Ruginski, I. T., Boone, A. P., Padilla, L. M., Liu, L., Heydari, N., Kramer, H. S., . . . Creem-Regehr, S. H. (2016). Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, *16*(2), 154–172.

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, *45*, 115–143.

Shah, P., & Carpenter, P. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*, 337–370.

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, *3*(3), 560–578.

Shah, P., Freedman, E. G., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visual spatial thinking*. New York, NY: Cambridge University Press.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, *91*(4), 690.

Simkin, D. K., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, *82*, 454–465.

Smallman, H. S., & St. John, M. (2005). Naïve realism: Misplaced faith in realistic displays. *Ergonomics in Design*, *13*, 14–19.

Smith, L. D., Best, L. A., Stubbs, A. D., Archibald, A. B., & Roberson-Nay, R. (2002). Constructing knowledge: The role of graphs and tables in hard and soft psychology. *American Psychologist*, *57*, 749–761.

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, *5*, 61–77.

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393–1400.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Tufte, E. T. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, *3*(3), 499–535.

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, *57*, 247–262.

Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton NJ: Princeton University Press.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36*(4), 301–335.

Wickens, C. D., & Carswell, M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, *37*, 473–494.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance*. Upper Saddle River, NJ: Prentice Hall Inc.

Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL™ teachers* (Research Memorandum 12–20). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*(3), 215–229.

Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, *21*, 442–463.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26.

Zhang, J. (1996). A representational analysis of relational information displays. *International Journal of Human Computer Studies*, *45*, 59–74.

Zwick, R., Sklar, J. C., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice*, *27*(2), 14–27.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*(2), 116–138.

3

Subscores

When to Communicate Them, What Are Their Alternatives, and Some Recommendations

Sandip Sinharay, Gautam Puhan, Shelby J. Haberman, and Ronald K. Hambleton

Subscores are scores based on any meaningful cluster of items on a test. For example, scores on the algebra and geometry sections on a mathematics test or programming applications and program design and development sections on a computer science test are often referred to as subscores. Brennan (2012) stated that users of test scores often want (indeed demand) that subscores be reported, along with total test scores due to their potential diagnostic, remedial, and instructional benefits. According to the National Research Council report "Knowing What Students Know" (2001), the purpose of assessment is to provide particular information about an examinee's knowledge, skill, and abilities and subscores may have the potential to provide such information. Furthermore, the US Government's No Child Left Behind (NCLB) Act of 2001 and the Every Students Succeeds Act (ESSA; Every Student Succeeds Act, 2015–2016) requires that state assessments provide more detailed and formative information. In particular, it requires that state assessments "produce individual student interpretive, descriptive, and diagnostic reports" (p. 26); subscores might be used in such a diagnostic report. As is evident, there is substantial pressure on testing programs to report subscores, both at the individual examinee level and at aggregate levels such as at the level of institutions or states. It is therefore not surprising that subscores are reported by several large-scale testing programs, such as SAT®, ACT®, Praxis, and LSAT.

The next section provides a review of the literature on how subscores are reported and on recommendations regarding how they should be reported. The Quality of Subscores section includes a discussion of how existing subscores often do not have satisfactory psychometric properties. The Techniques section includes a discussion of several methods for evaluating the quality of subscores and a review of the existing analyses of subscores with respect to their quality. Several alternatives to subscores are discussed in the penultimate section. The final section includes several conclusions and recommendations.

Existing Findings and Recommendations on Communicating Subscores

Goodman and Hambleton (2004) provided a comprehensive review and critique of score reporting practices from large-scale assessments. Figures 17–21 of Goodman and Hambleton (2004)

include examples of several operational score reports that include subscores. The subscores in the score reports examined by them were reported in one of three forms: as number-correct or raw scores, as percent-correct scores, or percentile raw scores.

Figure 3.1 of this chapter shows an example of one section of a sample *Praxis*® score report (shown in www.ets.org/s/praxis/pdf/sample_score_report.pdf). A total of five subscores with maximum possible score ranging from 15 to 37 are reported for this test with title *Praxis Elementary Education: Curriculum, Instruction, and Assessment*. The average range of subscores earned by the middle 50%of test takers is reported and can be used to compare how well an examinee did versus the other test takers who took this test. Although such a report might be useful to assess examinees' specific strengths and weaknesses in these five sub-content areas, there is a word of caution that the sample score report provides, which is that these subscores are based on small numbers of questions and are less reliable than the official scaled scores. Therefore, they may not be used to inform any decisions affecting examinees without careful consideration of such inherent limited precision.

The problems that Goodman and Hambleton (2004) noticed in the score reports that they examined include (a) reports that assume a high level of statistical knowledge on the part of the users, (b) use of statistical jargon such as "statistical significance" or "standard error" that confused and often intimidated the users, (c) misunderstanding or ignoring of technical symbols, concepts, and footnotes by the users of score reports, (d) reports that provided too much information that made it difficult for users to extract what they needed most, (e) including excessively dense graphics and displays that was daunting for readers, and (f) lack of descriptive information such as definitions and examples to aid interpretation of the results. Rick and Park (2017)

Test / Test Category *	Your Raw Points Earned	Average Performance Range **
ELEMENTARY EDUCATION: CURRICULUM, INSTRUCTION, AND ASSESSMENT (5017)		
I. READING AND LANGUAGE ARTS	33 out of 37	23–29
II. MATHEMATICS	26 out of 31	19–25
III. SCIENCE	15 out of 20	11–15
IV. SOCIAL STUDIES	14 out of 17	9–13
V. ART, MUSIC, AND PHYSICAL EDUCATION	13 out of 15	6–12

* Category-level information indicates the number of test questions answered correctly for relatively small subsets of the questions. Because they are based on small numbers of questions, category scores are less reliable than the official scaled scores, which are based on the full sets of questions. Furthermore, the questions in a category may vary in difficulty from one test to another. Therefore, the category scores of individuals who have taken different forms of the test are not necessarily comparable. For these reasons, category scores should not be considered a precise reflection of a candidate's level of knowledge in that category, and ETS recommends that category information not be used to inform any decisions affecting candidates without careful consideration of such inherent lack of precision.

** The range of scores earned by the middle 50% of a group of test takers who took this form of the test at the most recent national administration or other comparable time period. N/C means that this range was not computed because fewer than 30 test takers took this form of the test or because there were fewer than eight questions in the category or, for a constructed-response module, fewer than eight points to be awarded by the raters. N/A indicates that this test section was not taken and, therefore, the information is not applicable.

Figure 3.1 Section of a Praxis® test score report that includes subscores.

sought to extend the findings of Goodman and Hambleton (2004) by conducting a similar investigation using score reports from 23 US states and one US territory. They used the findings of Goodman and Hambleton as a basis for comparison and identified score reporting practices that have remained mostly unchanged over the last decade or so. Examples of such practices are presenting overall results both graphically and numerically and having score reports that are still about two pages in length. Rick and Park also identified practices that changed more visibly such as providing non-numerical or descriptive performance feedback, use of more color in score reports and more details about the precision of overall scores. Finally, they also identified new practices that were not commonly observed before such as contextualizing results in terms of "college and career readiness." For more information on these and other related score reporting practices, see Hambleton and Zenisky (2013) and Zenisky and Hambleton (2012, 2016).

As interest in score reporting continues to increase, researchers have been trying to find better ways to communicate test results via good score reports. Zapata-Rivera, van Winkle, and Zwick (2012) suggested a framework for designing and evaluating score reports—the framework was based on methodologies used in the following areas: assessment design (e.g., Mislevy, Steinberg, & Almond, 2003), software engineering (e.g., Pressman, 2005), and human-computer interaction (e.g., Nielsen, 1994) and included the following steps: (a) gathering assessment information needs, (b) reconciling these needs with the available assessment information, (c) designing score report prototypes, and (d) evaluating these score report prototypes with internal and external experts. Tannenbaum (this volume) provided a seven-step model to develop score reports, which he claims are integral to assuring the validity and utility of test score interpretation. Similarly, Slater, Livingston, and Silver (this volume) described a step-by-step process to designing good score reports. This includes (a) gathering information about the score report needed, (b) creating a schedule for the score report design, (c) beginning graphic designs, (d) getting client's reactions to initial designs, (e) gathering feedback from intended users of the score report, (f) revising the design based on information from end-users, and (g) finalizing the design. These processes can not only be used to design score reports for the overall test scores but for subscores as well.

Roberts and Gierl (2010) noted that integration and application of interdisciplinary techniques from education, information design (e.g., Pettersson, 2002), and technology are required for effective score reporting. They also presented a structured approach for developing score reports for cognitive diagnostic assessments. They provided guidelines for reporting and presenting diagnostic scores based on a review of educational score reporting practices and literature from the area of information design and presented a sample diagnostic report to illustrate application of their approach. Because subscores constitute a type of diagnostic scores, several recommendations of Roberts and Gierl apply to subscore reporting as well. For example, their sample diagnostic report included three sections: (a) a top section contained an overview of the contents of the report, (b) the middle section contained diagnostic information along with item-level performance, and (c) the bottom section contained a narrative summary of the examinee's performance across all the subareas. A score report that is supposed to include information on subscores could consist of similar sections as in Roberts and Gierl.

On the Psychometric Quality of Subscores

Despite the demand for and apparent usefulness of subscores, they have to satisfy certain quality standards in order for them to be reported. According to Haberman (2008), a subscore may be reported if it has high reliability and it is distinct from the other subscores. Similarly, Tate (2004) has emphasized the importance of ensuring reasonable subscore performance in terms of high reliability and validity to minimize incorrect instructional and remediation decisions. These concerns are in agreement with Standards 1.14 and 2.3 of Standards for

Educational & Psychological Testing (2014), which require proof of adequate reliability, validity, and distinctness of subscores. Monaghan (2006) pointed out that "While they want to be responsive to the desires of the educational marketplace, testing organizations are also very concerned about the appropriate use and interpretation of subscores." Just as inaccurate information at the total test score level may lead to decisions with damaging consequences (e.g., wrongly certifying someone to be a teacher or medical practitioner), inaccurate information at the subscore level can also lead to incorrect remedial decisions resulting in large and needless expenses for examinees, states, or institutions (Sinharay, Puhan, & Haberman, 2011).

Very low reliabilities of subscores on educational tests are common because these subscores are often based on only a few items. For example, the state test score report shown in Figure 20 of Goodman and Hambleton (2004) includes seven subscores that are based on only five items each. Such subscores are most often outcomes of retrofitting, where reporting subscores was not a primary goal but were later provided to comply with clients' requests for more diagnostic information on examinees. Furthermore, as Sinharay and Haberman (2008a) pointed out, real data may not provide information as fine-grained as suggested or hoped for by the assessment specialist. A theory of response processes based on cognitive psychology may suggest several skills, but a test includes a limited number of items and the test may not have enough items to provide adequate information about all of these skills. For example, for the iSkills™ test (e.g., Katz, Attali, Rijmen, & Williamson, 2008), an expert committee identified seven performance areas that they thought comprised Information and Communications Technology literacy skill. However, a factor analysis of the data revealed only one factor and confirmatory factor models in which the factors corresponded to performance areas or a combination thereof did not fit the data (Katz et al., 2008). As a result, only an overall Information and Communications Technology literacy score used to be reported for the test until the test was discontinued in 2017. Clearly, an investigator attempting to report subscores has to make an informed judgment on how much evidence the data can reliably provide and report only as much information as is reliably supported.

To demonstrate how low reliability can lead to inaccurate diagnostic information, Sinharay, Puhan, and Haberman (2010) considered the Praxis Elementary Education: Content Knowledge test. The 120 multiple-choice questions focus on four major subject areas: language arts/reading, mathematics, social studies, and science. There are 30 questions per area and a subscore is reported for each of these areas—the subscore reliabilities are between 0.71 and 0.83. The authors ranked the questions on mathematics and science separately in the order of difficulty (proportion correct) and then created a Form A that consists of the questions ranked 1, 4, 7, . . ., 28 in mathematics and the questions ranked 1, 4, 7, . . ., 28 in science. Similarly, Form B was created with questions ranked 2, 5, 8, . . ., 29 in mathematics and in science, and a Form C with the remaining questions. Forms A, B, and C can be considered roughly parallel forms and, by construction, all of the several thousand examinees took all three of these forms (see Figure 3.2 below for illustration). The subscore reliabilities on these forms range between 0.46 and 0.60.

The authors considered all the 271 examinees who obtained a subscore of 7 on mathematics and 3 on science on Form A. Such examinees will most likely be thought to be strong on mathematics and weak on science and given additional science lessons. Is that justified? The authors examined the mathematics and science subscores of the same examinees on Forms B and C. They found that:

- The percent of the 271 examinees with a mathematics subscore of 5 or lower is 34 and 39 respectively for Forms B and C.
- The percent with a science subscore of 6 or higher is 39 and 32 respectively for Forms B and C.
- The percent of examinees whose mathematics score is higher than their science score is only 59 and 66 respectively for Forms B and C.

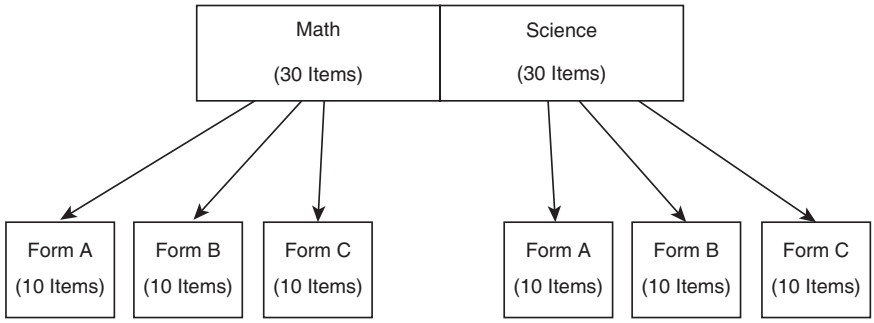


Figure 3.2 Graph showing the partition of the three sub forms from the original math and science forms.

This simple example demonstrates that remedial and instructional decisions based on short subtests will often be inaccurate. If subscores are used as a diagnostic tool, then the stakes involved in using subscores is less compared to scores that are used for high stakes use such as certification. Nevertheless, subscores should at least be moderately reliable (e.g., 0.80 or higher[1]) to be of any remedial use. Otherwise, users would mostly be chasing noise and making incorrect remedial decisions.

There are several statistical techniques starting from fairly basic statistical operations such as computing correlations and reliabilities to more sophisticated applications such as factor analysis and dimensionality analysis that have been used to assess whether the subscores are worth reporting. Some of these techniques are summarized in the next section.

Techniques to Evaluate When Subscores Are Worth Reporting

Use of Correlation or Reliability of Subscores

Researchers and practitioners often use simple rules to determine if subscores are worth reporting. Several researchers have used the correlations corrected for attenuation between the different subscores to decide whether it is reasonable to report subscores. If the disattenuated correlations among subscores are fairly high, then it essentially means that the subscores are not distinct from each other and therefore not worth reporting. For example, McPeek, Altman, Wallmark, and Wingersky (1976) and Haladyna and Kramer (2004) used the criterion that reporting of subscores is not warranted if the correlations corrected for attenuation among them are larger than 0.90. Similarly, researchers often judge the subscores to be useful if their reliabilities are sufficiently large. For example, Wainer et al. (2001, p. 363) commented that the subscores for three of the four sections of the item tryout administration of the North Carolina Test of Computer Skills are insufficiently reliable to allow individual student reporting.[2]

Application of Principal Component Analysis and Factor Analysis

A simple approach to evaluate whether the subscores are distinct enough would be to compute the eigenvalues from the correlation matrix of the subscores (or from the correlation matrix of the items) in a principal component analysis. If most of the eigenvalues computed from the correlation matrix of the subscores are smaller than 1 or if a scree plot of these eigenvalues shows that the eigenvalues abruptly levels out at some point, then the claim of several distinct subscores is probably not justified. On the contrary, the presence of multiple large eigenvalues

would support the reporting of subscores. For example, Sinharay, Haberman, and Puhan (2007) computed the eigenvalues from the 6 x 6 correlation matrix of 6 reported subscores from two forms of a test for paraprofessionals. They found that the largest eigenvalue was 4.3 for both forms while the remaining five eigenvalues were smaller than 0.5, suggesting that the test is essentially unidimensional and the claim of six distinct subscores is probably not justified. Similarly, Stone, Ye, Zhu, and Lane (2010) reported, using an exploratory factor analysis method on the inter-item correlation matrix, the presence of only one factor in the Eighth Grade Mathematics portion of the Spring 2006 assessment of the Delaware State Testing Program; hence, subscores were not worth reporting for the assessment.

Application of the Beta-Binomial Model

The method of fitting a mathematical model named the beta-binomial model (Lord, 1965) to the observed subscore distributions to determine if the subscores have added value over and beyond the total score has been suggested by Harris and Hanson (1991). Consider a test with two subscores. If the bivariate distribution of the two subscores computed under the assumption that the corresponding true subscores are functionally related provides an adequate fit to the observed bivariate distribution of subscores, the true subscores are functionally related and therefore do not provide any added value. Harris and Hanson used a chi-square-type statistic in their data example to determine the goodness of fit of the bivariate distribution of the two subscores under the assumption that the corresponding true subscores are functionally related to the observed bivariate distribution of subscores. However, as pointed out by Sinharay, Puhan, and Haberman (2010), the method of Harris and Hanson involves significance testing with a chi-square statistic whose null distribution is not well established.

Fitting of Multidimensional Item Response Theory Models

Another way to examine if subscores have added value is to fit a multidimensional item response theory (MIRT) model (e.g., Reckase, 1997; Ackerman, Gierl, & Walker, 2003) to the data. MIRT is a tool to model examinee responses to a test that measures more than one ability, for example, a test that measures both mathematical and verbal ability. In MIRT, the probability of a correct item response is a function of several abilities, rather than a single measure of ability. To examine if subscores have added value, one can perform a statistical test of whether a MIRT model provides a better fit to the data than a unidimensional IRT model. See von Davier (2008) for a demonstration of this sort of use of MIRT.

Using Dimensionality Assessment Softwares Such as DIMTEST and DETECT

The DETECT software (Zhang & Stout, 1999) uses an algorithm that searches through all of the possible partitions of items into clusters to find the one that maximizes the DETECT statistic. Based on results from a simulation study, Kim (1994) provided guidelines to interpret the DETECT statistic. According to these guidelines, if the DETECT statistic is less than 0.10, then the data can be considered as unidimensional. Values between 0.10 and 0.50 would indicate a weak amount of dimensionality, values between 0.51 and 1.00 would indicate a moderate amount of dimensionality, and values higher than 1.00 would indicate high level of multidimensionality. The DIMTEST software (Stout, 1987) implements a hypothesis testing procedure to evaluate the lack of unidimensionality in data from a test. It assesses the statistical significance of the possible dimensional distinctiveness between two specified subtests (the Assessment Subtest or AT and the Partitioning Subtest or PT). The test statistic T calculated by DIMTEST

represents the degree of dimensional distinctiveness between these two specified subsets. For example, if the testing practitioner wants to know if a test of general ability has distinct dimensions such as mathematics or reading that are dimensionally distinct from the rest of the items in the test, then the mathematics (or reading) items can form the assessment subtest and the remaining items can form the partitioning subtest; then, a significant value of the DIMTEST index *T* would indicate that the two subsets are dimensionally distinct.

Application of Haberman's Classical Test Theory Based Method

Haberman (2008) suggested a method based on classical test theory to determine if a subscore has added value. According to the method, a subscore has added value only if it can be predicted better by the corresponding subscore on a parallel form than by the total score on the parallel form (Sinharay, 2013). To apply the method, one examines whether the reliability of a subscore is larger than another reliability-like measure referred to as the proportional reduction in mean squared error (PRMSE) of the total score. Like the reliability coefficient, the PRMSE typically ranges from 0 to 1 with 0 and 1 indicating the lowest and highest degrees of trustworthiness, respectively. Using the Haberman method, a subscore is said have added value and may be considered worth reporting only when the subscore reliability is larger than the PRMSE of the total score, which happens if and only if the observed subscore predicts the true subscore more accurately than does the observed total score. Sinharay, Haberman, and Puhan (2007) discussed why the strategy suggested by Haberman is reasonable and how it ensures that a subscore satisfies professional standards. Conceptually, if the subscore is highly correlated with the total score (i.e., the subscore and the total score measure the same basic underlying skill), then the subscore does not provide any added value over what is already provided by the total score. A subscore is more likely to have added value if it has high reliability and if it is distinct from the other subscores. Applications of the Haberman method can be found in Lyren (2009), Puhan, Sinharay, Haberman, and Larkin (2010), Meijer, Boev, Tendeiro, Bosker, and Albers (2017), Sinharay (2010), Sinharay, Puhan, and Haberman (2010), and Sinharay, Puhan, and Haberman (2011). For example, Sinharay, Puhan and Haberman (2011) analyzed data from a teacher certification test with four subscores. They found the values of the reliabilities of the subscores to be 0.71, 0.85, 0.67, and 0.72, for the reading, math, social studies and science subscores, respectively. The corresponding values of the *PRMSE* for the total score are 0.72, 0.76, 0.74, and 0.79, respectively. Thus, only the mathematics subscore of the test had added value. Similar results were also reported by Puhan, Sinharay, Haberman, and Larkin (2010) for several other teacher certification tests.

How Often Do Subscores Have Adequate Psychometric Quality

In the previous section, we presented methods that can be used to examine the psychometric quality of subscores to determine if they are worth reporting. In this section, we will present research that shows how often subscores have adequate psychometric quality to be considered worth reporting. Both actual and simulated data were used in these studies.

Sinharay (2010) performed an extensive survey regarding whether subscores have added value over the total score using the Haberman (2008) method. He used data from 25 large-scale operational tests (e.g., P-ACT + English, P-ACT + Math, SAT Verbal, SAT Math, SAT, SweSAT, MFT Business, etc.), several of which reported subscores operationally. Of the 25 tests, 16 had no subscores with added value and among the remaining nine tests, all of which had at least an average of 24 items contributing to each subscore, only some of the subscores had added value. Sinharay also performed a detailed simulation study to find out when subscores can be expected

to have added value. The simulation study showed that in order to have added value, subscores have to be based on a sufficient number of (roughly 20) items, and be sufficiently distinct from one another—the disattenuated correlation between subscores has to be less than about 0.85.

Stone, Ye, Zhu, and Lane (2010) used an exploratory factor analysis method on the inter-item correlation matrix to report the presence of only one dominant factor for the 2006 eighth-grade Delaware Math Assessment even though the test blue print called for four content domains. Harris and Hanson (1991), using their method of fitting beta-binomial distributions to the observed subscore distributions, found subscores to have little added value for the English and mathematics tests from the P-ACT+ examination. Wainer et al. (2001) performed factor analysis and an examination of the reliability of the subscores on data from one administration of the Just-in-Time Examination conducted by the American Production and Inventory Control Society (APICS) certification. They concluded that the six subscales in the APICS examination did not appear to measure different dimensions of individual differences. However, they found a tryout form of the North Carolina Test of Computer Skills that has four subscales was not as unidimensional as the APICS examination, and an application of the method of Haberman (2008) to the data reveals that three of the four subscores have added value over the total score. Wainer, Sheehan, and Wang (2000) considered the problem of constructing skills-based subscores for the Education in the Elementary School Assessment that is designed for prospective teachers of children in primary grades (K-3) or upper-elementary/middle-school grades (4–8). They concluded, mostly from an analysis of reliability of the subscores, that the test's items were essentially unidimensional and therefore, skill-based subscores could not be supported. Ackerman and Shu (2009), using DIMTEST and DETECT, found subscores also not to be useful for a fifth-grade end-of-grade assessment.

Another study conducted by Sinharay and Haberman (2009) tried to answer the question "Is it possible to inform low-scoring test takers about the subareas in which those with similar total scores have been weak?" The assumption is that low-scoring examinees (i.e., repeaters) may perform differently on different subscores and therefore reporting subscores might be beneficial to them for planning remediation. The study tested this assumption using data from three forms of an elementary education test for prospective teachers. The authors divided the data into 20 groups ranked on the basis of their scaled scores and computed, for each examinee group, the average values of the four subscores on the test—a line joining the four average subscores was drawn for each group. Sinharay and Haberman did not observe any noticeable difference in the pattern of the average subscores in the different groups—-the line for each examinee group was roughly parallel to a horizontal line, which means that the test takers who obtain low scaled scores on the test perform almost equally poorly in all the subject areas on average. The authors concluded that it is not justified to share with low-scoring test takers the subareas where those with similar scores have been historically weak, simply because no such subareas exist. These studies collectively indicate that subscores on operational tests have more often been found not to be useful than to be useful and do not satisfy professional standards.

Alternatives to Simple Subscores

The studies in the previous section show that subscores on many high-stakes operational tests are probably not worth reporting because they are either unreliable or not distinct from the other subscores or a combination of both. One may wonder whether there exist ways to improve the psychometric properties of subscores so that they are worth reporting. Some psychometric techniques have been developed that can be used to enhance the reliability of subscores. But the usefulness of these techniques depends partially on the current state of the subscores. We group the current state of the subscores under the following three categories:

1. If an observed subscore is reliable and dimensionally distinct from the remaining subscores, then it may seem reasonable to simply report the observed subscores. A statistical technique to enhance the psychometric quality of these subscores may not be necessary.
2. If an observed subscore is unreliable and not dimensionally distinct from the remaining subscores, then it may not be justified to report the observed subscores or even a statistically enhanced subscore because a statistical technique cannot be expected to make up something that is simply not there. Examples of such tests are the battery of tests for measuring examinee and school progress that was considered in Sinharay and Haberman (2008b)—several correlations among the subscores were found to be larger than 1 after correction for attenuation.
3. Statistical techniques to improve the reliability of the subscores may have some utility if the observed subscore is moderately reliable and is moderately correlated with the remaining subscores. For example, for three teacher licensing tests considered in Puhan et al. (2010), no subscore was worth reporting according to the PRMSE criterion of Haberman (2008), but all augmented subscores (described shortly) had improved reliability and were worth reporting. In the next section we describe some techniques that have been suggested to improve the reliability of subscores.

Augmented Subscores and Weighted Averages

Wainer et al. (2000) suggested an approach to increase the precision of a subscore by borrowing information from other subscores—the approach leads to "augmented subscores" that are linear combinations of all subscores. Because subscores are almost always found to correlate at least moderately, it is reasonable to assume that, for example, the science subscore of a student has some information about the math subscore of the same student. In this approach, weights (or regression coefficients) are assigned to each of the subscores and an examinee's augmented score on a particular subscale (e.g., math) would be a function of that examinee's ability on math and that person's ability on the remaining subscales (e.g., science, reading, etc.). The subscales that have the strongest correlation with the math subscale have larger weights and thus provide more information on the "augmented" math subscore. Haberman (2008) suggested a weighted average that is a linear combination of a subscore and the total score where weights for the subscore and total score depend on the reliabilities and standard deviations of the subscore and the total score and the correlations between the subscores. Sinharay (2010) showed that the augmented subscores and weighted averages often are substantially more reliable than the subscores themselves. A possible limitation of augmented subscores or weighted averages is that it is hard to explain to users exactly what these scores mean (although some progress on this front has been made by Sinharay, 2018). For example, it may be difficult to explain to an examinee why her reported math score is based not only on her observed math score but also on her observed scores on other sub-sections such as reading and writing. Also, the test organizers and test-score users may not like the idea of borrowing from other scores. In addition, researchers such as Stone et al. (2010) raised the concern that augmented subscores and weighted averages may hide differences between the subscores by forcing the different augmented subscores of examinees to appear similar to each other. The similarity between the different augmented subscores of an examinee is a price to pay for their greater accuracy.

Objective Performance Index

The objective performance index or OPI (Yen, 1987) is another approach to enhancing a subscore by borrowing information from other parts of the test. This approach uses a combination of item response theory (IRT) and Bayesian methodology. The OPI is a weighted average of two

estimates of performance: (a) the observed subscore and (b) an estimate, obtained using a unidimensional IRT model, of the subscore based on the examinee's overall test performance. If the observed and estimated subscores differ significantly, then the OPI is defined as the observed subscore expressed as a percentage. One limitation of this approach is that because of the use of a unidimensional IRT model, it may not provide accurate results when the data are truly multidimensional, which is when subscores can be expected to have added value.

Estimated Skill Parameters From a Cognitive Diagnostic Model

It is possible to employ a psychometric model such as a cognitive diagnostic model (e.g., Fu & Li, 2007; Roberts & Gierl, 2010) or a diagnostic classification model (Rupp, Templin, & Henson, 2010) to report diagnostic scores instead of reporting subscores. These models assume that (a) solving each test item requires one or more skills, (b) each examinee has a discrete latent skill parameter corresponding to each of the skills, and (c) the probability that an examinee will answer an item correctly is a mathematical function of the skills the item requires and the latent skill parameters of the examinee. For example, a reading test may require the skills such as remembering details (skill1), knowing fact from opinion (skill 2), and speculating from contextual clues (skill 3) (McGlohen & Chang, 2008), and the probability that an examinee will answer a certain item on the reading test correctly, is determined based on the skills that item requires and if the examinee do have those skills. After a diagnostic classification model is fit to a data set, the estimated values of the skill parameters are the diagnostic scores that can be reported. Examples of such models are the rule space model (RSM; Tatsuoka, 1983), the attribute hierarchy method (AHM; Leighton, Gierl, & Hunka, 2004), the DINA and NIDA models (Junker & Sijtsma, 2001), the general diagnostic model (von Davier, 2008), and the reparametrized unified model (Roussos et al., 2007). The first two of these, the RSM and AHM, are slightly different from the other diagnostic classification models in nature because they do not estimate any skill parameters—they match the response pattern of each examinee to several ideal or expected response patterns to determine what skills the examinee possesses. While there has been substantial research on diagnostic classification models, as Rupp and Templin (2009) acknowledge, there has not been a very convincing case that unequivocally illustrates how the added parametric complexity of these models, compared to simpler measurement models, can be justified in practice. In addition, there have been few empirical illustrations that the diagnostic scores produced by these models are reliable and valid (see, e.g., Haberman & von Davier, 2007; Sinharay & Haberman, 2008a). Nonetheless, researchers have continued interest in CDMs and, as mentioned earlier, Roberts and Gierl (2010) provided guidelines for presenting and reporting diagnostic scores for assessments that employ CDMs, and several of those guidelines are applicable to reporting of subscores.

Estimates From a MIRT Model

Several researchers such as Luecht (2003), Yao and Boughton (2007), Yao (2010) and Haberman and Sinharay (2010) have examined the use of MIRT models (Reckase, 1997) to report subscores. Yao (2010) showed that MIRT models can be used to report a set of reliable overall as well as domain scores. Haberman and Sinharay (2010) also evaluated when subscores computed using a MIRT model have any added value over the total score or over subscores based on classical test theory and found that there is not much difference between MIRT-based subscores and augmented subscores (Wainer et al., 2001). Haberman and Sinharay also suggested reporting of estimated true subscores that are on the same scale as the number-correct subscores using MIRT models.

Scale Anchoring

Scale anchoring (e.g., Beaton & Allen, 1992; Hambleton, Sireci, & Huff, 2008) makes claims about what students at different score points know and can do and is an approach that can be used to report more information than the total score when subscores are not of adequate psychometric quality. Scale anchoring typically is carried out by (a) selecting a few dispersed points on the score scale (*anchor points*) that will be anchored, (b) finding examinees who score near each anchor point, (c) examining each item to see if it discriminates between successive anchor points, that is, if most of the students at the higher score levels can answer it correctly and most of the students at the lower level cannot, and (d) reviewing the items that discriminate between adjacent anchor points to find out if specific tasks or attributes that they include can be generalized to describe the level of proficiency at the anchor point (e.g., Phillips et al., 1993). The outcome from this review is a description of what students at various scale points know and can do (see, for example, Hambleton & Zenisky, 2018). Although scale anchoring seems promising, there can be confusion about the meaning of data related to score anchors, and care must be used in offering correct anchor score interpretations (Linn & Dunbar, 1992). Phillips et al. (1993) also described the danger of over-interpreting examinee performance at anchor points so that all examinees at a particular level are assumed to be proficient at all abilities measured at that level. Sinharay, Haberman, and Lee (2011) described statistical procedures that can be used to determine if scale anchoring is likely to be successful for a test. They used several data sets from a teacher certification program and concluded that scale anchoring is not expected to provide much useful information to the examinees for this series of examinations. Although the discouraging results for the data they considered do not necessarily imply that the same results will always be observed, they do indicate that success in scale anchoring is far from guaranteed.

Conclusions and Recommendations

While subscores are highly sought after by test users and are reported operationally for several large-scale assessments, not all subscores that are reported are of adequate psychometric quality. Based on the existing research, our recommendations on communicating subscores are provided in the following list. We do not focus (in the following list and elsewhere in this chapter) on how subscores should be communicated—such discussions can be found in several other chapters in this volume and in Goodman and Hambleton (2004) and Roberts and Gierl (2010). Instead, our recommendations mostly focus on when to communicate subscores.

1. Use of content blueprints as a basis for subscores does not necessarily guarantee that the different subscores will produce highly distinct subscores. For example, Sinharay, Haberman, and Puhan (2007) showed that subscores based on highly distinct content categories such as math, reading and writing still produced scores that were highly correlated to each other. Subscores based on psychometric approaches might be more useful. For example, Wainer, Sheehan, and Wang (2000) proposed a tree-based regression analysis where items are clustered in a way that minimizes within-cluster variation while simultaneously maximizing between-cluster variation. However, psychometric-based approaches may lead to subscores that would be difficult to interpret. They may also not be reproducible when applied to other data. Techniques such as evidence-centered design (e.g., Mislevy, Steinberg, & Almond, 2003) or assessment engineering practices for item and test design (e.g., Luecht, Gierl, Tan, & Huff, 2006) should be used to ensure that the subscores have satisfactory psychometric property.

2. Subscores that are reported should be of adequate psychometric quality. In other words, the reported subscores should provide evidence of adequate reliability, validity, and distinctiveness of the subscores. Any reported subscore, in order to be reliable, should be based on a sufficient number of carefully constructed items. Although there are no clear guidelines in the psychometric literature on how many items are needed to provide a reliable subscore, some research has shown that at least 20 multiple-choice items are typically needed in each subtest for the corresponding subscore to yield a moderately high reliability (e.g., Puhan, Sinharay, Haberman, & Larkin, 2010; Sinharay, 2010). Combining some subscores can increase subdomain test length and therefore might result in subscores that have higher reliability and hence added value. For example, subscores for "Physics: theory" and "Physics: applications" may be combined to yield one subscore for "Physics." Combining subscores to meet reliability considerations, however, can create interpretation problems. It is also important to ensure that the skills of interest are as distinct as possible from each other, though this is quite a difficult task to accomplish before seeing the data.
3. One can consider reporting of weighted averages or augmented subscores that often have added value (e.g., Sinharay, 2010) and often provide more accurate diagnostic information than the subscores do. Weighted averages may be difficult to explain to the general public, who may not like the idea that, for example, a reported reading subscore that is based not only on the observed reading subscore, but also on the observed writing subscore. However, several approaches to the issue of explaining such weighted averages can be considered. One is that the weighted average better estimates examinee proficiency in the content domain represented by the subscore than does the subscore itself. This result can be discussed in terms of prediction of performance on an alternative test. Sinharay, Haberman, and Wainer (2011) demonstrated using data from an operational test that the correlation between a subscore and the corresponding subscore on a parallel form is smaller than the correlation between the corresponding weighted average on the original form and the corresponding subscore on a parallel form; this finding is supported by theoretical results of Sinharay (2018). The issue can also be discussed in terms of common cases in which information is customarily combined. For example, premiums for automobile insurance reflect not just the driving experience of the policy holder but also related information (such as education and marital status) that predicts future driving performance. In most cases, this difficulty in explanation of the weighted averages is more than compensated for by the higher reliability of the weighted average.
4. Subscores should be reported on an established scale and also equated so that the definition of strong or weak performance in a subject area does not change across different administrations of a test. Note that if the subscores are based on only a few items, equating may not be of satisfactory psychometric quality for the subscores.[3] For example, if common items are used to equate the total test, only a few of the items will correspond to a particular subarea so that common-item equating (e.g., Kolen & Brennan, 2014) of the corresponding subscore is not feasible. Some research on equating of subscores and weighted averages has been done by Puhan and Liang (2011) and Sinharay and Haberman (2011).
5. Although we primarily discussed subscores for individual examinees, subscores can be reported at an aggregate level as well (e.g., institution or district levels). Longford (1990) and Haberman, Sinharay, and Puhan (2009) have suggested methods to examine whether aggregate-level subscores are of added value, and presented examples of situations when aggregate-level subscores do not have added value. So, examining the quality of the aggregate-level subscores before reporting them is important.
6. Finally, this chapter primarily focused on large-scale assessments. Although other types of assessments, such as formative assessments, are beyond the scope of this chapter, one

can apply some of the techniques and procedures discussed in this chapter to evaluate the utility of subscores in formative assessments. Just as subscore information in large-scale assessments can aid in training and remediation, subscores in formative assessments can help teachers plan instruction that takes into account the strengths and weaknesses of their students. For this purpose, they will need subscores that are reliable and distinct from the other subscores. Providing those subscores might be even more difficult in formative assessment, where it is impractical to administer long tests on an ongoing basis throughout the year.

Notes

1 The cutoff of 0.80 seems reasonable from a discussion in Nunnally (1978, p. 245).
2 Wainer et al. did not clearly mention what cutoff they used; however, given that the reliabilities of the four section scores are 0.52, 0.60, 0.77, and 0.85, it seems that they most likely used the criterion of 0.80 that is often attributed to Nunnally (1978).
3 That is another reason of not reporting subscores based on a few items.

References

Ackerman, T. A., Gierl, M. J., Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37–51.

Ackerman, T., & Shu, Z. (2009, April). *Using confirmatory MIRT modeling to provide diagnostic information in large scale assessment*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191–204.

Brennan, R. L. (2012). *Utility indexes for decisions about subscores* (CASMA Research Report No. 33). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Every Student Succeeds Act, Pub. L. No. 114–95 § 114 Stat. 1177 (2015–2016).

Fu, J., & Li, Y. (2007, April). *Cognitively diagnostic psychometric models: An integrative review*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*, 145–220.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based attribute diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1031–1038). Amsterdam: Elsevier.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions*, *24*, 349–368.

Hambleton, R. K., Sireci, S., & Huff, K. (2008). *Development and validation of enhanced SAT score scales using item mapping and performance category descriptions* (Final report). Amherst: University of Massachusetts, Center for Educational Assessment.

Hambleton, R. K., & Zenisky, A. (2018). Score reporting and interpretations. In W. van der Linden (Ed.), *Handbook of modern item response theory*, Volume 3 (pp. 127–142). London: Chapman and Hall.

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: Some new findings, research methods, and guidelines for score report design. In K. F. Geisinger (Ed.), *Handbook of testing and assessment in psychology* (pp. 479–494). Washington: American Psychological Association.

Harris, D. J., & Hanson, B. A. (1991, March). *Methods of examining the usefulness of subscores*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Katz, I. R., Attali, Y., Rijmen, F., & Williamson, D. M. (2008, April). *ETS's iSkills™ assessment: Measurement of information and communication technology literacy*. Paper presented at the conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York, NY: Springer.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.

Linn, R. L., & Dunbar, S. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*, 177–194.

Longford, N. T. (1990). Multivariate variance component analysis: An application in test development. *Journal of Educational Statistics*, *15*, 91–112.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*, 239–270.

Luecht, R. M. (2003, April). *Applications of multidimensional diagnostic scoring for certification and licensure tests*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Lyren, P. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research, and Evaluation*, *14*, 1–10.

McGlohen, M., & Chang, H-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*, 808–821.

McPeek, M., Altman, R., Wallmark, M., & Wingersky, B. C. (1976). *An investigation of the feasibility of obtaining additional subscores on the GRE advanced psychology test* (GRE Board Professional Report No. 74–4P). Princeton, NJ: Educational Testing Service. (ERIC Document No. ED163090).

Meijer, R. R., Boev, A. J., Tendeiro, J. N., Bosker, R. J., & Albers, C. J. (2017). The use of subscores in higher education: When is this useful? *Frontiers in Psychology*, *8*. doi:10.3389/fpsyg.2017.00305

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

Monaghan, W. (2006). *The facts about subscores* (ETS RDC-04). Princeton, NJ: Educational Testing Service.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

Nielsen, J. (1994). *Usability engineering*. San Francisco, CA: Morgan Kaufmann.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Pettersson, R. (2002). *Information design: An introduction*. Philadelphia, PA: John Benjamins Publishing.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales* (NCES 93421). Washington, DC: National Center for Education Statistics, US Department of Education.

Pressman, S. (2005). *Software engineering: A practitioner's approach*. New York, NY: McGraw-Hill Education.

Puhan, G., & Liang, L. (2011). Equating subscores under the non equivalent anchor test (NEAT) design. *Educational Measurement: Issues and Practice*, *30*(1), 23–35.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). Comparison of subscores based on classical test theory. *Applied Measurement in Education*, *23*, 1–20.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Rick, F., & Park, Y. (2017, April). How far have we come? A review and evaluation of changes in score reporting. In M. R. Roberts (Chair), *Looking back and moving forward on score reporting research and practice*. Symposium held at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, *29*(3), 25–38.

Roussos, L. A., DiBello, L. V., Stout, W. F., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnostic system. In J. Leighton & M. Gierl (Ed.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). New York, NY: Cambridge University Press.

Rupp, A. A., & Templin, J. L. (2009). The (un)usual suspects? A measurement community in search of its identity. *Measurement*, *7*(2), 115–121.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford Press.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174.

Sinharay, S. (2013). A note on added value of subscores. *Educational Measurement: Issues and Practice*, *32*(4), 38–42.
Sinharay, S. (2018). A new interpretation of augmented subscores and their added value in terms of parallel forms. *Journal of Educational Measurement*, *55*, 177–193.
Sinharay, S., & Haberman, S. J. (2008a). How much can we reliably know about what examinees know? *Measurement: Interdisciplinary Research and Perspectives*, *6*, 46–49.
Sinharay, S., & Haberman, S. J. (2008b). *Reporting subscores: A survey* (Research Memorandum RM-08–18). Princeton, NJ: Educational Testing Service.
Sinharay, S., & Haberman, S. J. (2009, September). *The value of subscores and the low-scoring test-takers in Praxis* (Inter-Office memorandum). Princeton, NJ: Educational Testing Service.
Sinharay, S., & Haberman, S. J. (2011). Equating of augmented subscores. *Journal of Educational Measurement*, *48*, 122–145.
Sinharay, S., Haberman, S. J., & Lee, Y. (2011). When does scale anchoring work? A case study. *Journal of Educational Measurement*, *48*(1), 61–80.
Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28.
Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Do adjusted subscores lack validity? Don't blame the messenger. *Educational and Psychological Measurement*, *71*, 789–797.
Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic subscores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*, 553–573.
Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40.
Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*, 266–285.
Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.
Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, *17*, 89–112.
Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.
Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140.
Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., et al. (2001). Augmented scores—'borrowing strength' to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Mahwah, NJ: Erlbaum Associates.
Yao, L. (2010). Reporting valid and reliability overall score and domain scores. *Journal of Educational Measurement*, *47*, 339–360.
Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*, 83–105.
Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the meeting of the Psychometric Society, Montreal, Canada.
Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL™ teachers* (Research Memorandum 12–20). Princeton, NJ: Educational Testing Service.
Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26.
Zenisky, A. L., & Hambleton, R. K. (2016). Test score reporting: Best practices and issues. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (pp. 585–602). New York, NY: Routeledge.
Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

4

Reporting Student Growth

Challenges and Opportunities

April L. Zenisky, Lisa A. Keller, and Yooyoung Park

The aim of score reporting in the context of K-12 educational testing is to provide stakeholders ranging from students and families to teachers, schools, districts, and states, as well as the general public, with information about student performance. The individual student score report documents that have served as the primary communication vehicle for such results follow something of a typical script across states, testing companies, and reporting vendors in terms of both content and format: High prominence is usually afforded to reporting of a scale score (often represented numerically and graphically) and a proficiency level classification, followed by or integrated with some normative results comparing a student's score to that of defined and relevant reference groups (e.g., school, district, state), and typically, there is also some kind of a subscale reporting, often using a number-correct metric. Individual state reports for individual students, of course, do vary from one another with the addition of elements such as results for released test items, narratives of performance strengths and weaknesses, and individualized guidance for next steps, as well as differences in structure and layout.

One element that is increasingly common on student test score reports is displays of student growth. Such displays, when included on score reports, draw on the test score history for each individual student and (sometimes) a statistical cohort group to compute and contextualize patterns of scores over time and often, to project future performance. There are numerous strategies and methods that have been developed and deployed operationally to compute indices of growth at both the individual and group (classroom, school, district, etc.) level. O'Malley, Murphy, McClarty, Murphy, and McBride (2011), in their overview of student growth models, characterized approaches to calculating student growth as three general types:

- *Growth to Proficiency models*: Student performance is compared to a yearly growth target, in order to reach a defined "proficiency" within a set number of years.
- *Value/Transition Tables*: Growth is considered relative to change in performance category assignment over years (e.g., movement from "Needs Improvement" to "Proficient").
- *Projection models*: Student performance is predicted using past and current student performance and the performance of prior cohorts in the target grades.

Each one of these approaches comes part and parcel with key statistical assumptions that guide appropriate interpretation and use, and these assumptions take on critical importance when these results are included on reports distributed to families, alongside other indicators of student performance.

The present chapter begins with an acknowledgement that student growth results are at present being included on reports for a wide range of stakeholders and are being used for purposes ranging from informational to high-stakes. At the outset is provided a very brief review of student growth approaches and models drawn from the psychometric literature, to set the stage for the practice of reporting growth operationally. While our purpose is not to litigate methodology or the underlying statistical assumptions of growth models, their use in various educational policy decisions has at times been challenged because of conceptual and computational complexity. This necessarily connects to questions pertaining to reporting, including the extent to which such results may be understood and ultimately used correctly by the various intended audience(s). From there, the chapter focuses on current practices in reporting growth, discussing strategies used and the implications for interpretation and use associated with different display approaches. The next portion of the chapter will report the results of a small-scale survey study carried out to investigate the extent to which various displays of student growth results can be understood and interpreted correctly. The last section of this chapter aims to apply the broader body of research on score reporting to the specific topic of growth reporting, synthesizing best practices in reporting and the results of the study presented here to evaluate the communication of growth results to consumers of educational test data, and ultimately to lay out a research agenda in this area.

A Brief Overview of Growth

The term "growth" has come to enter the current educational vernacular over the course of the past 10 to 15 years, albeit without much in the way of a standard or formalized meaning. At present, a variety of growth models have been developed to aid in the interpretation of scores from high-stakes assessments beyond simple status measures, and all of these models define growth in slightly different ways. Often, the various models do not make explicit what type of "growth" is being measured in the model; further, it is not uncommon that the statistically implemented definitions of growth do not conform to the common conceptualizations of growth, and using the term may at times be confusing or misleading to lay audiences (Keller, Colvin, & Garcia, 2016). Growth models are used for a variety of purposes and at many levels (for example, at the level of the individual student as well as aggregated to the school, district, or state).

The use of growth measurements in large-scale testing stemmed from a concern that status measures were insufficient to capture the work that schools were doing to advance students. With No Child Left Behind, status measures were primarily used to monitor school progress, through measures such as "percent proficient." Schools were expected to increase the percent of students proficient on statewide tests at various grade levels in various subjects from year to year. Such changes (and really, any observed improvements) in the number of students deemed proficient were indicators of progress of schools. In using such a system, however, there was no credit given to students who made progress but did not change from not proficient to proficient. This issue was seen as especially problematic for lower-performing schools who would have difficulty increasing the percent of students in the proficient category, but that were, nonetheless, improving the achievement of students. As a result, the family of approaches now known as growth measures was designed to give credit to schools that were making progress, and also to provide students and their families with more information about their progress.

While the idea of measuring academic growth as change is quite appealing, the reality of measuring academic growth is much more complex. Measuring physical gains, like height, for

example, is relatively easy, whereas measuring changes in academic performance is less so, for several reasons. First, the definition of the construct being measured is more complex. The construct of height is well-understood, clearly defined, and easy to compute. Changes in the construct are also easy to determine given that a measurement at time 1 can be easily compared to a measurement at time 2. For example, a child who is 36 inches at time 1 and 42 inches at time 2 grew 6 inches, and an inch is a readily-relatable unit for measuring human height. It might not be clear whether or not that is a lot for a child to grow in that time span, and other information would be required to provide context for that change in height. Some contextual factors might be the age of the child, the sex of the child, and the ethnicity of the child.

Turning to a different kind of construct to be measured, such as math achievement: In this context, the knowledge and skill of interest cannot typically be directly assessed (so a test instrument provides a proxy for that measurement), and thus it may be less clear what it is we are exactly measuring (although it does not seem controversial that our statewide tests are measuring academic knowledge). However, how to conceptualize that change is not as simple. The most intuitive way to conceptualize growth is to do as we do with the height example: Give a test at the beginning of the year and at the end of the year, and see how many more questions the student got right. Now, suppose the student got 10 items right (out of 50) in the fall and 40 items right in the spring. The difference in performance is 30 items, where the student was able to answer 30 more items correct in the spring data collection than in fall. This change could then be contextualized as is done with height to determine if that is a large change or not. However, a student cannot be administered the exact same test form in the fall and spring, as they might remember the questions, and further, the amount of testing time might be extraordinary to do this across subjects. (As an aside, there are well-known underlying statistical issues to using these simple gain scores (e.g. Cohen, Cohen, West, & Aiken, 2003)). Therefore, there is a desire to use existing tests, designed to measure status, to also measure the academic growth of students. Some ways growth could be, and currently are, conceptualized included:

- A higher score on the test compared to last time you took the test
- A change in proficiency category
- A greater mastery of the content
- Doing better than the other students in the class.

There are many resources available to read about the variety of growth models that are available for use, and that are in use (e.g. Castellano & Ho, 2013), and so they are not reiterated here.

One important defining characteristic of growth measures is the notion of whether the measure is a norm-referenced or criterion-referenced measure. Criterion-referenced measures would look at gains relative to the curriculum, such as how much more of the curriculum has the student mastered at "time two" as compared to what was mastered at "time one." These are the kinds of growth measures we might intuitively think about when we think about improvement. Norm-referenced measures, on the other hand, consider questions like "How did my performance, or change in performance, compare to my peers?" The definition of "peers" might vary across measures but can be either all students in the same grade, or all students with a similar score history, or all students with similar demographics.

Issues in Reporting Growth

To place growth reporting in the larger context of results reporting, it is important to note that while individual score reports have been furnished to students and families for as long as student testing has been taking place, there has been considerable evolution in the appearance,

contents, format, and distribution of score reports, and these changes have been particularly pronounced in the past 20 years or so. One especially notable publication that was evidence of such change and consequently further raised awareness of the importance of thoughtful and deliberate score reporting is the 1998 publication of National Educational Goals Panel (NEGP), *Talking About State Tests: An Idea Book for State Leaders*. While framed by educational goals that were set for the year 2000, this document is remarkable in establishing several clear principles for reporting that remain highly relevant to reporting practices today, divided among *strategic* and *content* recommendations. The NEGP suggested that states answer the following four questions through the reports distributed to families (NEGP, 1998, p. xi):

1. How did my child do?
2. What types of skills or knowledge does my child's performance reflect?
3. How did my child perform in comparison to other students in the school, district, state, and—if comparable data are available—the nation?
4. What can I do to help one of these children improve?

While these would seem to be obvious today, the NEGP's *Idea Book* helped to formalize these questions as guiding principles for report development, and ultimately helped launch a sea change in report development practices. With the 2002 passage and implementation of the No Child Left Behind Act and its subsequent reauthorizations, student testing was elevated to a place of significant public prominence in the US educational landscape, and consequently research attention on communicating test scores has increased dramatically. Rather than an uninteresting duty for testing agencies that is left to the end of test development, reporting today is increasingly viewed as a critical element of communication about what students know and can do, and agencies have increasingly devoted resources to advancing good test score reporting to facilitate action on the basis of the information presented. This has led to the development of several models for reporting (Zenisky & Hambleton, 2012; Zenisky & Hambleton, 2016, etc.) as well as a number of empirical studies about specific reporting elements including work by Zwick, Zapata-Rivera, and Hegarty (2014) and Zapata-Rivera, Kannan, and Zwick (this volume) on error, and consideration of the interaction of report contents and specific stakeholder groups (e.g., Goodman & Hambleton, 2004; Rick et al., 2016).

As noted at the outset of this chapter, typical test score reports for students contain a number of basic or common pieces of information. In recent years, the choice has been made by many state educational agencies in the United States to incorporate displays of results pertaining to growth on those reporting documents, in addition to the more familiar reporting elements that are generally present. This shift to include growth results on reports has been driven in part by policy decisions at the level of federal and state agencies and the availability of these results, but unlike other indicators of academic performance, growth results seem not yet to have found their way into many studies that evaluate the use and understanding of specific score report displays. In some typical approaches, growth is presented on individual score reports in the form of line graphs (e.g., Colorado, www.cde.state.co.us/accountability/understanding-growth-reports), but other displays in use include bar charts (Georgia, www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/GSGM/GSGM_EOG_SampleReport_16.pdf) and tables/text (http://understandthescore.org/score-report-guide/).

As with other elements of student score reports, report developers are faced with many content and design choices when formatting growth display results for inclusion on student score reports. In terms of communicating results for measures of student growth to educational stakeholders, regardless of the model or strategy used, there are several key considerations that should impact how these elements of reports should be implemented. Chief among

these considerations is *model choice*, because the nature of the statistical information to be communicated is a driving factor in what can and should be shared on a score report for families, educators, and/or educational administrators. Related to this is an accounting of the statistical assumptions of each model, which likewise define what is and is not possible from a reporting standpoint, in particular relative to interpretation and use. Next, reporting for growth model results will vary considerably based on the *appropriateness/relevance for individual reports and group reports*, and the landscape of approaches (text, graphical displays, and/or tables) encompasses a range of strategies. The final consideration in reporting such results is the issue of *score error for the growth model calculations*. The standard error or measurement for scale scores sporadically appear on student score reports at present, and this pattern seems to be continuing in the context of reporting growth results. The standard error of measurement on reports is already problematic for users (in general, users do not like the inclusion of the SEM, and often it is not understood), and growth scores are even less reliable than test scores and so the need for error bands is even greater than with single scores.

Part of the challenge of reporting growth is consistent with findings of other studies of reporting elements, in that some report elements pose particular difficulties in terms of data interpretation and use. Research has shown that misinterpretations are common when users are asked direct knowledge questions about various score report elements. To our knowledge there have been no published studies of reporting that have focused explicitly on growth reporting displays, though considerable efforts have been made by state education agencies to develop interpretive guides (across text, presentations, and video formats) to explain reports. Little is known about best practices for reporting growth, nor what elements of growth displays lend themselves to correct interpretations or misinterpretations by intended users. It should be noted that growth displays are being included in individual score reports sent to families as well as made available in dashboard-type interactive formats for teachers, school administrators, and district and state personnel, and the training and expectations for use in all of those scenarios are quite different.

To return to the guiding principles espoused by the questions posed by the NEGP in the 1998 publication, the inclusion of growth reports as a display element in individual score reports or students' needs to be evaluated for purpose. What do displays of growth mean, how do intended users understand them, and how does inclusion of those results on reports help intended users to move forward on the basis of that information? Ultimately, as with all aspects of student reports, when included growth results should accomplish a specific informational purpose, and evidence needs to be gathered to show that such results are understood and used as intended. As noted previously, there is not a lot of research surrounding best practices for reporting growth measures. Given the complexity of some of the growth measures in use, the means for communicating that information is crucial for proper understanding and use of the information. Research in this area will help practitioners learn of best ways to communicate this complex information in a way that is appropriate. In the next section, a small-scale study to begin to understand reporting practices in this area is described.

A Small-scale Study

To obtain information about how interpretable score reports are a small study was conducted online using Amazon Mechanical Turk (MTurk). Six score report displays that reported student growth were selected from publicly available score reports, and for each score report, items to evaluate the interpretability of these score reports were created and a random sampling of 2 displays and their associated statements (for agreement/disagreement) were presented to research participants.

Respondents were required to be at least 18 years old and live in the United States. 220 adult respondents living in the United States participated in the study. The median age grouping of the participants was between 25 and 34, which comprises about 45% of all participants in the study. Most of the respondents reported at least some college education, if not a degree, as 44% of the respondents indicated that they held a Bachelor's degree, about 26% of the respondents indicated they had education at the college level without degree, and 13% of the respondents held an associate's degree from a two-year college. A majority of the respondents (>75%) indicated having some course in statistics, and in terms of the highest level of statistics courses taken, about half (49.5%) of the participants indicated their undergraduate statistics courses were their highest level of statistics courses, followed by 24% of the participants with no statistics course experience, and followed next by almost 20% with statistics courses in high school. With regards to sex, 50% of the total respondents were male and 50% were female. By ethnicity, just over 81% of the respondents were white, nearly 12% were native Hawaiian or Pacific Islander, about 8% were Asian and 1% were African American. Finally, 5.9% of the respondents also reported a Spanish, Hispanic, or Latino background.

Six different displays of score reports that were publicly available were chosen for the study. All the displays used growth percentiles as their growth measure, making it possible to focus on evaluating the differences in the interpretability of displays, rather than metric differences. Student Growth Percentiles (SGPs: Betebenner, 2009) are one of the more popular measures of growth, as SGPs can be used for any test with multiple administrations, regardless of the type of scale that it is used. Statistically it is a complex model that utilizes quantile regression as the foundation, resulting in the potential for users to misinterpret the meaning of the growth percentile unless guidance and information about how to interpret the value is provided. The statistical details of the model can be found in Betebenner (2009). As noted by Goldschmidt et al. (2012) the SGP does not measure absolute growth in performance but provides a normative context to compare student test scores of students with similar score histories. Percentile ranks are assigned to students. For example, a student with an SGP of 70 performed better than 70% of his/her peers that had similar score histories. As such, many students can get an SGP of 70, but it does not mean that they have shown the same changes in performance.

Since the goal of the study was not to provide a critique of specific state's displays, but instead to evaluate broad interpretability of common growth reporting approaches, the images of the actual displays used are not provided. However, a mock-up of the various types of displays is provided in Figure 4.1, to provide context in understanding the results of the research. A description of the specific displays is also provided following Figure 4.1, and will reference this figure to provide context for the various types of displays.

Each participant was presented with two of the score displays, randomly selected from among the six displays prepared. Six statements were presented for each display, and survey participants were instructed to click on the one or more statements that they decided were true, given the information presented in the display. [At least one statement presented for each display was true.]

These display-specific statements were developed to be as parallel as possible across the displays, rather than have unique statements for each display. However, since some displays provided information that others did not, it was not possible to have completely parallel statements across displays. To this end, a framework was developed and used for constructing the specific statements across displays. The five information categories described below in Table 4.1 informed the formation of the six statements for each display, along with sample statements for each information category.

Although performance indicator results were not essential to this study, statements about the performance of the students were included to determine if the respondents could broadly

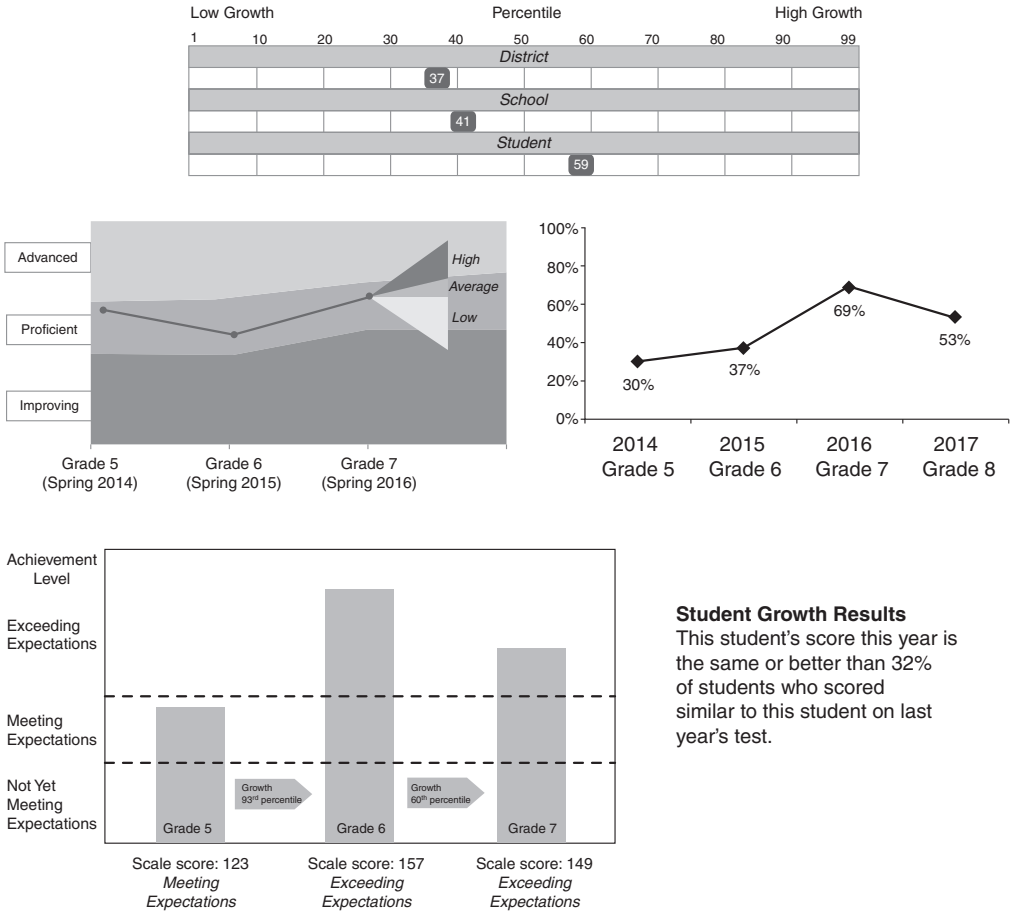


Figure 4.1 Mock-Ups of displays used (Clockwise from top: Table, Line, Text, Bar, Projection).

Table 4.1 Information categories of statements with examples

Information Category	Example Statement
Interpretation of performance	The student's performance was in the "Met Expectations" category.
Comparison of performance levels/scores	The student's performance level improved from Grade 4 to Grade 5.
Identifying the growth measure	The student's growth percentile in Grade 6 was 26.
Interpreting the growth measure	In 2017, the student's growth was the same or better than 52% of other grade 8 students.
Comparison of growth measures	The student's growth score increased from 2015 to 2016.

interpret the display as a starting point, as performance indicators are generally more familiar/straightforward test results for many prospective audiences. These statements will not be summarized in our results section but are provided here for context. In part these were included so that both statements were included on more and less familiar reporting metrics, lest the participants become frustrated in responding to the growth statements (which are generally less familiar displays) continuously.

The first display was a table-type display (Figure 4.1) that provided only information about SGPs and not on the student's performance. This display presented the growth of a student as well as the growth of the school and the district that the student belonged to, lending itself to the comparison of the individual student's growth to the average growth of his or her school and district, formatted as a table.

The second display was a line graph type display (Figure 4.1) that presented a student's performance history over the past three years, using performance level categories and growth percentile scores for each year. The growth percentiles for each of the three years were presented in a line graph, and the interpretation of the current year's growth percentile was provided in text next to the graph, with the SGP in bold print.

The third display included a bar graph display of performance, but the only growth-related information was included as a text-based display of growth information at the bottom of the score report, in the form of a single sentence (Figure 4.1). In general, this display focused primarily on the performance level of the student, both individually and within the context of others at the school, district, state, and cross-state levels, with growth information appearing relatively minimized.

The fourth display was a bar graph type display that presented a student's trajectory of achievement from the past two years to the current year in terms of scale scores and performance levels (Figure 4.1). The performance level of the student for each of the three years was presented both numerically and in a bar graph. Between the current year and the one year previous, the SGP was provided, along with the categorization of that growth, as either High, Low, or Typical.

The fifth display was a projection type display that focused on the changes of a student's achievement levels as a line graph, simultaneously using scale scores, and growth percentiles. The values for the scale scores and the SGPs were presented in a tabular form below the graph, and there were no values provided on the graph itself. The category of growth (High, Low, Typical) was provided in the table and in the graph through the use of color. Moreover, this display provided the projection of the student's future performance by indicting the levels of growth relative to the performance levels that might be obtained in the next year.

The sixth display was a projection type display, and was very similar to the fifth display, presenting the trajectory of the performance of an individual student as well as the projection of his or her future performance. However, in this presentation, the Y-axis provided values for the scale score as well as presenting the scores in a tabular form.

For each display, the number of statements of each type is provided in Table 4.2 to provide context. The number of statements of each type was dictated by the type of information provided by the display. So, while it was desirable to get all statements of all types for each display, some displays simply did not have the information to support those types of statements. This table should help provide some insight into the differences between the type of information provided in each display.

Table 4.2 Number of questions of each type for each display.

Display	Interpret Performance	Compare Performance	Identify Growth	Interpret Growth	Compare Growth
1: Table	0	0	0	1	5
2: Line	1	1	1	1	2
3: Text	2	0	0	4	0
4: Bar	1	1	0	3	1
5: Projection	1	1	1	1	2
6: Projection	1	1	1	1	2

Results

Each of the six displays was presented to a group of adults online, who were asked to respond to some factual statements about the display (by agreeing or disagreeing). As noted previously, every attempt was made to present statements as parallel as possible across the six displays, although there was some variation across displays due to the nature of the information presented in the respective displays. Statements were categorized in terms of cognitive level (identify, interpret, compare) and content (performance or growth). The table-type display only had information related to growth, and no information related to performance. Although the purpose of this study was to learn about best practices in reporting growth, some simple statements inquiring about general performance were included since this type of information was typically more straightforward and easy to interpret. By including these statements, we had some evidence as to whether the respondents could make the basic interpretations of the displays, and also to present statements on all aspects of the score report. This led to the five categories of items provided above. For each of these categories, the percent correct was computed for each display. The results are presented for the statements related to growth, as those are the statements that are relevant to this study.

The statements related to growth were combined into the three categories of Identifying Growth, Interpreting Growth, and Comparing Growth; the results were computed for each category instead of for each individual statement. When the Identifying Growth category was analyzed, there was data only for two types of displays: line and projection. For the line type display, the percent correct for this category was 78% and for the projection type it was 87% (both projection displays had exactly 87%). These differences were not statistically significant. For the Interpretation of Growth category, however, there were greater differences. The line, text, and bar-type displays produced results that were significantly higher in accuracy than the projection displays ($p<0.05$). Since the projection type had two versions, it is interesting to note that there was considerable variability within that display type. For example, for one projection-type display, the accuracy was 68% and for the other the accuracy was 52%. Given the complexity of interpreting the SGP, this is a very important finding; if the information was presented appropriately, the user was able to understand it; however, how that information was presented really did matter. This presentation was not just based on the type of graphic used, but also the way the information was displayed on the graphic.

In the category of Comparing Growth measures, respondents were asked to either compare different values of the SGP for a student across different years, or to compare the student's value of the SGP to the school/district/state SGP. In either case, the respondent was asked to compare two or more values of the SGP. The text type display did not contain information relevant to this category. Again, in this instance, there was a lot of variability among the displays with respect to how difficult the respondents found this task. On average, the projection type displays were most successful in presenting this type of information, with the highest accuracy (93%) as compared to the bar display (69%) and line display (52%). The differences between the projection type displays and the other two types was statistically significant ($p<0.05$). Within the projection type displays, there was some variability, although both displays had high rates of accuracy, 89% and 97%.

In summary, the short study indicated that there were differences in the respondents' skills in interpreting the statements about several displays of growth results. Since participants in the study were just adults in the general population, they were not stakeholders, and the results of this study might be a lower bound on those that might be obtained in cases where the respondents had a more acute interest in trying to understand the displays. Nonetheless, the study provided valuable information regarding how easily various displays of student growth could

be interpreted. As a general result, there was greater variability across the various displays when the statements asked were of a higher order of thinking (*Interpret* or *Compare*). Statements that required no real interpretation, such as identifying the performance level or growth level of a student, were typically easy for all respondents, regardless of the display used. In contrast, when asked to make interpretations, or compare measures, there was greater variability across the displays, indicating that some displays were more successful than others at communicating complex information.

Impact of Demographic Groups

When the results were compared across demographic groups, there were no statistically significant differences in the percent correct for any of the demographic groups. Specifically, there were no statistically significant differences in the percent correct depending on experience with statistics, gender, ethnicity, age or level of education. This lack of difference might be due in part to the small sample sizes in the groups. When looked at descriptively, there were some differences noted, and these provide some credibility to the results obtained. For most of the categories of statements, no interesting differences were noted across groups, however, for the interpreting growth category, some differences began to emerge. With respect to level of education, those with a professional degree (e.g. MD, or JD), the accuracy was higher than for those with less education. For those with the professional degree, the accuracy was approximately 90%, while for the other levels of education, it was closer to 70% with some categories slightly less than that. Similarly, those that took a statistics course at the graduate level were more accurate, approximately 87%, that those that took either no statistics courses, or statistics at the high school or undergraduate level, where the accuracy was approximately 67%.

Four Key Issues

As noted previously, growth model results reporting is something of a new frontier in reporting practices in K-12 testing, as these types of results have only started to become more widespread on reports in recent years, though the increase in use has been exponential. In reflecting on the displays available on various state education agency and testing company websites (which, though a non-scientific sample, offers some insight into state and local reporting practices), and considering the results of this small-scale study, it is clear that from a communication and reporting perspective, some important issues and considerations are emerging.

1. Complexity of Results

It is clear from the displays we have found that while the computation of many types of growth scores involves a series of statistical models and choices behind the scenes, and the mathematics of these approaches may not be readily accessible to many stakeholders (such as families and educators), it is possible to report the meaning of these scores in a way that can be understood by a lay audience. Not all of the approaches were successful in being able to communicate the meaning of the SGP, or how to interpret it, and as such, care must be taken in how that data are presented. Some specific features that are associated with greater understanding of the displays include:

- Clear definition of who is in the norm group for the SGP: In cases where that information was clearly presented, the respondents could correctly identify that the SGP reflected performance relative to some other group of students, not students in general.

- Prominence in layout and design: If this information is provided in text, the text should be of a large enough font size to draw attention to it and should be located centrally so that it is not ignored.
- Line graphs: In comparing the growth measures, the displays that utilized line graphs for comparing growth appeared to work better than other types of graphics, such as bar charts.

It is important to recognize that growth scores are a different kind of score for many intended users of the data, and as such present a communication challenge for testing agencies in informing users about the meaning of growth scores as well as appropriate interpretations and use. However, principles and insights gained from research on reporting scaled scores, or performance categories, is relevant. No information was gathered as to how to take action based on these growth scores, and as such, it is difficult to conclude how deeply the respondents understood the information. However, the data collected here indicate a different trend from that in previous studies (Clauser, Keller, & McDermott, 2016) where high school principals were not able to correctly interpret the SGP. This trend is encouraging that there are improvements being made in how this information is being communicated to the public.

2. Need for Interpretive Materials

Interpretative materials were not presented to the respondents in the small-scale study presented here, but clearly would aid in the interpretation of growth measures. In reviewing and surveying reporting strategies for growth reporting, it was evident that agencies responsible for reporting are at least in part aware of the potential for these reporting efforts to be misinterpreted. Two states, Hawaii and Virginia, have made explanatory/interpretive videos available on their state education websites that explicitly focus on providing stakeholders with details on growth models. These videos use features such as cartoon figures and analogies (bus, high jump) to make the concepts accessible to viewers. This is an innovative approach to reporting in general (and growth reporting specifically) that should be commended. Other states have released annotated report displays, illustrative guides, and PowerPoint presentations to provide further details on the mechanics and appropriates uses of these data. These approaches are innovative strategies, and we look forward to seeing further findings about report clarity and usefulness for users. Future research could, and should, focus on how these materials are used by stakeholders, and how their use affects the interpretation of score reports. These materials would be highly useful so long as the stakeholders are using them and applying the information correctly. Based on the study conducted here, interpretative materials should also provide guidance on how to *compare* different growth scores. These comparisons might be from subject to subject, from student to school/state/district or across years. These types of statements appeared to be especially problematic for the respondents in the study.

3. Error Reporting

Reporting error in test scores is an element of score reporting that has been implemented to varying degrees in different testing contexts, even for the relatively straightforward scale scores typically reported. It is also a topic receiving increased research attention (Zapata-Rivera, Kannan, and Zwick, this volume). Some states do provide standard error information in their current K-12 reporting efforts for scale scores, and others do not. Where information about score imprecision is not included, the reasoning behind that decision may be rooted in concerns about

adding complexity to reports and/or a belief that the presence of such information implies that a test is unreliable because error is associated with it. However, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) suggest that providing some information about the errors associated with test scores is necessary and responsible, and this guidance, in our opinion, should be extended to the statistical calculations of growth, which themselves are extrapolated from test scores. None of the displays that were used in this study contained any reference to errors in the growth scores, or how stable the growth scores were. Adding this type of information would bring the typical practice in line with the *Standards* and should be included. Future research on how stakeholders interpret the errors would be necessary to determine best practices around this reporting, although research conducted on reporting errors in test scores in general is germane.

4. Report Development Processes

Underlying all of the observations about growth reporting here is the idea that reports do matter, and that report development should follow a logical sequence of events that includes the solicitation of feedback from intended audiences. This argument has been advanced by Zenisky and Hambleton (2015); see also (Zenisky & Hambleton, 2012), in their proposed model for report development. As with all score reporting efforts, displays, and strategies for growth reporting should begin with a data-gathering phase including a statement of the purposes of any report along with the intended audiences, followed by report preparation, tryout with intended users, and a final phase of monitoring and improvement.

Summary

This brief chapter does not end with conclusions, because the conversation about growth model reporting is only beginning. As the use of growth model results for multiple audiences and purposes is increasing (for individuals, for groups of students, and for reporting of teacher quality), the importance of how these data are communicated will only increase. Rather, we raise the following broad questions, as a call for more research on how growth model results are used and understood.

- How can growth score results with their added complexity over single occasion test score be displayed so as to be readily understood and used by relevant audiences?
- How can errors in growth scores be communicated?
- How should growth scores be used by stakeholders? Are there specific actions that stakeholders should take based on these growth scores? These questions might help to clarify how they should be reported.
- Can tools be developed for interested stakeholders to plug in the information from their scores reports to get actionable information?
- How much do stakeholders use interpretative materials? While this applies to score reporting in general, given the additional complexity of growth measures, it might be even more relevant here. Are there ways to make interpretative materials more accessible?

There are many more questions that can be imagined, and we further note that the displays used in the study here are not representative of all of the approaches being used in the states. Accordingly, this topic remains an important direction for continued work on the reporting of growth.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Betebenner, D. W. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Clauser, A. L., Keller, L. A., & McDermott, K. A. (2016). Principals' uses and interpretations of student growth percentile data. *Journal of School Leadership*, *26*(1), 6–33.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Washington, DC: Council of Chief State School Officers.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*, 145–220. doi:10.1207/s15324818ame1702_3

Hambleton, R. K., & Zenisky, A. L. (2016). Advances in score reporting: Advances since 2008. Keynote address at the 10th meeting of the International Test Commission, Vancouver, BC, Canada.

Keller, L. A., Colvin, K. F., & Garcia, A. (2016). Growth: Measurement, meaning, and misuse. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 318–334). New York, NY: The Guilford Press.

National Education Goals Panel. (1998). *Talking about tests: An idea book for state leaders*. Washington, DC: U.S. Government Printing Office. Retrieved from http://govinfo.library.unt.edu/negp/reports/98talking.PDF

No Child Left Behind Act of 2001, P.L. 107–110, 20 U.S.C. § 6319 (2002).

O'Malley, K. J., Murphy, S., McClarty, K. L., Murphy, D., & McBride, Y. (2011). *Overview of student growth models*. (Test, Measurement, & Research Services White Paper). Iowa City, IA: Pearson. Retrieved from www.pearsonassessments.com/hai/Images/tmrs/Student_Growth_WP_083111_FINAL.pdf

Rick, F., Slater, S., Kannan, P., Sireci, S., Zenisky, A., & Dickey, J. (2016). Parents' perspectives on summative test score reports (Center for Educational Assessment Research Report No. 937). Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26.

Zenisky, A. L., & Hambleton, R. K. (2015). Test score reporting: Best practices and issues. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 585–602). New York, NY: Routledge.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*(2), 116–138.

5

Communicating Measurement Error Information to Teachers and Parents

Diego Zapata-Rivera, Priya Kannan, and Rebecca Zwick

Clearly communicating assessment results to the intended users so they can make appropriate use of this information is a central issue for assessment validity (Kane, 2013; Tannenbaum, this volume). The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) contains several guidelines on score reporting issues, including the need to provide interpretations of assessment information that are appropriate for the intended audience, evidence to support interpretations for intended purposes, information about recommended uses, and warnings about possible misuses. These standards address the responsibilities of test-developers in appropriately communicating assessment results, and the rights of test-users and test takers to understand and make use of this information appropriately.

Research on score reporting has produced some guidelines and iterative frameworks for designing score reports that meet the *Standards* (Goodman & Hambleton, 2004; Hambleton & Zenisky, 2013; Hattie, 2009; Underwood, Zapata-Rivera, & VanWinkle, 2007; Wainer, 2014; Wainer, Hambleton, & Meara, 1999; Zapata-Rivera, VanWinkle, & Zwick, 2012; Zenisky & Hambleton, 2016). These guiding principles usually involve steps for evaluating score reports with the intended audience. Additional information about some of these frameworks can be found in Brown, O'Leary, and Hattie (this volume), O'Donnell and Sireci (this volume), and Tannenbaum (this volume).

Because the characteristics of score reporting audiences vary, different approaches to facilitating comprehension of assessment information for the intended audience have been explored. Zapata-Rivera and Katz (2014) suggest focusing on the *needs, knowledge,* and *attitudes* of the audience as an important step in the process of designing interactive score report components. These components may include the use of graphical representations, written explanations, examples, on-demand help, video tutorials designed to address misconceptions, and navigation approaches.

Given the importance of information about the precision of the test scores (i.e., measurement error) to inform decision making, the *Standards* include language on the need for providing interpretations on what the scores represent, their precision, and how they are intended to be

used. Prior work has shown that communicating measurement error information with test-users and test takers can be challenging (Hambleton & Slater, 1997; Kannan, Zapata-Rivera & Leibowitz, in press; Lukin, Bandalos, Eckhout, & Mickelson, 2004; Zapata-Rivera, VanWinkle, & Zwick, 2010; Zapata-Rivera & Zwick, 2011; Zwick et al., 2008). Zwick, Zapata-Rivera, and Hegarty (2014) described user difficulties in comprehending verbal and graphical representations of measurement error. There is no one-size-fits-all visual representation that is uniformly understood by all users, and no universal line of text that clearly conveys what standard errors are and why they are important (Zenisky & Hambleton, 2016).

In this chapter, we examine research on designing and evaluating score reports for teachers and parents. In particular, we focus on issues regarding the communication of measurement error information with these audiences. Finally, challenges and opportunities for continuing research in this area are identified and discussed.

Teachers and Parents as Two Different Audiences

Even though both teachers and parents want to receive information about students' test performance, their needs and uses of assessment information may differ. For example, teachers may be interested in both classroom and individual-level information while parents are likely to be interested only in their individual child's performance. Teachers may use the assessment results to inform grouping and instructional planning, while parents may use this information to support their conversations with their child's teacher and to obtain appropriate support for their child. In the past, score reports for parents were sometimes regarded as simpler versions of those produced for teachers. While reports resulting from this approach may have provided parents with test-relevant information, these reports were not necessarily designed with parents' needs and levels of understanding in mind (Barber, Paris, Evans, & Gadsden, 1992).

Needs, Knowledge, and Attitudes of Teachers

Work on applying audience analyses to the design of score reports for teachers has considered their needs for assessment information, their prior knowledge about assessment, and their attitudes toward score report information (Zapata-Rivera & Katz, 2014). In particular, teachers need information that can be used to guide instruction (Underwood, Zapata-Rivera, & VanWinkle, 2007). This requirement has been referred to as "who needs to be taught what next" (Brown, O'Leary, & Hattie, this volume). The questions teachers are interested in including: How did the class perform on the test? What are my students' strengths and weaknesses? How does a particular student's score compare to other students' scores? How difficult were the tasks for students? And what should I do next to help an individual student or the class as a whole?

In general, cognitive aspects of the user such as perception, attention, working memory and prior knowledge play an important role on users' comprehension of graphical representations (Hegarty, this volume). In the field of score reporting, we have found that after reading interpretive materials, teachers usually have the knowledge required to understand most of the information typically included in score reports (e.g., scores, score means, percentiles) (Zapata-Rivera, VanWinkle, & Zwick, 2012). However, the use of technical language may interfere with proper understanding of score report information (Hambleton & Slater, 1997; Underwood, Zapata-Rivera, & VanWinkle, 2007; Zapata-Rivera, VanWinkle, & Zwick, 2012). In addition, teachers may have limited knowledge of the concept of measurement error and how to use it to inform their decisions (Zwick, Zapata-Rivera, & Hegarty, 2014; Zapata-Rivera, Zwick, & Vezzu, 2016).

In terms of attitudes, teachers may value clear and direct answers to their assessment questions, since many of them have limited time to explore assessment results (Zapata-Rivera,

Hansen, Shute, Underwood, & Bauer, 2007) and may place too much trust in the precision of scores (Zapata-Rivera, VanWinkle, & Zwick, 2012; Zwick et al., 2008). Information about teacher attitudes toward assessment in general can be found in Goertz, Oláh, and Riggan (2009), Mandinach and Gummer (2016), and Marshall and Drummond (2006).

Needs, Knowledge, and Attitudes of Parents

Parents and guardians are keen to understand how their child has performed on a test, what the scores mean, and what they can do to help support their child to improve performance in the future (Kannan, Zapata-Rivera, & Leibowitz, in press). In order to support parents' interpretations and uses of the information conveyed in score reports, it is important that their unique needs, pre-existing knowledge, and attitudes are taken into consideration.

Research on identifying parents' needs have found that, across the board, parents are most interested in understanding what each score on the report means, how their child performed against set standards, their child's performance level (e.g., basic, proficient, and advanced) and the implications of that placement (A-Plus Communications, 1999; Kannan, Zapata-Rivera, & Leibowitz, in press; Munk & Bursuck, 2001; NEGP, 1998). Beyond that, research evaluating score reports with a diverse subgroups of parents (Kannan, Zapata-Rivera, & Leibowitz, in press) has found that what parents list as their second-most important need varies across demographic subgroups. While parents with a college degree wanted to see how their children were performing in relation to other students (i.e., normative comparisons) in their own and other schools within their district or state (A-Plus Communications, 1999; Kannan, Zapata-Rivera & Leibowitz, in press), parents with no college degree listed areas that needed improvement and ways to help their child as the second-most important information (Kannan, Zapata-Rivera, & Leibowitz, in press).

With regard to comprehension of information presented in score reports, parents typically do not have the assessment-related and technical background required to appropriately interpret and use the test results presented for their child (Barber et al., 1992). Parents struggle to fully understand some of the information that is typically included in individual student reports (ISRs). These include information about their child's performance in subareas, their child's growth across the years, and the measurement error involved in their child's scaled score (Kannan, Zapata-Rivera, & Leibowitz, in press; Rick et al., 2017). These pieces of information have ranked much lower on parents' listed needs (A-Plus Communications, 1999; Kannan, Zapata-Rivera & Leibowitz, in press; Munk & Bursuck, 2001; NEGP, 1998) when compared to information about their child's overall score and performance level placement.

In terms of attitudes, Barber et al. (1992) found that only 53% of the 105 parents surveyed thought that the assessment contributed to their child's education. A recent poll by Phi Delta Kappa (PDK) and Gallup (2015) showed that even though 71% of public-school parents across the country felt that using tests to measure what students have learned was important for improving public schools in their community, 67% of respondents felt that there was too much emphasis on testing in their children's schools—a perception that has remained somewhat persistent through the last couple of decades (see A-Plus Communications, 1999; Bennett, 2016; PDK & Gallup, 2015). Additional information about parent attitudes about assessment can be found in Harris and Brown (2016).

Score Reports for Teachers and Parents

In this section, we describe the types of score reports that are usually designed for two key audiences—teachers and parents.

Common Features of Score Reports for Teachers

Score reports for teachers usually include classroom-, individual-, and task/item-level score reports. These reports can be available as online interactive reports or printed documents. As an example of the type of information included in teacher score reports, we describe the score report prototypes that were designed and evaluated as part of the Cognitively Based Assessment of, for, and as Learning (CBAL™) research initiative (Bennett & Gitomer, 2009). Even though this section focuses on teacher score reports used as part of the CBAL research initiative, most operational score reports for teachers include a subset of the score report elements presented here. More information about these reports can be found in Zapata-Rivera, VanWinkle, & Zwick (2012).

- Individual-level reports provide information to answer the following questions: *How did student X do on the test? How did students in the same grade do on the test? And what should be done next?* Information in the CBAL individual report includes a general description of the test and the sections of the report, a personal identification section with information such as student name, grade, teacher name, subject, and test date; a section with appropriate and inappropriate uses of the assessment information presented in this report; a performance summary section with information such as scaled scores, confidence bands, performance levels, a distribution of scores for students in the same grade, and additional materials (e.g., explanations of statistical terms used in the report); a current test performance section that includes information such as raw scores and links to relevant information (e.g., skill definitions, sample tasks, and explanations of statistical terms used in the report); and a "What to do Next?" section that provides a summary of how the student did, information about the next performance level and recommendations for teacher follow-up.
- Classroom-level reports respond to the question, *how did students in my classroom perform on the test?* The CBAL classroom report includes an introduction, a section with appropriate and inappropriate uses of this report and classroom performance information presented as a sortable table including score and performance level information for each student in the class and a graph showing the score distribution of the class among performance levels. This report also includes links to explanations and definitions of some terms.
- Task/item-level reports provide information to answer the question, *how did my students do on this task?* The CBAL item-level reports include an introduction section, a section with appropriate and inappropriate uses, and a table with item/task difficulty information. In addition, each task/item is linked to information about related content and process skills.

Several aspects of the report (e.g., navigation, additional explanations, and lists of appropriate and inappropriate uses) are organized according to the needs and attitudes of this audience in order to facilitate finding the information needed to inform decisions and minimize opportunities for misuse. Even though these teacher reports were designed to provide score report information after each of the interim assessments that comprised a larger assessment system, CBAL has developed other reporting systems for formative purposes that can serve as learning tools allowing teachers to assign tasks to students and provide immediate feedback (Zapata-Rivera, 2011).

Common Features of Score Reports for Parents

Score reports for parents and guardians are always intended to provide results for their individual child—these reports can either be based on formative or summative assessments (Kannan,

Bryant, Zapata-Rivera, & Peters, 2017). Results presented to parents may be evaluative and present summary information about their child's performance on a standardized assessment at the end of a learning period. Alternatively, the results provided may be intended to support decisions about placement on advanced classes or remediation for their children (Kannan, Bryant, Zapata-Rivera, & Peters, 2017). Finally, though score reports designed for parents have traditionally been static, printed score reports, they could also be interactive score reports with layers of drillable information to accommodate the varied needs of this extremely diverse stakeholder group who vary in education level, language proficiency, socio-economic background, and an array of other variables (Kannan, Zapata-Rivera & Leibowitz, in press).

Overall, parent score reports based on summative or end-of-year assessments are intended to answer the following questions: *How did my child do on the test? How did other students in the same grade (in his school, and in the state and district) do on the test? What are my child's general strengths and weaknesses? Has my child shown any improvement since last year? How can I help my child or where can I get more help?* Parent score reports based on formative assessments are typically designed to answer the questions such as: *What are my child's specific strengths and weaknesses? Does my child have any specific weaknesses that we should work on with the teacher? Where can I get more help to support my child's growth in this area?*

Research (e.g., Goodman & Hambleton, 2004; Hambleton & Zenisky, 2013; Kannan, Zapata-Rivera, & Leibowitz, in press; NEGP, 1998; Zapata-Rivera & VanWinkle, 2010; Zenisky & Hambleton, 2012) has suggested that score reports designed for parents should include a range of test- and performance-related information, such as (a) a description of the purpose of the test; (b) a snapshot or at-a-glance summary in the beginning; (c) a personal identification section with information such as student name, grade, teacher name, subject, and test date; (d) cues, hints, definitions, and extended descriptions when technical language is used; (e) scores represented as graphics with colored bars, so parents find it easy to make comparisons; (f) norm-referenced comparative information for other students in the same grade (within the school, district, and state); (g) a description of the nature and precision of scale scores in an unambiguous manner (i.e., measurement error); (h) the student's performance across the subareas tested, with a detailed description of each subarea; (i) a listing of the types of knowledge and skills (as well as examples of items) that the student has mastered and currently struggles with; and (j) information about next steps, and where to get additional help for their child. Finally, it has also been recommended (Goodman & Hambleton, 2004; Kannan, Zapata-Rivera, & Leibowitz, in press) that significant efforts should be made to limit overall use of technical language by simplifying and streamlining the text. If possible, score reports for parents should be made available in multiple languages.

Communicating Information About Measurement Error

Representing and communicating uncertainty is a topic of interest in in several disciplines (Correll & Gleicher, 2014; Demmans Epp & Bull, 2015; Hopster-den Otter, Muilenburg, Wools, Veldkamp, & Eggen, 2018, Ibrekk & Morgan, 1987, Spiegelhalter, Pearson, & Short, 2011). Clearly communicating uncertainty is important in order to support evidence-based decision making. An understanding of the level of uncertainty of a particular event can play an important role when making decisions based on scientific data (e.g., deciding whether to evacuate before a hurricane or comparing medical treatments; Fischhoff & Davis, 2014).

In educational assessment, appropriate communication of uncertainty is particularly important when educational decisions are to be made on the basis of test scores. Measurement error information associated with the test scores can provide the knowledge test-users need to make a particular decision. For example, a test user may wonder whether two test scores are

meaningfully different. Although, in some cases measurement error information is not presented due to possible misinterpretations, educating the general public on how to understand and use this information may increase transparency and confidence (Newton, 2005).

Communicating Information About Measurement Error to Teachers

In this section, we describe two studies in which researchers investigated ways of communicating measurement error information to teachers. Both of these studies are an exploration of *comprehension* and *preference* aspects of the intended audience and used a variety of research methods (e.g., usability studies, pilot studies, and large-scale studies). Both comprehension and preference data provide important insights on the types of misconceptions teachers have and the types of supports that can be put in place to help them understand and use measurement error information appropriately.

Zwick, Zapata-Rivera, and Hegarty (2014) explored the use of verbal and graphical representations of measurement error intended to help teachers understand and make appropriate decisions based on test-score information. The research participants, 148 teachers and 98 introductory psychology students, were asked to view score reports that included varying representations and descriptions of measurement error. Verbal descriptions included analogies between measurement error in test scores and error in measuring weight and blood pressure. Graphical representations of measurement error included a standard error bar and a tapered confidence band (see Figure 5.1). Participants were randomly assigned to one of four conditions (two verbal descriptions crossed with two graphical representations). As suggested in the literature on score reporting, participants were asked both preference and comprehension questions (Wainer, Hambleton, & Meara, 1999; Zenisky & Hambleton, 2012). Results showed that participants who reported greater comfort with statistics tended to have higher comprehension scores and to prefer the tapered confidence band. Several misconceptions about measurement error were identified such as the belief that test scores are perfectly precise or that the level of certainty was constant across the confidence band. Some participants assumed that confidence bands for test scores must be based on multiple observed scores from a single individual or from a group of test takers. Participants mentioned the need for explanations on the meaning of confidence bands and the use of information that could be used to support decision making (information that was intentionally omitted in the study).

A follow-up study exploring the effectiveness of a short, web-based tutorial in helping teachers to better understand the measurement error information in test-score reports was carried out (Zapata-Rivera, Zwick, & Vezzu, 2016). Participants were 145 K-12 teachers across a variety

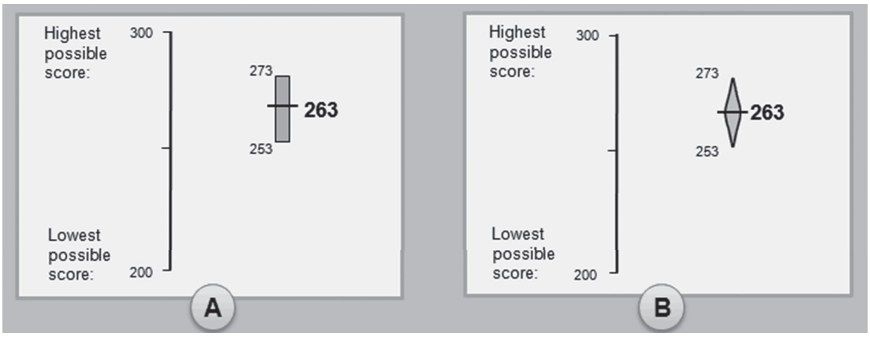


Figure 5.1 A standard error bar and a tapered confidence band.

of subject areas including mathematics, English language arts, science, and foreign language. Two short video tutorials were created. The basic tutorial included simple definitions, examples of the causes of measurement error, illustrations of confidence bands, and explanations of how to interpret them. The enhanced version included additional screens showing how a confidence band is obtained. Results showed that participants who were assigned to the tutorial conditions (basic and enhanced) significantly outperformed those assigned to the control condition (no tutorial) in the comprehension questionnaire. The proportion of variance in comprehension scores that was attributable to experimental condition (eta-squared) was .23, but the difference between the two tutorial conditions was not statistically significant. Results of a usability questionnaire administered to those participants in the tutorial conditions showed that most would like to use this type of tutorial in the future and found the tutorial useful (96%), easy to understand (93%), and engaging (90%). The majority believed they learned a lot from it (85%) and reported that they understood what a confidence band represents (97%). These results showed that potential of instructional materials like these to provide teachers with clear information that helps them understand score report information and use it in appropriate ways.

Lessons learned in terms of the research methodology employed with teachers were applied to a different audience, parents. The next section discusses work on communicating measurement error information with parents.

Communicating Information About Measurement Error to Parents

Whether it is useful to include information about measurement error (or score precision) in ISRs primarily intended for parents has been a controversial issue. In particular, the *Standards* (AERA, APA, & NCME, 2014) and several researchers (e.g., Faulkner-Bond, Shin, Wang, & Zenisky, 2013; Zapata-Rivera, Zwick, & Vezzu, 2016) have specifically recommended that a description of the nature and precision of scale scores be presented in an unambiguous manner in ISRs. Although Wainer, Hambleton, and Meara (1999) have recommended that it is best to omit information about measurement error on ISRs unless this information can be presented in a way that leads to accurate interpretations and appropriate uses.

Until recently, there was very little evidence in the research literature about the steps taken to explain or quantify error and uncertainty when reporting test results (to any stakeholder group). In their survey of international score reports, Bradshaw and Wheater (2009) found that while descriptive information (e.g., overall score, grades) was easy to find on most reports, it was almost impossible to find any explanations of reliability or measurement error in most of the reports they reviewed. However, more recently there has been increasing research in how to represent and communicate measurement error information with parents (Kannan, Zapata-Rivera, & Leibowitz, in press; Kannan, Bryant, Zapata-Rivera, & Peters, 2017; Zapata-Rivera, Vezzu, & Biggers, 2013; Zapata-Rivera et al., 2014).

In practice, there has been vast variation across states in the amount and nature of information about score precision that is provided in ISRs for standardized assessments. Several states do not provide information about error or precision of scores in their score reports. For example, after reviewing score reports for 41 states, Faulkner-Bond et al. (2013) found that only two states provided information about measurement error for their English Language Proficiency (ELP) assessments. Other states that do provide information about measurement error typically do not provide a clear explanatory text. More recently, ISRs designed for parents have started to include information about measurement error. However, studies evaluating the interpretation and use of this information by parents, as a diverse and heterogeneous stakeholder group, are minimal.

Consistent with the suggestion to understand stakeholder needs (e.g., Hambleton & Zenisky, 2013), we have conducted studies at ETS (Kannan, Zapata-Rivera, & Leibowitz, in press;

Kannan, Bryant, Zapata-Rivera, & Peters, 2017) to understand how to appropriately communicate measurement error information to parents. Similar to the studies with teachers (Zwick, Zapata-Rivera, & Hegarty, 2014; Zapata-Rivera, Zwick, & Vezzu, 2016), the studies with parents involved the use of both comprehension and preference questions and a variety of research methods (e.g., cognitive laboratories and usability and experimental studies comparing various graphical and verbal representations).

In an early study with parents, Kannan, Zapata-Rivera, and Leibowitz, (in press) used cognitive laboratories to investigate the extent to which potential users could understand and interpret various pieces of information (including measurement error) presented in a hypothetical ISR. Participants were 35 parents from diverse subgroups (disaggregated by education level and English language proficiency). Results from that study suggested that parents across four subgroups defined by education level (i.e., those with and without a college degree) and English language proficiency, struggled to understand the information presented about measurement error in particular. Even though parents were allowed to refer to the hypothetical score report in answering the comprehension questions, about 50% or more of the parents in each subgroup, even those with college degrees, were not able to accurately read back or reiterate the information presented about score precision. Parents in this study pointed out that even though the information about "precision/error" was understandable and perhaps even useful, it could be confusing and overwhelming to parents in general, and that they were not sure if parents should/would care about this information.

Therefore, in a follow-up study, Kannan, Bryant, Zapata-Rivera, and Peters (2017) used a between-subjects experimental design to evaluate parents' comprehension of measurement error information. Specifically, they sought to determine whether parents understood information about measurement error and whether they would find this information useful in making appropriate inferences about their child's performance. 196 parents of middle school children were randomly assigned to three conditions in an online experiment: (a) a condition where no error information was presented; (b) a condition where measurement error was presented graphically with a bar around the score (one standard error of measurement above and below the observed score) and a standard footnote typically used in state standardized assessment reports; and (c) a condition where measurement error was presented graphically as in the previous condition but with a more detailed (enhanced) footnote describing the various factors that could affect a child's score on any given test administration. The researchers did not, however, describe how these error bars are computed (or the percent confidence) to the participants in any of our study conditions. Once participants answered the comprehension questions, they had the opportunity to examine all three different representations of measurement error and indicate which representation they would prefer included in their child's score report. Results from this study suggest that parents are highly receptive to information about measurement error, and that that between 58% and 79% of parents across all three between-subject study conditions (irrespective of the type of information they received during the rest of the study) preferred the representation with the most information (i.e., the enhanced error representation). Moreover, when provided with a detailed written explanation, parents were more likely to understand this information (with higher overall comprehension scores) than when such explanation was not provided.

The results from the second parent study (Kannan, Bryant, Zapata-Rivera, & Peters, 2017) can be interpreted as evidence that parents are not only trying to understand the information presented about measurement error (indicated by the higher comprehension scores for the "standard" and the "enhanced" conditions), but also want to try and use this information to better understand their child's performance on standardized assessments. Overall, from these results, we glean that it is important to provide clear and detailed information to parents so that they are able to easily understand this information.

Discussion

Here we offer a discussion of several important themes that have emerged from the score reporting research.

Sharing Research Methods and Materials

Even though each audience has its own characteristics, it is possible to use similar methodological approaches (e.g., usability studies, cognitive laboratories, focus groups, interviews, and controlled studies) and data collection materials (e.g., preference and comprehension questionnaires) in studying them. In some cases, insights gained from doing research with one audience can inform future work with other audiences. For example, teachers may suggest ideas about materials they have used or would like to use to share assessment information with parents or students.

Preference and Comprehension

When creating score reports for different audiences, it is important to consider both preference and comprehension issues since it is possible that the preferred display may not be the best-understood one. In terms of interactive, computer-based reports, usability studies and cognitive laboratories provide interesting information about issues that may hinder the interaction. Also, these research approaches provide useful data on the cognitive processes that users exhibit when trying to understand the assessment results provided in the report. These studies also provide an opportunity to pilot-test data collection materials that will be used in large-scale studies.

More Research Needed

More research is needed in the area of score reporting. Score reports may include "legacy" score report elements that do not clearly communicate assessment information to particular audiences. Research on the effectiveness of particular score report elements to communicate assessment information should be conducted.

Other potential areas of research include exploring the trade-off between achieving simplicity of reports (e.g., by hiding information that might result in misinterpretations) and using instructional and training materials, such as video tutorials, to facilitate understanding of important assessment information. More research on exploring the effectiveness of instructional materials for teaching different assessment concepts to different audiences is also needed.

Summary

Research on score reports involves exploring how to present assessment information to different audiences. When designing and evaluating score reports and additional materials, it is important to take into account the needs, knowledge, and attitudes of the audience and to pay attention to both preference and comprehension issues to capture a complete picture of the benefits and drawbacks of the score report elements being studied.

The work on communicating measurement error information with teachers and parents provides a good use case where the characteristics of the audience have been taken into account to design different types of graphical representations and supporting materials. Although the final reports for parents and teachers may look completely different, this work shows that it is possible to apply similar research methods and materials with different audiences.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

A-Plus Communications. (1999). *Reporting results: What the public wants to know*. A companion report to Education Week's Quality counts'99. Retrieved from www.edcounts.org/archive/sreports/qc99/opinion/edweekresults.pdf

Barber, B. L., Paris, S. G., Evans, M., & Gadsden, V. L. (1992). Policies for reporting test results to parents. *Educational Measurement: Issues and Practice*, *11*(1), 15–20.

Bennett, R. E. (2016). *Opt out: An examination of issues* (Research Report No. RR-16–13). Princeton, NJ: Educational Testing Service. Retrieved from http://dx.doi.org/10.1002/ets2.12101

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K—12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Assessment issues of the 21st century* (pp. 43–61). New York, NY: Springer.

Bradshaw, J., & Wheater, R. (2009). *National foundation for educational research: International survey of results reporting (OFQUAL 10/4705)*. London: Office of Qualifications and Examinations.

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization & Computer Graphics*, *20*(12), 2141–2151.

Demmans Epp, C., & Bull, S. (2015). Uncertainty representation in visualizations of learning analytics for learners: Current approaches and opportunities. *IEEE Transactions on Learning Technologies*, *8*(3), 242–260.

Faulkner-Bond, M., Shin, M., Wang, X., & Zenisky, A. L. (2013, April). *Score reports for English proficiency assessments: Current practices and future directions*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

Fischhoff, B., & Davis, A. L. (2014). Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences*, *111* (Supplement 4), 13664–13671.

Goertz, M. E., Oláh, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction*. Philadelphia, PA: Consortium for Policy Research in Education.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145–220.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Hambleton, R., & Zenisky, A. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In *APA handbook of testing and assessment in psychology*. Washington, DC: American Psychological Association.

Harris, L. R., & Brown, G. T. L. (2016). Assessment and parents. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 1–6). Singapore: Springer.

Hattie, J. (2009, April). *Visibly learning from reports: The validity of score reports*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA.

Hopster-den Otter, D., Muilenburg, S. N., Wools, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2018). Comparing the influence of various measurement error presentations in test score reports on educational decision-making. *Assessment in Education: Principles, Policy & Practice*, 1–20.

Ibrekk, H., & Morgan, M. G. (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, *7*(4), 519–529.

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

Kannan, P., Zapata-Rivera, D., & Leibowitz, E. A. (in press). Interpretation of score reports by diverse subgroups of parents. *Educational Assessment*.

Kannan, P., Bryant, A. D., Zapata-Rivera, D., & Peters, S. (2017, April). *Evaluating parent comprehension of measurement error presented in score reports*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Antonio, TX.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, *23*(2), 26–32.

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*, 366–376.

Marshall, B., & Drummond, M. J. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education*, *21*, 133–149.

Munk, D. D., & Bursuck, W. D. (2001). What report card grades should do and communicate. *Remedial and Special Education*, *22*(5), 280–287.

National Education Goals Panel, NEGP. (1998). *Talking about tests: An idea book for state leaders*. Washington, DC: U.S. Government Printing Office. Retrieved from http://govinfo.library.unt.edu/negp/reports/98talking.PDF

Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, *31*(4), 419–442.

Phi Delta Kappa, & Gallup. (2015). The 47th PDK/gallup poll of the public's attitudes toward the public schools: Testing doesn't measure up for Americans. *Phi Delta Kappan*, *97*(1). Retrieved from http://pdkpoll2015.pdkintl.org/

Rick, F., Kannan, P., Slater, S., Sireci, S., Zenisky, A., & Dickey, J. (2017, April) *Parent perspectives on summative score reports*. Paper presented at the 2017 annual meeting of the National Council for Measurement in Education.

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393–1400.

Underwood, J. S., Zapata-Rivera, D., & VanWinkle, W. (2007). *Growing pains: Teachers using and learning to use IDMS* (Research Memorandum 08–07). Princeton, NJ: Educational Testing Service.

Wainer, H. (2014). Visual revelations: On the crucial role of empathy in the design of communications: Genetic testing as an example. *Chance*, *27*, 45–50.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36*, 301–335.

Zapata-Rivera, D. (2011, April). *Designing score reports that help teachers make instructional decisions*. Paper presented the annual meeting of the Educational Research Association conference (AERA), New Orleans, LA.

Zapata-Rivera, D., Hansen, E. G., Shute, V. J., Underwood, J. S., & Bauer, M. I. (2007). Evidence-based approach to interacting with open student models. *International Journal of Artificial Intelligence in Education*, *17*, 273–303.

Zapata-Rivera, D., & Katz, I. (2014). Keeping your audience in mind: Applying audience analysis to the design of score reports. *Assessment in Education: Principles, Policy & Practice*, *21*, 442–463.

Zapata-Rivera, D., & VanWinkle, W. (2010). *A research-based approach to designing and evaluating score reports for teachers* (Research Memorandum 10–01). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2010, May). *Exploring effective communication and appropriate use of assessment results through teacher score reports*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL™ teachers* (Research Memorandum 12–20). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., Vezzu, M., & Biggers, K. (2013, May). *Supporting teacher communication with parents and students using score reports*. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

Zapata-Rivera, D., Vezzu, M., Nabors Olah, L, Leusner, D., Biggers, K., & Bertling, M. (2014, April). *Designing and evaluating score reports for parents who are English language learners*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Philadelphia, PA.

Zapata-Rivera, D., & Zwick, R. (Eds.). (2011). *Improving test score reporting: Perspectives from the ETS score reporting conference* (Research Report 11–45). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*(3), 215–229. doi:10.1080/10627197.2016.1202110

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26.

Zenisky, A. L., & Hambleton, R. K. (2016). A model and good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 585–602). New York, NY: Routledge.

Zwick, R., Sklar, J., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice*, *27*(2), 14–27.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*(2), 116–138. doi:10.1080/10627197.2014.903653

Part II
Practical Applications

6

Score Reporting Issues for Licensure, Certification, and Admissions Programs

Francis O'Donnell and Stephen G. Sireci

Testing the knowledge, skills, and abilities of people has a long history, almost as long as recorded history itself. For example, the story of Adam and Eve in the book of Genesis in the Bible could be considered the earliest recorded "test." Later, in the same book, there is a second reference to a test when God "tested" Abraham by asking him to sacrifice his son Isaac (around 2100 BC). With respect to large-scale testing, the Book of Judges (12:4–6, around 1400 BC) describes the one-item "test" developed by Gileadites to identify the enemy Ephraimites who were hiding among them (the test was to pronounce the word "shibboleth"). Although these events are of historical interest in considering when testing first occurred, there are no formal records of the results of these performance assessments.

The focus of this chapter is on the results of testing—score reports—for licensure, certification, and admissions testing programs. We will focus on current score reporting practices, and so we will not attempt to trace the history of score reporting practices to their origin. However, our review of these practices suggests that the first recorded results of credentialing testing may have literally been carved in stone. For example, "score reports" from the Le Thanh Tong dynasty in Vietnam (1484 AD) can still be seen today in the Temple of Literature in Hanoi. All candidates who passed the rigorous steps to be selected to work in the central imperial government under the Emperor had their names and hometown carved into huge steles in the shape of a turtle, which emphasized their longevity and wisdom. According to Nguyen (2009), this feudal examination model originated from the testing approach used in China for civil and military testing, and there are similar steles in the Temple of Literature in Beijing.

Today, score reports are not carved in stone. In fact, in many instances they are not even printed on paper. Instead, many testing programs in credentialing and admissions testing programs provide score reports in digital formats via a URL that examinees, parents, and other stakeholders access and interact with to acquire various levels of detail regarding their performance on a test. In this chapter, we describe current score reporting practices in credentialing and admissions testing, discuss some of the practical issues and validity issues involved in reporting test results in these areas, and provide suggestions for future research and practice.

Before beginning our review, it is important to define the terms we use for different testing contexts. *Admissions* tests refer to tests that have the primary purpose of providing information to those who make admissions decisions at various schools, such as selective high schools, colleges, universities, and postsecondary schools (e.g., medical schools, law schools, business schools, other graduate programs). *Licensure* tests refer to exams developed as part of a professional licensure requirement that is needed for practice within a profession. Examples include the Uniform Certified Public Accountants Exam for accountants, the National Bar Exam for lawyers, the United States Medical Licensure Exam for medical doctors, and the National Council Licensure Examination for nurses. *Certification* tests refer to tests used to award certificates to candidates to certify competence or excellence, independent of a licensure requirement. Examples of certification exams include those used in the technology industry to certify competence in working with hardware or software (e.g., Microsoft, Cisco, Hewlett Packard exams), career and technical education (e.g., automotive, culinary exams), and accomplished teaching beyond the licensure stage (e.g., National Board of Professional Teaching Standards). Because the issues and practices in licensure and certification testing are so similar, the more general term *credentialing testing* can be used to describe both contexts.

Current Practices in Score Reporting in Credentialing and Admissions Testing

There is great variety in the content and design of score reports for credentialing and admissions tests. Within the same field, some reports use a simple "letter" format, while others consist entirely of tables with numbers. To describe current practices in score reporting for admissions and credentialing tests, we conducted online searches of score reports from admissions testing and licensure testing programs, and we contacted 60 certification programs from a list of organizations accredited by the National Commission for Certifying Agencies (Institute for Credentialing Excellence, 2017).

Through these efforts we were able to locate score reports from 38 testing programs: 22 certification programs, eight admissions programs, and eight licensure programs. We did not use random sampling to select these score reports and so they cannot be considered representative of these areas. Nevertheless, they illustrate a wide variety of examples of current score reporting practices across multiple domains. The 22 certification score reports came from programs in nursing specialty areas (five reports), fitness and exercise (three reports), pharmacy specialties (two reports), occupational therapy and rehabilitation (two reports), other health-related specialties (seven reports), culinary arts (one report), real estate management (one report), and safety (one report). The eight reports for admissions tests came from programs designed to inform admissions into professional school (three reports), middle school (two reports), undergraduate colleges and universities (two reports), and graduate school (one report). Lastly, the eight licensure reports represented programs supporting licensure in financial services (three reports), teaching (three reports), and health sciences (two reports).

In the next section we summarize the features of those reports and the types of information presented for each context. The reports typically provide information pertaining to examinees' performance on the exam overall, their performance in specific subdomains, and general interpretive guidance. A summary of the elements included in these reports, stratified by testing context, is presented in Table 6.1. Although there are some similarities in the types of information presented, there are also notable differences across these testing contexts. For the certification context, where we had the most responses, different types of information were reported depending on whether the candidates passed or failed the exam, so separate results for "pass" and "fail" reports are presented. We begin with a summary of the information provided in the reports from admissions testing programs.

Table 6.1 Types of Information Included in 38 reviewed score reports.

	Testing Context			
Information Provided	**Admissions**	**Certification**		**Licensure**
	(n = 8)	**Pass** (n = 22)	**Fail** (n = 22)	(n = 8)
Overall Results				
Numerical score	8 (100%)	13 (59%)	19 (86%)	6 (75%)
Performance levels[1]	2 (25%)	22 (100%)	22 (100%)	8 (100%)
Performance level descriptions	2 (25%)	0 (—)	0 (—)	0 (—)
Information about precision	5 (63%)	0 (—)	0 (—)	1 (13%)
Visual display	7 (88%)	4 (18%)	7 (32%)	2 (25%)
Subdomain Results				
Numerical score(s)	4 (50%)	7 (32%)	15 (68%)	1 (13%)
Performance levels	1 (13%)	1 (5%)	8 (36%)	6 (75%)
Performance level descriptions	1 (13%)	0 (—)	3 (14%)	5 (63%)
Information about precision	1 (13%)	2 (9%)	6 (27%)	3 (38%)
Visual display	3 (38%)	7 (32%)	19 (86%)	6 (75%)
Interpretive Information				
Statement of test purpose	1 (13%)	0 (—)	0 (—)	2 (25%)
Guidance on next steps	5 (63%)	16 (73%)	19 (86%)	3 (38%)
Details about where to find additional resources	6 (75%)	7 (32%)	15 (68%)	6 (75%)

[1] For certification and licensure reports, "pass" and "fail" were considered performance levels.

Score Reporting Features in Admissions Testing Programs

All eight reports for the admissions tests presented results for multiple subject areas (e.g., quantitative reasoning, verbal reasoning) and included numerical total scores for each area (which we refer to as "overall scores"). Six reports (75%) presented composite scores, or total scores across subject areas, in addition to single-subject overall scores. Among those, three (50%) presented both types of scores with the same level of detail. However, there were two reports in which performance levels were only provided for overall scores, and one report in which information about precision was reported for overall scores but not composite scores (note that the ways in which composite scores were presented are only described in text; the "Overall Results" section of Table 6.1 focuses on overall scores). Subscores were provided in four out of the eight (50%) admissions testing reports.

There were several patterns in how overall scores were presented. In every case, both a scaled score and its corresponding percentile were included. One report also provided stanines. Notably, the score reports for the two undergraduate admissions tests categorized overall scores in relation to college readiness benchmarks, a unique feature reflecting the programs' goal. Almost all reports (seven, or 88%) included a visual display of overall performance—either a table with numbers or a horizontal bar where a symbol or band denoted the location of the test taker's score. Five reports (63%) included information about overall score precision: two used written explanations only, two used both a written explanation and a visual representation of measurement error (e.g., a score band where the width of the band reflected the amount of imprecision), and one provided a "personal score range" that incorporated the standard error of measurement (SEM) for the overall score.

Three of the four reports that presented subscores (75%) provided the number and percentage of items answered correctly within select content areas. One report used scaled scores to

present subdomain feedback, and that was the only report that included performance levels for subscores (in relation to college readiness) as well as information about subscore precision. Regarding visual displays, tables were used in three out of the four reports (75%) with subdomain results.

With respect to explanatory text on the reports, only one of the eight reports included a statement about the purpose of the test from which results were derived. Several reports (five, or 63%) provided guidance on desirable next steps, such as sending scores to academic institutions, requesting additional reports, and deciding whether to retake the test. One report stood out in that it provided information to support this decision by including five questions to determine if retesting would be beneficial, and a pie chart showing the percentage of test takers who saw an increase, a decrease, or no change in their composite score upon retesting.

Score Reporting Features in Certification Testing Programs

Almost all of the certification agencies sampled provided separate reports for passing and failing test takers (20 of 22 agencies, or 91%). As shown in Table 6.1, the two kinds of reports were substantially different. For example, "fail" score reports were more likely to include a numerical overall score (19, or 86%) than "pass" score reports (13, or 59%). Overall scores were almost always scaled scores and none of the reports included percentiles, reflecting the fact that the primary purpose of certification programs is to determine whether a candidate performs above or below a standard (i.e., criterion-referenced performance), and not to support norm-referenced comparisons among candidates with the same pass/fail designation. Performance levels were always included, either explicitly (e.g., "Result: PASS") or implicitly (e.g., "We regret to inform you . . ."). None of the certification reports included information about the precision of overall scores.

All 22 score reports for failing candidates included subdomain feedback compared to only eight reports (36%) for passing candidates. Reports for candidates who did not pass either provided qualitative subdomain feedback in relation to performance levels or subscores; only one report combined both approaches. Across the eight (36%) reports that used performance levels, descriptions for such levels were only included in three. Additionally, information about the precision of subdomain results was addressed in six reports (27%) for failing candidates and two reports for passing candidates (but only eight of the reports for passing candidates provided subdomain results). Visual displays were used by 19 reports (86%) for failing candidates and seven (32%) for passing candidates, with the most common displays being tables and horizontal bar graphs.

A substantial number of score reports provided guidance on next steps, for both passing (16, or 73%) and failing (19, or 86%) candidates. Reports typically presented information on maintaining the newly-earned certification and obtaining certification materials or applying to retake the test and using the score report to guide remediation, depending on the testing outcome. Among reports for candidates who did not pass, one unique feature was the use of sympathetic language—words acknowledging the disappointment associated with failing an exam and, sometimes, encouraging candidates to consider retesting. A much higher number of report for failing candidates (15, or 68%) included information about where to find additional resources than repots for candidates who passed (seven, or 32%), which is expected since those who did not pass have a greater need for additional information about the testing program.

Score Reporting Features in Licensure Testing Programs

Although differential reports were produced for passing and failing candidates for the majority of score reports from the certification tests we sampled, the same was not true for the score reports

from licensure programs. Among the eight licensure reports we collected, six included the same elements regardless of the testing outcome, and two were only for failing candidates (i.e., those two programs did not produce reports for passing candidates). Thus, we did not divide the licensure reports into "pass" and "fail" as we did for the reports from certification exams.

The majority of licensure score reports (six, or 75%) included numerical overall scores. The exceptions were the two reports made specifically for failing candidates, which focused on subdomain results. When present, overall scores were either scaled scores or percent-correct scores. All reports indicated candidates' performance level in terms of a pass or fail outcome, but explicit descriptions of what it means to perform in the "pass" or "fail" range were not included in any report. Only one report provided information about overall score precision, and only two (25%) used visual displays to present overall results (in both cases, tables were used).

Unlike reports from the two other contexts, six of the eight score reports for licensure programs presented subdomain results in relation to performance levels rather than numerically, as was done by only one program. Additionally, almost all reports that used performance levels (e.g., *Lower/Borderline/Higher Performance*) included descriptions of those levels (five out of six, or 83%). Only three included details about subscore precision. In terms of visual displays for subdomain results, five licensure reports used tables and one used a graphic with horizontal bands representing performance.

As was the case for admissions and credentialing, only two of the licensure reports included descriptions of the purpose of the test. Unlike the two previous contexts, however, only three of the reports (38%) provided guidance on next steps. In all three cases, the suggested next steps involved using subdomain feedback to devise a study strategy, with a warning that candidates would be best served by reviewing all content areas to some extent prior to retesting.

Interpretive Materials for Admissions and Credentialing Score Reports

Almost all examples of score reporting include interpretative material to help examinees and other stakeholders understand the content of the reports. The most common type of interpretive material associated with score reports is an interpretive guide, which traditionally is a static document mailed to stakeholders along with a score report. Among the reports we reviewed, most admissions reports (six, or 75%), licensure reports (six, also 75%), and certification reports for failing candidates (15, or 68%) included text about where to find additional information. Score report users were typically referred to a website about the testing program, a web page about understanding score reports, or the candidate handbook (for certification reports). Unfortunately, it was not possible to gather all interpretive materials for every score report reviewed.

With the growing popularity of online report delivery systems and the use of websites to disseminate test-related content, several new types of interpretive materials have been created. Ferrara and Lai (2016) conducted a review of documentation practices that included certification and licensure programs. They found that test takers received information supporting score interpretation and use not only through interpretive guides, but also through candidate bulletins and handbooks. They also found that candidate bulletins typically described the purpose of the test as well as test day instructions and information about interpreting score reports. This possibly explains why so few score reports for credentialing programs in our review included statements of test purpose. Additionally, it suggests that such programs may use candidate bulletins as primary avenues to provide interpretive information pertaining to score reports rather than interpretive guides.

In addition to "hard copy" guides for interpreting score reports, some testing programs are also using videos posted on their websites to help stakeholders understand the report. For

example, both the American Board of Internal Medicine (2015a) and the USMLE program (2017) use videos slightly under 5 minutes in which different parts of a score report are shown along with voice-over narration. A video explaining the *SAT score report* (College Board, 2016) uses a similar approach, but it has more dynamic effects and lasts under two minutes—perhaps reflecting SAT test takers' format preferences. Another approach is seen in a video published by the *ACT program* (2016), which combines score report screenshots and voice-over narration with scenes where a YouTube personality in the same age demographic as most ACT test takers appears in front of the camera to share information.

Some testing programs also provide other interpretive material via their websites. For example, the American Board of Internal Medicine (2015b), and the Graduate Management Admissions Council (2015), provide on-demand interpretive information through features such as hyperlinks embedded in static score reports and buttons (e.g., "more information") in interactive score report delivery platforms.

Research on Score Reporting

Developing successful score reports involves both art and science. The "art" refers to the creative design process that is important for effective communication. The "science" involves considering the various studies that have been done to investigate what information people can perceive and comprehend, as well as the different types of information desired by the consumers of test results. In the previous section, we described the features of current score reports in admissions, licensure, and certification testing. The content of these reports has been determined through research focusing on the information score report users desire. Based on this research, several models for guiding report development have been proposed. In this section, we review this research, which provides recommendations and promising methods for developing, enhancing, and evaluating score reports for credentialing and admissions testing programs. Our review of this literature is stratified by three testing contexts: testing in grades K-12, licensure testing, and certification testing.

Applicable Research From K-12 Contexts

Although outside the primary areas of our review—admissions, licensure, and certification testing programs—models for score report development derived from research in K-12 settings are relevant to these contexts. There are two prominent models for designing and evaluating score reports: the Zapata-Rivera (2011) model and the Hambleton and Zenisky (2013) model. Both were presented as part of work that focused on educational reporting but apply to other contexts. The models offer a series of steps to guide score report development, prioritizing thoughtful planning steps before any prototypes are created and encouraging an iterative approach—using information from later stages to revise and repeat earlier stages as needed. In Table 6.2, we provide a brief summary of each model. The first two columns in Table 6.2 list the steps for score report development involved in each model. The third column lists sample tasks that can be carried out to evaluate score report prototypes based on research by Clauser and Rick (2016) following the Hambleton and Zenisky model.

In essence, the Zapata-Rivera (2011) model proceeds as follows: identify the information needs of the intended audience for a score report (Phase 1); consider how those needs match the information provided by the assessment (Phase 2); design/revise score report prototypes (Phase 3); and gather internal and external feedback on the prototypes (Phase 4). In turn, the Hambleton and Zenisky (2013) model consists of the following steps: lay the groundwork for developing reports (Phase 1); design prototypes (Phase 2); gather feedback, making revisions, and

seeking additional feedback as necessary (Phase 3); and establish a process to evaluate whether implemented reports continue to be used and interpreted as intended (Phase 4).

Besides the models, multiple methods and general suggestions from research on K-12 score reporting have applications to reporting for credentialing and admissions programs. Some of these models and methods were summarized by van den Heuvel, Zenisky, and Davis-Becker (2014) and by Rick and Keller (2015). These reviews highlight the importance of recognizing that the most important information to include in a score report and the best way to present it vary widely depending on the intended audience(s), the purpose of the assessment, and the psychometric properties of the data from which scores are derived. Thus, incorporating general principles of good report design (e.g., Hullman, Rhodes, Rodriguez, & Shah, 2011; Jacoby, 1997; Tufte, 1983, 1990; Wainer, 1997) is important, but is no substitute for collecting direct feedback. For that reason, much of the research on score reporting has focused on methods for gathering feedback from stakeholders, and the types of feedback that are needed. In the next section, we review this research with respect to admissions and credentialing testing programs.

Score Reporting Research From Credentialing Contexts

Several published studies and conference presentations from admissions and credentialing contexts have described procedures for gathering feedback on operational or draft score reports. Jones and Desbiens (2009), for example, used a four-question survey to investigate how well 53 residency applicants could interpret their United States Medical Licensing Exam (USMLE) scores. The survey simply asked, "What was your score?" and "What percentile does this

Table 6.2 Summary of two score report development models and sample tasks.

Zapata-Rivera (2011) model	Hambleton and Zenisky (2013) model	Sample tasks from Clauser and Rick (2016)
Phase 1: Gather assessment information needs Phase 2: Reconcile those needs with available assessment information	Phase 1: a. Articulate score reporting considerations throughout test design decisions b. Identify intended audiences c. Complete needs assessment for each intended audience d. Review the literature and relevant documents	1a. Reviewed and expanded an existing internal document outlining intended inferences from score reports 1b. Identified target users in the process of expanding the inferences document 1c. Postponed until Phase 3 1d. Conducted a literature review
Phase 3: Design/revise score report prototypes	Phase 2: Create draft reports	2. Developed eight report prototypes. In collaboration with staff, selected three for the next phase.
Phase 4: Evaluate report prototypes internally and externally	Phase 3: Gather feedback on proposed reports (revise and repeat as necessary)	3. Conducted a focus group with medical students (target audience) to collect feedback on the prototypes. Then, made revisions and gathered input on the revised prototypes through cognitive interviews. Lastly, sent a survey to staff to elicit additional input.
	Phase 4: Once reports become operational, evaluate stakeholder feedback in terms of accessing, interpreting, and using the reports	4. Not reached (Phase 3 efforts are still in progress)

represent?" in relation to two exams in the USMLE series. At the time, examinees received both a three-digit and two-digit score, the latter resulting from the need to meet licensing authorities' requirement that the passing score would always be 75. They found that 30 (57%) of the residency applicants incorrectly perceived their two-digit score on the Part I exam as a percentile, and 31 (58%) applicants made the same wrong assumption for scores on the Part II exam. Thus, this quick and straightforward data collection approach provided evidence supporting the researchers' hypothesis that some examinees misunderstood one of the two scores provided on score reports (two-digit scores were discontinued in 2011).

In another effort related to USMLE score reports, Rick and Clauser (2016) conducted cognitive interviews with 12 medical students as they interacted with three report prototypes. There was interest in understanding what report features best supported adequate interpretations and remediation plans, so all prototypes displayed the performance of an examinee who did not pass. Participants received the prototypes in varying order and were asked to "think aloud" while considering three guiding questions: "What do I see? What does this mean to me? What can I do with this information?" (p. 6).

After analyzing the content of the interviews, Rick and Clauser (2016) concluded that when examinees received subdomain feedback both in relation to the national average and in relation to their own overall performance, they found it easier to interpret the former. There were 52 correct and two incorrect "compared to the national average" interpretations, while there were 43 correct and 11 incorrect "compared to your own overall performance" interpretations. Some students aptly combined both types of feedback, but others had trouble understanding that "overall performance" referred to their performance level across all subdomains. In terms of remediation, there were seven times as many "adequate" plans (50 statements) as "inadequate" plans (seven statements). Remediation plans were deemed adequate when students expressed that they would spend more time on their weakest areas without ignoring other areas in which their performance was also displayed as less than ideal. This type of plan is important because the exam is designed to be integrative. Rick and Clauser noted that sentences explaining that the exam is integrative and students should review all subdomains prior to retesting which were included in all score report prototypes—likely contributed to the high number of adequate remediation plans.

In the area of teacher certification, Klesch (2010) demonstrated the benefits of obtaining input from multiple stakeholders and adjusting data collection methods according to the level of detail needed at each point. First, she created three examinee score reports based on the K-12 literature. Then, she gathered feedback from 16 educators through individual meetings that included an interview and a questionnaire with preference- and comprehension-based questions. The meetings were conducted via an online video conferencing tool, which helped recruit a geographically diverse sample.

The interviews and questionnaires revealed several trends about how teachers interpret score reports. For example, several teachers suggested eliminating abbreviations and pure statistical terms; one teacher mentioned that even the use of "N" to represent "number" could be confusing. In addition, teachers had a strong preference for seeing raw scores when possible, and some even attempted to compute percent-correct scores from scaled scores to understand them better, which leads to a false result in most cases. After this stage of data collection, Klesch (2010) gathered feedback from six educational testing professionals through focus groups. Upon completion of the study, she offered a number of conclusions about teachers' preferences and information needs, including "confidence intervals were not immediately understood or seen as useful, while the performance of passing examinees provided an important contextual framework; [and] scaled scores need more explanation in how they are related to and derived from raw scores" (p. 138).

Score Report Research in Admissions Testing Programs

In addition to providing subscores, one approach for providing diagnostic information to end-users of test results involves using "item mapping" to enhance performance category descriptions (PCDs). Hambleton and Sireci (2008) led an effort to develop "clear, meaningful, and instructionally relevant" (p. 3) PCDs for the SAT using item mapping and assistance from content experts. Their process involved first calibrating previously administered SAT items from seven forms onto a common IRT scale. Then, equipercentile equating was used to find the IRT score intervals corresponding to six intervals on the SAT scale (e.g., 200 to 290, 300 to 390, and so on), which would be the focus of the PCDs. Next, content experts who were familiar with the SAT program were recruited to collaboratively develop PCDs using the item mapping information. Booklets were prepared in which experts could see which items test takers in a given interval were likely to answer correctly (based on at least a 65% probability of a correct answer), how response probabilities differed across intervals, and other relevant information. In a series of two to three meetings, panels drafted and finalized PCDs, which were reviewed by consultants and staff and sent back to the content experts for a last round of feedback. This process was conducted separately for mathematics, critical reading, and writing.

According to Hambleton and Sireci (2008), nearly every one of the 20 content experts who participated in the process provided comments during the final round of feedback, but all changes suggested were editorial in nature, reflecting widespread consensus over the substance of the PCDs. This result suggests that clearly communicating the goals of item mapping and PCD development, and providing experts with several avenues to offer input into the process, are helpful steps in developing effective PCDs. For assessments that are largely unidimensional, carefully developed PCDs can provide valuable diagnostic information to improve stakeholders' understanding of their performance and, if necessary, inform remediation plans.

Finally, Powers, Li, Suh, and Harris (2016) described efforts to improve ACT score reports through the addition of "reporting categories." Starting in late 2016, scores on reporting categories such as functions, algebra, and geometry replaced ACT subscores. According to Powers et al., the advantage of reporting categories is that they are more closely aligned with college and career readiness standards and are provided along with readiness benchmarks that help students better prepare themselves for college.

Research on Score Report Quality

In terms of enhancing score reports, there is a considerable body of research on the psychometric quality of subscores across a number of credentialing and admissions testing programs (Haladyna & Kramer, 2004; Lyren, 2009; Puhan, Sinharay, Haberman, & Larkin, 2008; Wedman & Lyrén, 2015). A full review of those studies is beyond the scope of this chapter, so we focus on recommendations for communicating—rather than computing or evaluating—subscores (readers seeking a more in-depth discussion of subscores are referred to Sinharay, Puhan, Haberman, & Hambleton, this volume).

One important study in this area was conducted by Luecht (2003), who compared four methods of computing subscores for credentialing tests and discussed points to consider when deciding to report them. To ensure that subscores are interpreted and used as intended, he emphasized the need to match numerical information with the appropriate display. Luecht provided several recommendations based on a review of long-established resources about creating good graphics, including: "show the data or legitimate patterns that represent the data," "avoid distortions," "encourage visual comparisons," and "make sure that the graphic(s) is/are closely integrated with statistical and verbal descriptions of results" (pp. 18–19). He concluded that

diagnostic feedback should help candidates understand their strengths and weaknesses in an unambiguous way, and how well subscores and related graphs are understood by their intended audience should be tested empirically rather than assumed.

One of the most prevalent challenges associated with reporting subscores is communicating their precision. Omitting information about measurement uncertainty may be misleading (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014), but presenting too many technical details may be equally problematic. Phelps, Zenisky, Hambleton, and Sireci (2012) offered several examples of how certification and licensure assessment programs report reliability and measurement uncertainty. Inspecting nine score report samples and ancillary documents from accounting, law, medicine, nursing, and teaching programs revealed a mix of approaches. Two organizations provided diagnostic feedback in the form of subscores with confidence bands, one organization included text about possible sources of measurement error, and two organization did not present information about precision, but used computerized-adaptive testing algorithms to ensure that pass/fail decisions were based on a pre-specified level of precision. Many organizations only provided information about score and subscore imprecision in published papers, invited presentations, and technical documents, some of which were not directly available to the public.

Considering score reports along with other supporting materials, Phelps et al. (2012) found that licensure programs tended to provide less information about precision than educational testing programs, and smaller programs provided fewer details than larger programs. Ferrara and Lai (2016) had similar observations and noted that larger programs are likely better able to provide information about precision and other technical aspects due to higher testing volumes as well as potentially more resources to support scoring procedures and report development. Ferrara and Lai also added that licensure programs tend to offer more technical information to test takers than certification programs, and that might be due to the usually higher stakes associated with obtaining a license versus a certificate.

Validity Issues in Score Reporting

In the world of testing, validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). This definition, from the *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*), makes it clear that validity does not refer to an inherent property of a test, but rather to how test scores are used and interpreted (see Tannenbaum, this volume, for other related definitions).

The interpretation of a test score begins with a person viewing a score report. Thus, the design and dissemination of score reports directly affect the degree to which a test has its intended effects. For this reason, the AERA et al. (2014) *Standards* mention the importance of properly reporting test results in several chapters.

In the chapter on "*Test administration, scoring, reporting and interpretation*," the *Standards* point out, "Reports and feedback should be designed to support valid interpretations and use, and minimize potential negative consequences" (AERA et al., p. 119). This recommendation sums up the guidance provided by the models for score report development (e.g., Hambleton & Zenisky, 2013; Zapata-Rivera, 2011), and seems to be adhered to by the admissions and credentialing score reports we reviewed.

The *Standards* also point out the importance of providing explanations of score reports to prevent misinterpretations. As they suggest, "Interpretive material should be provided that is readily understandable to those receiving the report" (AERA et al., 2014, p. 112). This standard

suggests that sufficient supporting material should be provided so that the score reports are easily comprehensible to those who receive them. In some cases, for example when admissions test scores are reported to parents, translations of the interpretive information may be necessary (e.g., College Board, 2017).

The *Standards* also provide suggestions for conducting research to help develop explanatory material to accompany score reports. As they put it,

> While test users are primarily responsible for avoiding misinterpretation and misuse, the interpretive materials prepared by the test developer or publisher may address common misuses or misinterpretations. To accomplish this, developers of reports and interpretive materials may conduct research to help verify that reports and materials can be interpreted as intended (e.g., focus groups with representative end-users of the reports.
>
> (p. 119)

Based on our review of the literature, many testing programs are also adhering to this guideline as the development of their score reports and interpretive information has been informed by research, much of which involved gathering perceptual and preference data from their stakeholders (e.g., Jones & Desbiens, 2009; Klesch, 2010; Rick & Clauser, 2016).

Other validity issues discussed in the AERA et al. (2014) *Standards* with respect to score reports are ensuring that score reports are corrected whenever errors are found, and ensuring privacy and confidentiality in score reporting. The *Standards* also point out that when composite scores are formed from different components of a test, it should be clear to end-users how the composite was developed. For example, they state, "If tests will be combined into a composite, candidates should be provided information about the relative weighting of the tests" (p. 182).

With respect to score reporting in credentialing testing, the *Standards* explicitly encourage reporting information to candidates who do not pass the exam, but they also point out that the psychometric property of any "diagnostic" scores should be established. For example, in the "Workplace and Credentialing" chapter, the AERA et al. (2014) *Standards* state,

> Candidates who fail may profit from information about the areas in which their performance was especially weak. This is the reason that subscores are sometimes provided. Subscores are often based on relatively small numbers of items and can be much less reliable than the total score. Moreover, differences in subscores may simply reflect measurement error. For these reasons, the decision to provide subscores to candidates should be made carefully, and information should be provided to facilitate proper interpretation.
>
> (p. 176)

Thus, validity issues in reporting scores for credentialing exams are not limited to the reporting of the pass/fail distinction. Like other score reports, the validity of all the information provided should be supported by both theory and evidence. For example, the theory that dictated the definition of the construct (e.g., unidimensional or multidimensional) should be consistent with the scores that are reported, and evidence that the reported scores have sufficient reliability for their intended purpose should be provided. Of course, the amount of evidence needed for a given interpretation is related to the stakes associated with the use of the test score. Thus, reliability expectations for the pass/fail score will be higher than reliability expectations for subscores reported for diagnostic purposes. Nevertheless, all reported scores need to have sufficient evidence that they provide useful information and are understood by end-users.

The AERA et al. (2014) *Standards* describe five sources of validity evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use"

(p. 13). A description of these five sources is beyond the scope of the present chapter and so readers are referred to the *Standards* and to other descriptions in the validity literature (e.g., Sireci & Gandara, 2016; Sireci & Soto, 2016; Tannenbaum, this volume). However, it should be noted all five sources of evidence are relevant to the evaluation of score reports.

Recently, O'Leary, Hattie, and Griffin (2017) argued that a sixth source of validity evidence should be added to the list—evidence of appropriate interpretability of test scores. Specifically, they argued that "alignment between intended interpretations and use of scores and actual interpretations and use of scores is critical" (p. 16). O'Leary et al. recommend that test validation should include evaluating the degree to which end-users of tests correctly interpret test results. They pointed out such research has been lacking in validity arguments developed for testing programs, and they lamented, "It is almost absurd to think that the intended interpretations and uses of test scores might fail because there is a lack of alignment with the actual interpretations made and uses enacted by the audience" (p. 16). We revisit this perspective in the final section of this chapter.

Lastly, it is important to note that in addition to AERA et al. (2014), the National Commission for Certifying Agencies (NCCA, 2014) and the International Test Commission (ITC, 2014) also provide score reporting guidelines that apply to admissions and/or credentialing testing programs. The three organizations hold similar views of the responsibilities of testing programs towards their intended audiences, and their recommendations complement rather than conflict with each other. Davis-Becker and Kelley (2015) provide an excellent summary of the key ideas found across guidelines from AERA et al., NCCA, and ITC, as well as suggestions for how credentialing programs can meet those guidelines.

Looking Forward: Future Research and Practices in Score Reporting

In this chapter, we discussed research and practices in score reporting for admissions and credentialing exams, and we contrasted these practices with professional standards for testing. In general, the score reporting practices we reviewed were consistent with the AERA et al. (2014) *Standards* and with other guidelines for best practices in this area (e.g., Hambleton & Zenisky, 2013; NCCA, 2014; Zapata-Rivera, 2011). However, tests in these areas are under increasing scrutiny and also are experiencing significant growth. Thus, we expect score reporting to receive more attention, and to become more interactive, including links to extensive interpretive material available online (e.g., American Board of Internal Medicine, 2015a; USMLE, 2017).

Perhaps the most important standard in the AERA et al. (2014) *Standards* is, "A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation" (p. 23). Given that score reports are the seeds from which test score interpretations grow, in the future, we would like to see research-based evidence to support the reliability and utility of *all* information included on a score report.

We also predict the perspective of O'Leary et al. (2017) is likely to gain traction, and research on score reporting will be more commonly conducted as part of developing a validity argument to support the use of a test for a particular purpose. They claimed that "Broadening validity evidence to incorporate a notion of evidence of interpretability could be achieved quite simply by including evidence of score report interpretability as one of the forms of validity evidence" (p. 20). We are not sure that a new category of validity evidence is needed, because research on test score interpretability could be couched within the source of validity evidence known as validity evidence based on testing consequences (AERA et al., 2014). Nevertheless, regardless of how such evidence is categorized, we agree it is essential, and we hope to see more of it in the near future.

In summary, score reports for admissions and credentialing programs appear to have stakeholders' needs in mind and strive to provide the information they need in a comprehensible format. However, reporting the results from these assessments is complex, and so test developers and testing agencies continue to improve their reports based on research to facilitate proper score interpretation, and minimize misinterpretations. The addition of URLs to score reports and accompanying interpretive material that indicate how to get additional information is an important trend that is likely to facilitate proper interpretation of test results. We hope future research in this area will confirm that hypothesis.

References

ACT. (2016, September 21). *2016–17 enhanced score report* [Video File]. Retrieved from www.youtube.com/watch?v=HPO7cGClQ1A

American Board of Internal Medicine. (2015b). *MOC examination score report*. Retrieved from www.abim.org/~/media/ABIM%20Public/Files/pdf/exam/score-report.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Board of Internal Medicine [InfoABIM]. (2015a, June 26). *ABIM exam score report update* [Video File]. Retrieved from www.youtube.com/watch?v=41tXCmn87Ik

Clauser, A. L., & Rick, F. (2016, April). *Designing and evaluating score reports for a medical licensing examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

College Board. (2016, May 6). *Understanding your SAT score report* [Video File]. Retrieved from www.youtube.com/watch?v=Aq07MvKOc1o

College Board. (2017). *Cómo interpretar los resultados del PSAT 10, 2017*. Retrieved from https://collegereadiness.collegeboard.org/pdf/understanding-psat-10-results-parent-tutorial-spanish-2017.pdf

Davis-Becker, S., & Kelley, J. (2015). *Score reporting: Where policy meets psychometrics*. Washington, DC: Institute for Credentialing Excellence.

Ferrara, S., & Lai, E. (2016). Documentation to support test score interpretation and use. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 603–623). New York, NY: Routledge.

Graduate Management Admission Council. (2015). *Enhanced score report demo*. Retrieved from www.mba.com/us/the-gmat-exam/gmat-exam-scores/your-score-report/enhanced-score-report-demo.aspx

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & the Health Professions*, *27*(4), 349–368.

Hambleton, R. K., & Sireci, S. (2008). *Development and validation of enhanced SAT score scales using item mapping and performance category descriptions* (Final Report). New York, NY: College Board.

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: Some new findings, research methods, and guidelines for score report design. In K. F. Geisinger (Ed.), *American Psychological Association handbook of testing and assessment in psychology* (pp. 479–494). Washington, DC: American Psychological Association.

Hullman, J., Rhodes, R., Rodriguez, F., & Shah, P. (2011). Research on graph comprehension and data interpretation: Implication for score reporting. In D. Zapata-Rivera & R. Zwick (Eds.), *Test score reporting: Perspectives from the ETS score reporting conference*. (Research Report 11-–45). Princeton, NJ: Educational Testing Service.

Institute for Credentialing Excellence. (2017). *NCCA accredited certification programs*. Retrieved from www.credentialingexcellence.org/nccadirectory

International Test Commission. (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*, 195–217.

Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data*. Thousand Oaks, CA: Sage Publications.

Jones, R., & Desbiens, N. (2009). Residency applicants misinterpret their United States medical licensing exam scores. *Advances in Health Sciences Education*, *14*(1), 5–10.

Klesch, H. S. (2010). *Score reporting in teacher certification testing: A review, design, and interview/focus group study*. Doctoral dissertation. Retrieved from ProQuest LLC

Luecht, R. M. (2003, April). *Applications of multidimensional diagnostic scoring for certification and licensure tests*. Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, IL.

Lyren, P. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research & Evaluation*, *14*(4), 3–12.

National Commission for Certifying Agencies. (2014). *Standards for the accreditation of certification programs*. Washington, DC: Institute for Credentialing Excellence.

Nguyen, T. C. Q. (2009). *Khoa cu Vietnam (Tap Thuong): Thi Huong*. Hanoi, Vietnam: Literature Publishing House.

O'Leary, T. M., Hattie, J. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues & Practice*, *36*(2), 16–23. doi:10.1111/emip.12141

Phelps, R. P., Zenisky, A., Hambleton, R. K., & Sireci, S. G. (2012). On the reporting of measurement uncertainty and reliability for U.S. educational and licensure tests. In D. Opposs & Q. He (Eds.), *Ofqual's reliability compendium* (pp. 605–641). Coventry: Office of Qualifications and Examinations Regulation.

Powers, S., Li, D., Suh, H., & Harris, D. J. (2016). *ACT Reporting category interpretation guide: Version 1.0* (ACT Working Paper 2016, 05). Iowa City, IA: ACT.

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2008). *Comparison of subscores based on classical test theory methods* (ETS Research Rep. ETS RR-08–54). Princeton, NJ: Educational Testing Service.

Rick, F., & Clauser, A. (2016, April). *What score report features promote accurate remediation? Insights from cognitive interviews*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Rick, F., & Keller, L. (2015, April). *Score report development and evaluation: Adapting strategies from K-12 to licensing contexts*. Paper presented at the annual meeting of the New England Educational Research Organization, Portsmouth, NH.

Sireci, S. G., & Gandara, M. F. (2016). Testing in educational and developmental settings. In F. Leong et al. (Eds.), *International test commission handbook of testing and assessment* (pp. 187–202). Oxford: Oxford University Press.

Sireci, S. G., & Soto, A. (2016). Validity and accountability: Test validation for 21st-century educational assessments. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 149–167). New York, NY: Routledge.

Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

United States Medical Licensing Examination. (2017). *Understanding your USMLE step 1, 2 CK and 3 score report*. Retrieved from www.usmle.org/transcripts/

van den Heuvel, J. R., Zenisky, A., & Davis-Becker, S. (2014, April). *Applying lessons learned in educational score reporting to credentialing*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Wainer, H. (1997). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics*, *22*, 1–30.

Wedman, J., & Lyrén, P. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research & Evaluation*, *20*(21), 1–14.

Zapata-Rivera, D. (2011). Designing and evaluating score reports for particular audiences. In D. Zapata-Rivera & R. Zwick (Eds.), *Test score reporting: Perspectives from the ETS score reporting conference* (Research Report 11–45). Princeton, NJ: Educational Testing Service.

7

Score Reports for Large-scale Testing Programs

Managing the Design Process

Sharon Slater, Samuel A. Livingston, and Marc Silver

People pay to take a test—or pay to create a test and require other people to take it—because they want or need the test results for some purpose. The information on the score report is the product they are paying for, not the test itself. The most carefully developed, research-tested procedures for assessment design, item development, and psychometric analysis will be wasted if the score report does not communicate the test results in a way that encourages proper interpretation and use.

This chapter is written for people whose responsibilities include the development of score reports for a large-scale testing program. Typically, in this design process, the people who design the report are not the ones who will make the final decision as to whether the design is approved. We refer to the person or group who will make that decision as the "client." In some cases, the client will be a person or group of staff members at the same organization as the designers of the report—usually a testing company or agency. In other cases, the client will be a member of another organization that has contracted with the testing company for development work that includes the design of the report. Our purpose in writing this chapter is to provide the reader with the benefits of our experience in the design process. We think this information will be useful to both score report designers and to clients, and ultimately to the consumers of score reports.

In the past decade, there have been several publications containing recommendations for score report design (Hambleton & Zenisky, 2013; Hullman, Rhodes, Rodriguez, & Shah, 2011; Tannenbaum, this volume; Zapata-Rivera, 2011; Zapata-Rivera & Katz, 2014; Zapata-Rivera, VanWinkle, & Zwick, 2012; Zenisky & Hambleton, 2016). However, these published recommendations do not always translate smoothly to practice. Score report designers find their options limited by the available funding, technology, and display space. When clients' wishes conflict with what the report designers recommend, the process of designing score reports becomes even more complicated. In the following pages, we discuss several factors to consider during the score report design process, including: communicating with the client and negotiating what score information will be included on the report, deciding what additional descriptive information to include on the report, and deciding how to present that information.

We guide readers step-by-step through a score report design process that we have found to be successful in K-12, higher education, and business settings. This design process relies heavily

on principles of graphic design and user experience, and incorporates recommendations based on the score reporting literature.

Form a Team of Experts

Before beginning the score report design process, it is important to have the right people in place to do the work. We recommend assembling a team of people whose purpose, as a group, is to design and evaluate score reports. The mission of the group should be to develop score reports that are technically accurate and easily understood by the intended user. The members of this team should have different kinds of expertise, to bring different perspectives to the process. Some kinds of expertise are required at one stage of the process; other kinds of expertise are required at other stages:

- Graphic designers know how to create visually pleasing arrangements of report elements.
- User experience practitioners understand how design and wording decisions can add to or detract from the usability of the report, and they can recommend improvements. The term user experience, often referred to as UX, seems to have many definitions, and often refers to computer systems or web design. However, in our context, user experience experts keep in mind all aspects of the end-user's perception of the score report when considering the design. These aspects include how users will interact with the report, how well the information in the report is communicated, how the look and feel of the report is perceived, and how easy the report is to use and understand.
- Cognitive science researchers know what research studies have shown about communicating information so the audience can easily understand the intended message. (See the chapter in this volume by Mary Hegarty on how findings from Cognitive Science and Information Visualization can inform score reporting.)
- Psychometricians understand the limitations of each type of score and can recommend scores that will be adequately supported by the data.
- Assessment developers can identify the abilities that the test measures and describe them in language that the intended audience is likely to understand.
- Information technology (IT) staff can determine what content and format are technically feasible for the production of the report (online or printed).
- Accessibility experts ensure that the score report can be easily and correctly interpreted by those with visual impairments and/or those using assistive technologies, such as screen readers.

Depending on the resources available, it may not be possible to assemble a design team with all the necessary types of expertise. Sometimes a team member can fill more than one of these roles; for example, the graphic designer can also be the user experience expert if he or she has the necessary knowledge, or the IT expert may be sufficiently well-versed in accessibility features to fill that role. In the case where professionals with the types of expertise listed above are not available within an organization, we recommend hiring consultants to provide the missing skills. With the various perspectives represented, it is important for team members to respect each other's expertise. They must each realize that what appears best from their point of view could be impractical for some reason they hadn't considered. For example, a text change recommended to improve the usability of a report may cause an unintended change to the interpretation of the scores, prompting an objection from the psychometrician or assessment developer. The team must work together to create solutions that are acceptable to all.

In this chapter, we describe the process that we use for designing score reports for an external client. The procedure is somewhat simpler if the client is part of the same organization as the people actually designing the report. In that case, the design team will know more about the decision makers and the factors that will influence their approval of the design. For the remainder of the chapter we will refer to three separate entities with involvement in the score report design process:

- The *design team* is the group described above;
- The *client* is the person or group that the organization paying for the test designates as responsible for the score report being designed. This person or group represents the audiences' interests and is often staff from a state department of education or a credentialing agency;
- The *program team* consists of staff from the testing agency, including the program manager who is the person responsible for communicating with the client and keeping the project on schedule. In the case where there is no external client, the program team is the client, as well.

Clear, ongoing communication among these three groups is essential. From beginning to end, it is important for the design team to work as directly as possible with both the program team and the client staff. Both the design team and the program team need to understand what information the client would like to include on the score report. That information will nearly always include some kind of overall test score. It may also include classification levels, subscores, graphs, photographs or illustrations, and written text intended to help score users to understand the results or to help test takers interpret their scores or improve their performance. Our experience has taught us that it is important for the design team (or at the very least, the graphic designer) to get feedback about the score report designs directly from the client. It is also wise to find out what individual will have to approve the final design and to involve that person in the design process as early as possible. If not, the design team may not correctly understand the decision maker's wishes, leading to wasted design effort, valuable time lost, and frustration for all involved.

Communication works a little differently with every testing program. One strategy that works particularly well is to have a one-day or two-day working session completely devoted to designing the report. The participants are the graphic designer, a psychometrician, and an IT staff member from the design team, one or two members from the program team, and key staff on the client side. At this score report design retreat, the participants work together, brainstorming, and sketching ideas, with the designer modifying the draft score report designs (mockups), in real time as changes are suggested. Having such a work session makes it possible to make significant progress quickly. However, travel costs and scheduling constraints often make this type of a meeting impractical. More typically, we hold regularly scheduled conference calls as part of the design process, as often as weekly. In one case, our main communication with the client was only indirectly through the program manager by way of one-line emails sent from a smartphone. The result was a report design that the client considered unacceptable, followed by substantial rework and increased costs. Clear and documented communication with the client is important in order to avoid such an outcome.

Score Report Design Process: Step by Step

Ideally, score report design begins at the very beginning of the test development process. An early step in the process should be the creation of a prospective score report showing what

information the test is intended to provide. The test can then be designed to provide that information. This is good advice, whether the test development process is following evidence-centered design principles (Tannenbaum, this volume; Zapata-Rivera et al., 2012; Zieky, 2014) or not. However, this ideal situation often is not what occurs in practice. More commonly, the test development process is well underway before the test designers begin thinking about the score report. Often, we find ourselves developing a score report for a test that is already fully developed, or nearly so. Regardless of when the process of designing score reports begins, the following step-by-step procedure can be applied. Figure 7.1 is a graphic representation of the score report design process, which is described in detail below.

Design Process: Step 1

Gather Information About the Test and the Scores to Be Reported

The process begins with the design team asking the program team for information about the test and how it will be used. This "kick-off" meeting is an opportunity for the program team to provide background and context about the client's needs before the design team and client meet directly. Our design team has developed a questionnaire to help guide this initial discussion. It begins with open-ended questions intended to help us learn about the test for which the score report will be designed:

- What kinds of knowledge or skills does the test measure?
- Who will be the main users of the score report?
- What decisions will be made with the information from this score report?
- Is there anything unusual about the testing program that may create difficulties for score report users?
- What are the critical deadlines for designing the score report? (State board meetings, IT programming schedules, score reporting deadlines?)

Once the design team understands these general score reporting needs, they ask more specific questions like those listed below. The design team should have as much of this information as possible for each type of score report to be designed. The program team may not have

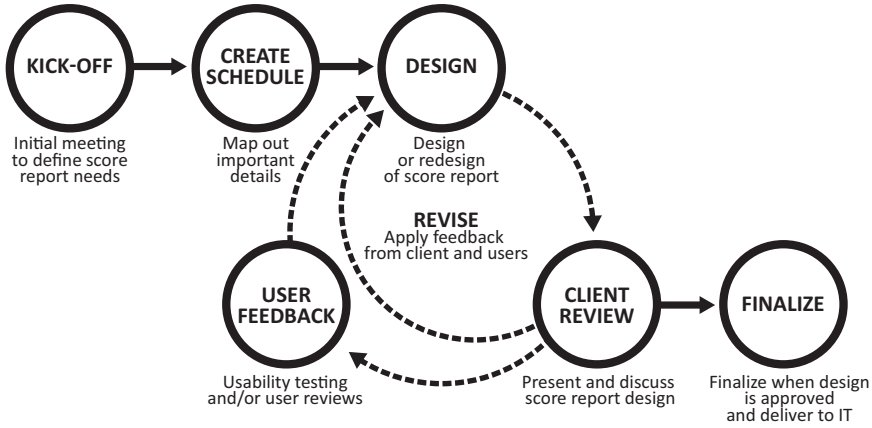


Figure 7.1 Iterative score report design process.

answers to all of these questions, but they can usually provide enough information about the client's wishes to get the design work started.

1. *Is this a new score report or a revision?*
2. *Is there an existing score report? If not, are there any sketches for suggested report formats?*
3. *Who are the test takers?*
4. *Who is the primary user of this score report?*
5. *Will there be other users of this score report?*
6. *What will the primary user want to know first? What is most important to the user of this score report?*
7. *What scores will this report include?*
8. *Will any reliability statistics (e.g., standard error of measurement) be included on this report? If so, what statistics, for which scores?*
9. *Will this score report include comparative data, such as percentiles, group averages, or previous year scores?*
10. *Will this score report include any other kinds of information (proficiency levels, growth, etc.)?*
11. *Will this score report be produced in-house or by an outside vendor?*
12. *Will this score report be delivered on paper? Electronically via email? Online?*
    - *If the report will be produced on paper, is the report limited to a number of pages/sides? If so, how many?*
    - *If the report will be online, what size screen is the reader assumed to have (e.g., desktop, tablet, smartphone)?*
    - *If the report is online, is conditional text required/desired (e.g., "If your score is between X and Y that means . . .") ?*
    - *If the report is online, will the reader have the option to choose which information is displayed?*
    - *If the report is online, will the reader be able to choose among different report formats?*
13. *Are there certain colors that must be used? Any specific branding or logos?*
14. *Is a sample score report required for a website, brochure, marketing materials or interpretive guide?*
15. *Are any needs assessments, usability studies, or focus group sessions planned?*
16. *Is there any other important information to consider in designing this score report?*

Using the questionnaire to focus the discussion, the design team typically can gather the information needed to begin the design process by talking with the program team for an hour or two. After that information-gathering session, the graphic designers will have the information they need to begin their work of creating preliminary mockups of the score report. For any questions the program team cannot answer, the design team can offer suggestions or make recommendations. They can offer to create different versions of the score report to show how various options would look (e.g., versions with and without percentiles or group averages, versions with and without standard errors shown graphically).

Design Process: Step 2

Create a Schedule for the Score Report Design

In Step 1 the design team gathers information about critical deadlines for the score report design. These dates will drive the creation of the score report design schedule in Step 2. A score report design that must be completed in two months will have a very different schedule from a

score report design that must be completed in six months. In scheduling, it is generally a good idea to start with the targeted end date for the design process and work backwards. The end of the design process is usually the date at which the IT staff needs to begin coding the systems for production of the score report (to be distributed electronically, printed on paper, or both). Although there does not seem to be such a thing as a "typical" score report project, Table 7.1 shows an example of a high-level schedule for the score report design process.

If time permits, we begin by scheduling two weeks or so for the designer to create preliminary concepts, followed by three rounds of mockups that will be shared with the program team and the client in an iterative process of review and revision. Additional iterations of the review-and-revision sequence can be added as needed. More complex reports usually require more iterations, but the time available will place a limit on the number of iterations possible. We have finalized designs in as few as three iterations, or "rounds," but some projects have involved over 20 rounds for a single report design. The number of rounds will depend somewhat on the complexity of the score report, but it will also depend heavily on the number of decision makers involved in the process and on how effectively the program team can manage the schedule and keep the work on track. The schedule in Table 7.1 allows about three months (counting business days only) for design of a score report; but we have seen projects with shorter timelines and projects that took much longer to complete.

Design Process: Step 3

Begin Creating Graphic Designs

One thing that can be helpful at the outset is to get a "napkin sketch"—a very rough drawing of the score report—from the client. Creating the napkin sketch forces the client to think about some of the goals and issues that the design team will face. The graphic designer then uses the napkin sketch, and/or the answers to the questions asked in the kick-off meeting to create several sample score report concepts, usually using a professional drawing program such as Adobe InDesign® or Adobe Illustrator®. These concepts are not mockups of the full report. They

Table 7.1 Sample schedule for design of one score report with three rounds of mockups.

Score Report Design Schedule Activities	Duration
Kick-off meeting to gather information about the score report	1 day
Designer creates multiple design concepts for the score report	10 days
Design team reviews concepts and provides feedback to designer	2 days
Designer creates Round 1 mockups	4 days
Design team reviews Round 1 mockups and provides feedback to designer	2 days
Designer revises Round 1 mockups based on design team comments	4 days
Program team reviews Round 1 mockups	5 days
Design team and program team meet to discuss feedback on Round 1	1 day
Designer revises based on feedback to create Round 2 mockups	4 days
Design team reviews Round 2 mockups and provides feedback to designer	2 days
Designer revises Round 2 mockups based on design team comments	4 days
Program team and the client review Round 2 mockups	5 days
Design team, program team and the client meet to discuss feedback on Round 2	1 day
Designer revises based on feedback to create Round 3 mockups	4 days
Program team and the client review Round 3 mockups	5 days
Design team, program team, and the client meet to discuss feedback on Round 3	1 day
Finalize score report design or continue additional rounds until complete	
Total days	55 days

are examples of the various options for each section of the score report. This is the time for the graphic designer to try out different options for types of graphs, for the size, placement, and arrangement of text and numerical information, for the use of color and icons, and for page composition. The choice of options to try will be based on the design team's understanding of client preferences and on previous experience with similar testing programs. At this stage in the process, the goal is to produce options for mockups that the client can then react to directly.

Existing reports can provide both good and bad examples of graphics, color, layout, and so on. In a K-12 assessment report to parents, it is important to use familiar graphics, attention-getting color, and brief, simple text. For score reports that will be used by institutions, it may be acceptable to use more complicated language and include more statistical information. If possible, the choice of language used and technical information included should be informed by research that has determined specific audience needs, pre-existing knowledge about, and attitudes toward assessments (e.g., Kannan, Zapata-Rivera, & Leibowitz, in press; Underwood, Zapata-Rivera, & VanWinkle, 2007; Zapata-Rivera et al., 2012). However, research-based guidance in designing score reports for various situations is not always available.

To help the design team keep the needs of the intended audience in mind, sometimes the program team or the client will provide market research. Such studies can describe what a particular audience wants or needs to know from the score report. At other times, the design team must rely on the client's judgments or intuitions about what the users of the score report will want to know. Throughout the design process, our design team uses the principles listed below, which are consistent with guidelines in the score reporting literature (Hattie, 2009; Kannan et al., in press; Underwood et al., 2007; Zapata-Rivera, 2011; Zapata-Rivera et al., 2012; Zenisky & Hambleton, 2016). In addition to guiding design efforts for new score reports, these principles can be applied to evaluate existing score reports. See Figures 7.2a and 7.2b for examples of two ways the same score report information can be displayed. Figure 7.2b was designed following the principles below.

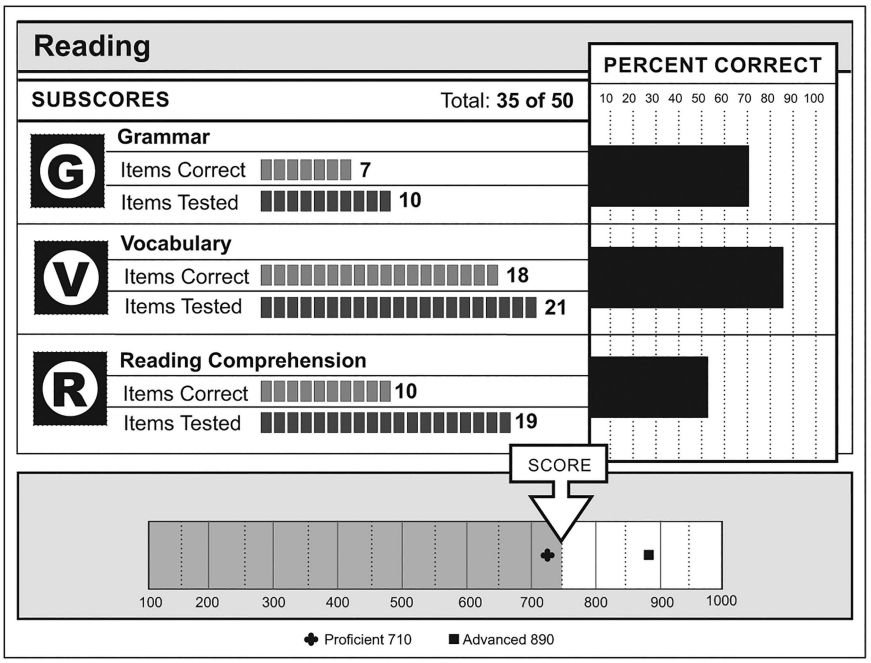


Figure 7.2a Example of a score report that does not follow the design principles listed in Step 3.

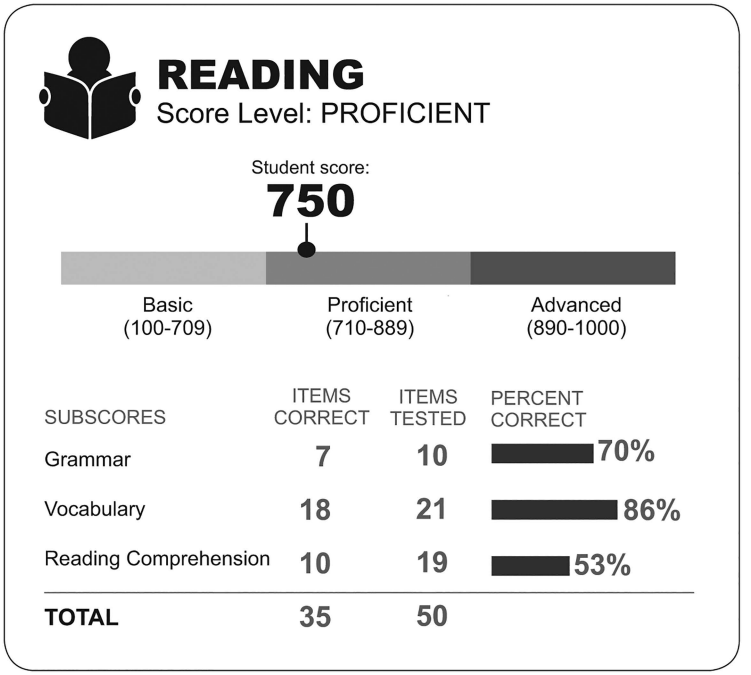


Figure 7.2b Example of a score report based on the design principles listed in Step 3.

Principles to guide the score report process include:

- Emphasize the most important information in the score report. Too many score reports emphasize unimportant features (logos, illustrations, etc.). The most important parts of the report should command the most attention.
- Design the report so that viewers can see and understand the most important information in 10 seconds or less. What is the first question that someone looking at the score report will want the answer to? It may be "What was my score?" or "Did I pass?" or "How did I perform in comparison to other people taking the test?"
- Create a strong visual hierarchy that guides the viewer's eye appropriately through the report. Items of information that viewers will need to compare should be close together in the report.
- Eliminate visual clutter—anything printed on the report that does not convey useful information. Avoid repeating elements of the report, except where repeating them makes the report clearer and easier to read. Each element on the score report should earn its space.
- Avoid using lines that add to visual complexity. Instead, use shaded areas to delineate space without adding clutter.
- Use visual embellishments such as icons only when they make it easier for users to correctly interpret the report. Be especially careful to avoid any visual elements that may have a negative connotation to some viewers of the report. (For example, a thumbs-up symbol is offensive in some parts of the world.)
- Use colors to help convey information. Colors, used meaningfully, can make the report easier to interpret. However, make sure that the report can convey all its information even

if the viewer cannot accurately differentiate colors or if the report is printed or photocopied in black and white.
- Follow Web Content Accessibility Guidelines (WCAG; Caldwell, Cooper, Reid, & Vanderheiden, 2008) for accessibility. Make sure there is adequate contrast between background and text, and that viewers using assistive technologies can easily read and interpret the report.
- If the score report will be printed, take into account the accuracy of the software and printers that will be used for production. For example, if the placement of printed elements (e.g., text, symbols) can vary by 1 millimeter, the report should not include elements that require more precision than that.
- Make sure that the report design can accommodate unusual but possible conditions, such as very long names or very low or high scores. (We sometimes design the sample score reports for a student named "Verylongfirstname Extremelylonglastname" to make sure to leave enough room.)
- Make sure the report can be economically reproduced. Avoid designs that will require large amounts of paper, toner, or ink, especially if schools or families will be printing the reports. Reduce the number of pages where possible.
- Ensure that any language in the report is appropriate to its intended audience, at the proper reading level. Avoid technical language that might be difficult for non-experts to understand, particularly if the score users may have limited English language skills.

At this stage, it is important to make sure that IT reviews the designs prior to showing them to clients. This step will ensure that the client does not fall in love with something that is not technically feasible.

Design Process: Step 4

Get the Client's Reactions to the Initial Designs

Once score report mockups have been created and the program team has had a chance to provide feedback, we share the mockups with the client to get feedback. Typically we present three different versions of each report. Often the client will like some features of one version and other features of another version. Sometimes after seeing the score report elements on the page for the first time, the client may have new ideas for ways to present a certain piece of information. We try to find a way for the designer to hear feedback from the client first-hand, with the opportunity to ask questions and get a better understanding of what changes the client wants, and why. A few minutes of dialog between the designer and the client can save hours of unnecessary work for the designer and days in the production schedule.

Design Process: Step 5

Gather Feedback from Intended Users of the Score Report

Once the client and the program team are happy with the score report mockups—which can take a number of rounds of design—it is wise to gather input from the people who will actually use the score report. In our experience, clients often prefer to gather score user feedback using actual mockups that have been designed and reviewed by them first (in Steps 1–4). It is unusual for a testing program to have the resources to gather feedback from score report users more than once. Typically, score reports are designed based on the client's understanding of what their score users will want and need, and those mockups are shown to score users in a focus

group setting or usability study in which their feedback can be collected and applied to later revisions to the score report design.

In the user experience world, this type of audience review is recommended for very early in the design process, which is also consistent with the score reporting literature. Zapata-Rivera and Katz (2014) suggest performing an "audience analysis" prior to designing score reports. Zenisky and Hambleton (2016) also recommend conducting a needs assessment at the beginning of the score report design process and to evaluate stakeholders' interpretation and use of score reports at every step of an iterative multistep report development process. While we agree with these recommendations and would like to see early and frequent input from score report users, this is not what we typically see in practice.

Whether early or later in the design process, the best way to know what the intended users want and need in a score report is to ask them directly. In this step, you can learn what score report recipients want to know and how well they understand the information on your score report. It is important to find out what they like and dislike, and even more important to find out whether they are interpreting the results correctly. Questions like, "What was the student's overall score?" or "What does the percentile mean?" give us a good idea about whether the score report design is clearly communicating the information and whether any parts of the score report are confusing to users. Comprehension questions like these are uncomfortable to ask and can make users feel as if they are being tested on something they have not studied. Before we ask any comprehension questions, we emphasize that our purpose is to find out if the report is clear and understandable. We tell the users that if they have trouble answering any of these questions, we will know that we need to revise the report. We are seeing more and more studies that are including comprehension questions in their assessment of interpretability of score reports (Hambleton & Slater, 1997; Kannan, Bryant, Zapata-Rivera, & Peters, 2017; Kannan, Zapata-Rivera, & Leibowitz, in press; McJunkin & Slater, 2017; Rick et al., 2016).

Unfortunately this step of the process—feedback from users of the report—is sometimes omitted due to tight reporting schedules, cost limitations, or both. When there is time and money for this important step, there are a few different ways to gather information from people like those who will use the score report. Options for gathering feedback from score report users range from broad-based surveys to focus groups, to one-on-one usability studies or cognitive labs with end users.

Surveys are generally the most cost-effective way to gather information on what users want in the score report. Even if the response rate on a survey is not very high, those who do respond will tend to be those with strong opinions about the topic, and their responses can provide useful feedback. If the survey includes open-ended questions, the schedule will have to allow time for coding the responses. Overall, surveys typically do not provide much rich data (apart from open-ended questions) or offer the opportunity to ask follow-up questions of the user. However, they are convenient to use and can be an efficient way to collect feedback.

Small focus group sessions enable the use of more open-ended questions. These sessions provide an opportunity for discussions among the group members, which can be observed. They let the facilitator and observers see the users interact with the score report and witness their reactions to various sections of the score report. However, focus group sessions must have a clearly structured protocol to be effective. They require more time than surveys, as staff will need to conduct the sessions, consolidate the findings, and prepare the report. The number of people (and geographical locations from where you can recruit) will also be limited. The biggest drawback, however, may be the influence of group dynamics on the results. There is always the possibility of a single focus group member imposing an opinion on the others. As a result, the members of that focus group may all say they had the same reactions to the report, when in fact they did not. Or if the vocal focus group member's personality is objectionable, the rest of the

group may be inclined to express disagreement with that person's opinion, regardless of their own underlying opinions.

One way to get many of the benefits of focus groups, but without the group dynamics, is to hold individual interview sessions with users, one-on-one. This format is the traditional one used for usability testing (Silver, 2005), and cognitive labs. A usability consultant or facilitator can introduce tasks and ask questions while observers on the design, program, and/or client teams view the sessions in person or remotely, with video and audio recordings made to show the participant's facial expressions (to show reactions such as confusion), and the screen or paper prototype that is being tested. This strategy is more time-intensive than focus groups, but it provides the opportunity for the participants to think out loud and express their initial thoughts, interpretations, and confusions. It enables them to provide information such as which parts of the report they don't understand, which version of the report they prefer, and what additional information they would like to see in the report—without being influenced by others. The usability test results are often written up in a report that summarizes issues and often makes recommendations for design improvements. This format provides actionable data that the design team can use to improve the score report design. For example, if you see the same misconception or point of confusion in several sessions, then you know you have observed something important. Dickey, Rick, Sireci, and Zenisky (2015) provide a review and bibliography that offers guidance on how to develop protocols for gathering this type of feedback for score report design.

Whatever strategy is employed for gathering user feedback, it is a good idea to test for accessibility by trying out the score report on score users with visual disabilities. In addition, it is important to keep a record of score user feedback in writing. That written feedback could be in the form of notes taken during a focus group session, a transcription of a recording from a usability study, or ideally, a more formal report summarizing all score user feedback.

Design Process: Step 6

Finalize the Design

At this stage in the process, the graphic designer applies the feedback gathered from Steps 4 and 5 to finalize the score report design. If many of the participants express similar views about the proposed score report, the necessary changes will be obvious. But if various groups react differently, someone—usually the client—may have to decide which changes the design team should implement. It can take multiple rounds of revision to get to the point where the client is ready to approve the design. This iterative nature of the score report design process is illustrated as the loop in Figure 7.1.

Once all of the decisions about the score report elements and the accompanying text have been made, the design process ends, and the final score report mockup can be handed over to the IT group which will begin coding the systems that will produce the report. During the production phase, much work remains to properly translate the mockup into the version of the report that will be provided to score users. Having IT staff involved throughout the design process is important, to make sure the proposed designs are feasible from a production standpoint and to make sure the IT staff understands what the client and score users want from the final report. The IT staff can also keep everyone involved, including the client, informed about the time needed for coding the production version of the report. To minimize the risk of missing reporting deadlines, it is important to give the IT staff the time they need to program the production version of the report. For operational score reporting to work well, there are many special circumstances that their coding must anticipate. For example, what message should appear

on the score report if a student misses only the mathematics portion of the test? What message, if any, should appear if a student received accommodations like extra time or additional help with language? The design schedule must accommodate the time needed for IT staff to evaluate possible circumstances and to build solutions to respond when the circumstances arise. The design team must stay engaged during this step to ensure that the production version of the score report continues to meet the requirements that were identified during the design process.

Interactive Score Reporting

To this point, the chapter has focused on the design process for a single static—or fixed—score report. Here the term "static" means that the same information will be presented on each score report in a noninteractive way. This is the type of paper score report that is mailed to a score users' home or may be available online as a PDF file or a link sent via email. Interactive score reports or score reporting systems, on the other hand, are always presented electronically (via computer, tablet, or smartphone). An interactive report includes a set of displays with hyperlinks and interactive menus that enable the user to select the information to be presented and to move from one type of view to another.

Demand for interactive reports has increased in recent years, just as personal-use, mobile technology (tablets, smartphones, etc.) has become more prevalent in our daily living. Some score users want access to score reporting information immediately and some may require the ability to tailor reporting to answer specific questions. Designing an interactive score report or an entire score reporting system is much more complicated than designing a single static score report. Score reporting systems house a collection of interactive reports that allow users to interact with the score report information. These systems encourage users to explore the data by clicking on aspects of a report to drill down for more information, sorting the data to show the score information in a particular order, or changing the way in which scores are displayed (tabular vs. graphical). This type of functionality requires sufficient time to design each individual display within the reporting system. You will need to understand and test how users will interact with the information in each report display and build intuitive ways for users to navigate between the various displays.

Most, but not all, of the principles for designing a static score report apply to each display included in an interactive score report. The guidelines described above for score report design should be applied to the design of each display in an electronic score reporting system. In a sense, one can think of each screen of the reporting system as a static score report, because the information on each screen is intended to communicate some aspect of the testing results to the score user. However, designing an interactive report involves the added complexity of determining how a user will navigate through such a system. When working through design of the navigation, there are a number of questions that must be considered:

- What will the user want to see first?
- How often and in what ways will the user want to interact with the system?
- How can the user drill down to more specific or detailed information?
- How might the user want to sort the information on the screen?
- Where might the user want to gain access to other resources?
- How can the user easily navigate back to a previous screen or to another part of the system?

This type of design work is where the user experience staff and cognitive science researchers on the team have a good deal of input. It is here that usability testing and evaluations of the comprehension of the information presented are a critical piece of the design process.

Lessons Learned

In addition to following the steps outlined above, there are a number of other lessons we have learned about the score report design process. In the section below, we share some of these important lessons we have learned as we have engaged in the design of score reports for large-scale assessments. Some of these constraints and pitfalls have been mentioned above, but they bear repeating due to their impact on the design process.

What Constraints Are Commonly Faced?

1. **Cost**. This is usually a constraint for any testing program. More pages, color, and customization included in a score report usually results in increased cost. Most states still require paper student score reports to be mailed home to parents. When hundreds of thousands or even millions of score reports need to be printed and mailed, we are often limited to the front and back of an 8 ½ x 11-inch piece of paper and at the most two or three colors in addition to black.
2. **Availability of data**. Sometimes the information included in the report is limited by the availability of the data. For example, a client may want to include comparisons to average scores for all test takers, but the scores for some test takers may need to be reported before others have taken the test. In this situation, one possible solution is to use the data from previous years. This solution works well when there is not much year-to-year change in the performance of the group of all test takers. However, there often is substantial year-to-year change, especially on a new or revised test.
3. **Tight schedule**. Another limitation we often need to work around is the schedule for design. Sometimes we need to work very quickly to create designs in time to meet client deadlines, and this time pressure often results in a decision not to gather feedback from score report users. In reality, there simply may not be enough time to conduct needs assessment and usability testing with the intended audience for the score report or score reporting system.
4. **Display space**. Not enough space is another limitation. It may be necessary to print the entire report on two sides of an 8 ½ x 11-inch piece of paper. There are often multiple subjects and scores to report, with text descriptions needed to explain the scores or additional information requested by the client. Sometimes score report designers are required to include so much information that the only solution is to put the text in very small print, making it unlikely that the score report recipients will read any of it. The font size in a score report should be no smaller than a 9- or 10-point font for the main areas of the report, and no smaller than a 7- or 8-point font for minor details or footnotes.

What Are Some Pitfalls to Avoid?

Vocal, Inexperienced "Designers"

Because a good score report is easy to read and understand, people may believe that it is easy to design a score report relatively quickly and that anyone can do it. Sometimes vocal individuals without experience in score report design (either on the client side or on the program team) insist on using a certain graphic or explanation that they like, instead of one that will typically be better understood by the end-user. In cases like this, it is quite helpful to be able to point to results from market research or a needs assessment with the actual users for the testing program to support selection of a particular design treatment.

Not Involving the Right Decision Makers Early in the Design Process

This pitfall often results in going back to the drawing board at a late stage in the design process. These decision makers are often busy people, which makes it difficult to get time with them to provide input on the score report. However, it is important to make sure they agree with at least the general designs before too much time is invested, and to get their approval in writing if possible. We had one situation in which a client rejected a design early on in the design process, only to come back to us later suggesting that we try a design similar to the one we originally presented. It turned out that the key decision maker for the client had not been included in the early decision to reject the initial design. In the interim, the design team produced several more versions of the report, getting the client's reactions to each version. In that situation, hundreds of hours of work and weeks or months can be wasted.

Clients Insisting on Including Too Much Text

Clients are passionate about their assessments and want to provide as much information as possible into the limited space available. It seems that when there is white space on a page, some have an irresistible urge to fill that space with words. This is contradictory to what score report users may want or need. In fact, in one of our focus groups, parents specifically stated that they preferred fewer words, bullets rather than long sentences, and more white space (Rick et al., 2016), yet we are often asked to add more descriptive text wherever space allows. One way to potentially avoid this problem is to present the client with two versions of the score report, one that you recommend for better readability and one that includes all of the text that they want to include. It would be better yet to present both versions to test score users and get feedback from them about which version is easier for them to read and interpret.

Educational Jargon on Score Reports

Score recipients want to know their score, whether they "passed," how they compare to their peers, and what they need to do to improve. They don't want to be confused by psychometric terms or other educational jargon. They may not care about the alignment of scores to standards. They usually don't know what it means to say that an assessment is "vertically scaled." They typically don't have the concept of a "standard error of measurement" or a "confidence interval," and if you tell them that their scores contain "error," they will want to know why you couldn't get it right. If the report includes any of these concepts, it is important to explain them in layman's terms, and those explanations can be difficult to create.

Clients Insisting on Including Something in the Report That is Against Your Professional Judgment

If your design team disagrees with the client about an aspect of the score report, it is wise to document the team's misgivings and recommendations in writing. Supporting your concerns with citations from the literature may be helpful (if they exist), but clients may be more swayed by recommendations based on regulations, court decisions, or examples from similar reports that were favorably viewed by the public. In one case, a client wanted to use a color that did not pass color contrast standards for visual accessibility. In another instance, a client wanted to simplify an explanation to the point that it was no longer entirely accurate. Share your concerns with the client in writing, so the client will be fully aware of the issues involved. If you can substantiate your concerns with results from research evaluating score report user understanding,

also include those in the letter to the client. Then let the client decide how to proceed. In the end, the report belongs to the client. The design team should never include misinformation on a report, but no client has ever asked us to do so. What is more common is for a client to request that the report include information that the design team does not think is useful. In general, it is wise to document your recommendations to the client. If a problem occurs because the client decided not to follow your recommendations, it could be very useful to have a record of those recommendations. It could be even more useful if your recommendations to the client were accompanied by an explanation in which you anticipated the problem that actually occurred.

The Role of Score Report Design Research

Research to determine how score report users comprehend and interpret test results could potentially help testing organizations to produce better score reports. However, much of the research done so far is based on small samples. Even when sample sizes are adequate, the results often depend heavily on the population the participants represent. What works well with one population of score recipients may not work well with another. In practice, score report design is most often based on visual design and user experience principles, driven by client preferences, with little research available to guide specific decisions about the design. However, there has been a noticeable increase in research on the topic of score reporting in recent years, and the results may help practitioners to improve score report design in the future. The chapters in this volume describe much of this research.

Conclusion

In the years that we have been working as a team to design score reports, we have learned a few things, and we expect to learn more each time we go through the process. In starting out, we focused our attention almost entirely on the graphics and the text included on the report. While these are clearly critical, we now realize that documented communication among all involved and strict adherence to the schedule for the work are just as important. Getting feedback from score report users is a design step that we are seeing more often, but unfortunately it still is often overlooked in score report design budgets. We will continue to emphasize to our clients the importance of obtaining user feedback. Hopefully over time, gathering feedback from score report users will become more the rule than the exception. In this chapter, we have offered our best advice about the score report design process, based on what we know now. In the years to come, we hope to learn how to improve the process further.

References

Caldwell, B., Cooper, M., Reid, L. G., & Vanderheiden, G. (2008, December). *Web content accessibility guidelines* (WCAG) 2.0. Retrieved from www.w3.org/TR/WCAG21/

Dickey, J. A., Rick, F., Sireci, S. G., & Zenisky, A. L. (2015). *Developing user needs assessment protocols for score report design: Literature review and annotated bibliography* (Center for Educational Assessment Research Report No. 906). Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report no. 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research based approach to score report design. In K. F. Geisinger (Ed.), *Handbook of testing and assessment in psychology* (Vol. 3, pp. 479–494). Washington, DC: American Psychological Association.

Hattie, J. (2009). Visibly learning from reports: The validity of score reports. *Online Eduational Research Journal*. Retrieved from www.oerj.org/view?action=viewPDF&paper=6

Hullman, J., Rhodes, R., Rodriguez, F., & Shah, P. (2011, December). *Research on graph comprehension and data interpretation: Implications for score reporting*. (Research Report No. RR—11–45). Princeton, NJ: Educational Testing Service.

Kannan, P., Zapata-Rivera, D., & Leibowitz, E. A. (in press). The interpretation and use of score reports by diverse subgroups of parents. *Educational Assessment*.

Kannan, P., Bryant, A. D., Zapata-Rivera, D., & Peters, S. (2017, April). *Evaluating parent comprehension of measurement error presented in score reports*. Paper presented at the meeting of the National Council on Measurement in Education, San Antonio, TX.

McJunkin, L., & Slater, S. (2017, April). *Educator interpretation and application of assessment results in a dynamic reporting system*. Paper presented at the meeting of the National Council on Measurement in Education, San Antonio, TX.

Rick, F., Slater, S., Kannan, P., Sireci, S., Zenisky, A., & Dickey, J. (2016). *Parents' perspectives on summative test score reports* (Center for Educational Assessment Research Report No. 937). Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Silver, M. A. (2005). *Exploring interface design*. Clifton Park, NY: Thompson/Cengage Learning.

Underwood, J. S., Zapata-Rivera, D., & VanWinkle, W. (2007). *Growing pains: Teachers using and learning to use IDMS* (Research Memorandum 08–07). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D. (2011). Designing and evaluating score reports for particular audiences. In D. Zapata-Rivera & R. Zwick (Eds.), *Test score reporting: Perspectives from the ETS score reporting conference* (pp. 32–62). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, *21*, 442–463.

Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL™ teachers* (Research Memorandum 12–20). Princeton, NJ: Educational Testing Service.

Zenisky, A. L., & Hambleton, R. K. (2016). A model and good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.) *Handbook of test development* (2nd ed., pp. 585–602). New York, NY: Routledge.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicologia Educativa*, *20*, 79–87.

8

Effective Reporting for Formative Assessment

The asTTle Case Example

Gavin T. L. Brown, Timothy M. O'Leary, and John A. C. Hattie

Assessment should have a purpose. As Zumbo (2009) stated, in the context of discussing validity, 'it is rare that that anyone measures for the sheer delight' (p. 66) going on to concede that measurement is 'something you do so that you can use the outcomes' (p. 66). Within educational contexts, there are many ways testing might be expected to used and improve schooling (Haertel, 2013), as well as many ways users might anticipate using test results (Hopster-den Otter, Wools, Eggen, & Veldkamp, 2016). One key use, perhaps the primary use, of educational assessment is the support of student learning (Popham, 2000). Given such improvement purposes for tests, validity requires that reports on student performance be well aligned to the test (and the test well aligned to the intended curricular goals) and well designed to ensure understanding (Tannenbaum, this volume).

In any system that expects teachers to monitor and respond to student learning, teachers are important users of test information. In such systems, the teacher's role is primarily to mediate test score information into appropriate instructional decisions (e.g., pace of progress, student grouping, task and activity design, selection of curricular resources, etc.). The focus of this chapter is on the communication of test results to teachers in ways that foster interpretations and actions that align with those intended. Shepard (2001, 2006) makes it clear that most educational assessment is carried out in classrooms by teachers and that significant improvements are needed in how testing might continue to play a part in that process. Teachers are expected to make a series of qualitative interpretations about observed student performances, as well as interpretations of test scores (Kane, 2006). These interpretations occur as teachers interact with students in the classroom and are not simply recorded for later interpretation. While modern directions in assessment design focus on ensuring that a robust theory of learning or cognition is present (Pellegrino, Chudowsky, Glaser, & National Research Council, 2001), it seems more appropriate in evaluating test reports for teachers to focus on theories of effective communication and instructional action.

Within educational settings, the first goal of a diagnostic test score report should be to ensure that the test reports inform teachers' decision-making about '*who needs to be taught what next*' (Brown & Hattie, 2012). Extensive research on feedback (Hattie & Timperley, 2007) shows that

in order to close the gap between where students are and intended curriculum goals and standards, tests have to describe diagnostically the current status (strengths and weaknesses) of a student and point to action that the teacher and/or the student can take to improve learning so as to maximise the probability of attaining the success criteria of the lessons. It is to reduce this gap between where they are and where we want them to be that leads to the importance of assessment. This means that effective educational tests have to provide more than total score or rank order information. In order to make instructional decisions about curriculum, reports need to specify, among other things, how scores can be used (AERA, APA, & NCME, 2014), though relatively little is contained in the *Standards* about ensuring that report readers make appropriate interpretations. Test developers seldom provide validation evidence as to what report readers see in the reports and what they do with the information (Hambleton & Zenisky, 2013; Hattie, 2014; Hattie & Brown, 2010). Yet, it is these two issues which will determine if test reports contribute to improved outcomes.

As a consequence, the second goal of such a test report is, or should be, to improve the quality of teacher instruction and student learning (Popham, 2000). This agenda has been made increasingly explicit with greater policy and research emphasis on a variety of approaches to assessment including: formative evaluation (Bloom, Hastings, & Madaus, 1971), school-based assessment (Torrance, 1986), classroom assessment (Crooks, 1988), performance assessment (Darling-Hammond, 1994), alternative assessment (Birenbaum, 1996), assessment for learning (Black & Wiliam, 1998), and assessment for teaching (Griffin, 2014). What these approaches have in common is that they situate the design, administration, scoring and interpretation of evaluative processes in the midst of the instructional environment, rather than external to it. This improvement-oriented process has to take place early enough so as to make a difference to outcomes (Scriven, 1991) and is methodologically catholic in that it does not privilege or denigrate tests versus other methods (e.g., performances, portfolios, peer or self-assessment, etc.). These approaches all focus on generating data and decision-making about learning outcomes, much in the manner of total quality management (Deming, 1986), by the people closest to and directly responsible for educational practices and processes (i.e., teachers and school leaders).

Parallel to this is the need for reports on test data to reach the teacher soon after the test has been administered so that the information is relevant to where the learning was when it was tested. There can be no doubt that a report that arrives from a central test agency some three months or so after the test date is unlikely to be valid or effective. As we have argued before:

> the potential for that information to actually shape meaningful learning activities is practically nil—the students have changed class or grade, the teachers have moved on to new material, the class may have been successfully taught that content, and so on.
>
> (Hattie & Brown, 2008, p. 195)

Prompt feedback to the teacher as to which children have which needs or strengths is a *sine qua non* in ensuring that standardised tests serve educational rather than administrative or policy goals. Indeed, another feature of rapid reporting to teachers is the assurance it gives that they are the first to read the reports; delayed reporting may have been monitored and inspected by superiors before it arrives, more so in jurisdictions that prioritise testing for school accountability. Rapid reporting allows teachers early access to both pleasing and disturbing data and the chance to respond to it before external stakeholders inspect the results (Brown & Hattie, 2012; Hattie & Brown, 2008). Hence, rapid reports to teachers connects the test information to their current teaching context and raises the probability that teachers will actively respond to the data.

An important policy consideration that will support accurate teacher interpretations and decisions from test reports has to do with consequences or stakes associated with the test.

Good tests can lead to educators discovering some very discomforting news (e.g., the class or school is well below expectations and averages). In an environment where there are negative consequences (e.g., league tables), there can be strong incentives to game or cheat the test to avoid 'unfair' consequences. Hence, a low-stakes environment, creating a sense of psychological safety, is often needed to ensure 'bad' news in a test report is read and acted upon (Hattie & Brown, 2008). Helping teachers embrace the 'bad news' of poor scores so that correct diagnosis of need and prescription of appropriate instruction are maximised is the legitimate goal of test reports. Hence, effective test reporting depends, in part, on the existence of a non-punitive professional environment anchored on educators using data to improve curriculum, instruction, and learning (Lai & Schildkamp, 2016).

Defining Score Reports

Hambleton and Zenisky (2013), the foremost of contemporary score report theorists, described score reports as the vehicle 'to convey how scores can be understood appropriately in the context of the assessment and what are the supported actions that can be taken using the results' (p. 482). Rankin (2016) defined a score report as communicating data, through tables, graphs and words in order to achieve a purpose, typically helping to turn data into actionable information, for an intended audience. Thus, score reports are the tangible communication used to disseminate scores, which are the summarised results or output of some observable phenomena (test performance), to an intended audience. A score report may be a stand-alone single report, a series of reports, it may be bespoke or automatically generated, it may be a static online reporting environment or even a dynamic online reporting system. A score report may be any combination of the above or much more. More than simply the manner in which the outcomes of testing is reported, score reports are the thin lens through which the outputs from the complex process of assessment are communicated to its audience. Indeed, score reports are, arguably, far more than simply the output of assessments; they are part of the assessment they are reporting (O'Leary, Hattie, & Griffin, 2017b).

Score reports are then of fundamental importance to the intended outcomes of testing. More than simply the afterthought to the test development process, score reports are the integral link or interface in the communication between test developers and test score users. Effectively, score reports are decision support tools (Dhaliwal & Dicerbo, 2015) and shoulder the responsibility for supporting accurate user interpretation and use of test scores. As such, their design should be focussed upon optimising user interpretation and use (Zapata-Rivera & Katz, 2014). How well a score report does, or does not, communicate its message and subsequently influence the decision and actions of their intended audience is then critical, and, arguably, as important to the notion of validity as the other psychometric properties traditionally considered when undertaking validation (Hattie & Brown, 2010). Indeed, score reports are, arguably, far more than simply the output of assessments; they are part of the assessment itself.

Accepting that score reports are the mechanism through which performance is conveyed to an intended audience, it is evident that score reports are a form of feedback to those receiving the reports. Within educational contexts, diagnostic or interim assessments are the vehicle through which teachers receive feedback about the students in their class to assist in answering the question '*who needs to be taught what next*' (Brown & Hattie, 2012). In order to close the gap between where students are and the intended curriculum goals and standards, tests have to describe diagnostically the strengths and weaknesses of a student and point to action that the teacher and/or the student can take to improve learning (Hattie & Timperley, 2007).

The Challenges of Score Reporting

Interpretation and use of scores are of critical importance to validation efforts and any subsequent claims about validity (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). However, interpretation and use of scores does not transpire purely because testing occurs. Interpretation and use are the conclusion of the complex process of testing and occurs solely because of audience engagement and interpretation of the output of test score reports. In fact, how well a score report does, or does not, communicate its message and subsequently influence the decision and actions of its intended audience is critical and as important to the notion of validity as the other psychometric properties traditionally considered when undertaking validation (Hattie, 2010).

Unfortunately, however, validity theory and validation practice rarely incorporate explicit references or guidance about how to deal with the actual (as opposed to the intended) interpretations made by report users, nor the consequential actions of score users' engagement with score reports. The literature on score report design date back almost three decades. That literature persistently identifies that test users have difficulty in understanding test scores as intended, across a range of report formats (Goodman & Hambleton, 2004; Hambleton & Slater, 1997; Jaeger, 1998; Van der Kleij & Eggen, 2013). The last 25 years has seen significant contributions to the design of test reports from the information display literature (Bertlin, 1983; Cleveland, 1994; Few, 2012; Kosslyn, 2006; Tufte, 1990, 2001; Wainer, 1997). For example, Tufte (2001) identified seven principles of graph design which are pertinent to any effort to represent test scores graphically:

1. Show the data
2. Direct the reader to think about data being presented rather than some other aspect of graph
3. Avoid distorting the data
4. Present data using the minimum of ink
5. Make large data sets coherent
6. Encourage the reader to compare different pieces of data
7. Reveal the underlying message of the data.

As a consequence of significant work (Aschbacher & Herman, 1991; Hambleton & Slater, 1997; Hambleton & Zenisky, 2013; Hattie, 2010; Impara, Divine, Bruce, Liverman, & Gay, 1991; Jaeger, 1998; Linn & Dunbar, 1992; Rankin, 2016; Zapata-Rivera & Van Winkle, 2010; Zenisky & Hambleton, 2012, 2015), there has been an evolution of guidelines relating to score reporting. These guidelines have been integrated with explicit notions of user validity (MacIver, Anderson, Costa, & Evers, 2014) and of score report interpretability as an aspect of validity (Van der Kleij, Eggen, & Engelen, 2014). The ongoing advancement of score reporting guidelines has seen a progression from recommendations about what and how to produce score reports through to iterative design methodology (Hambleton & Zenisky, 2013; Zapata-Rivera & Van Winkle, 2010).

Hattie (2010) enunciated 15 principles for the design of test reports which align in part with Tufte and also extend to address issues arising when test reports are embedded within software systems. For example, he recommends in accordance with Tufte that reports (Principle 6) minimise the amount of 'numbers' and maximise the amount of interpretations, (Principle 8) have a major theme, (Principle 10) minimise scrolling, be uncluttered, and maximise the 'seen' over the 'read'. In terms of deploying test reports within a software system, he recommends (Principle 3)

that readers of reports need a guarantee of safe passage from where they are in the system to where they want to go and (Principle 4) report readers need a guarantee of destination recovery; that is, the system must intuitively allow them to navigate among the various reports and tools within the human-computer interface. He also recommends (Principle 7) that reports be restricted in the amount of information displayed (i.e., the answer is never more than 7 plus or minus 2).

Current best practice is captured in the Hambleton and Zenisky model (2013) and comprehensively described by Zenisky and Hambleton (2015) in the *Handbook of Test Development* (Lane, Raymond, & Haladyna, 2015). This model is an iterative process of score report development and refinement. The process is conceptualised as a four step or phase model. The first phase is about laying an appropriate ground work. The second phase is about report development. The third phase is about field test and redesign. Finally, the fourth is about evaluation and maintenance. One of the key aspects that makes this model best practice is that it is focused on an ongoing process of improvement and refinement and not simply static guidelines. Consistent with the Hambleton and Zenisky model, Hattie (2010) recommended that (Principle 1) the validity of test reports be determined by the reader's correct and appropriate inferences and/or actions in response to the report, (Principle 2) evidence be obtained to demonstrate how readers interpret reports, (Principle 5) the focus be on maximising interpretations not displaying numbers, (Principle 11) reports be designed to address specific questions, (Principle 12) provide justifications that the test is fit for the specific applied purpose, and (Principle 15) reports be thought of as actions to take, not just screens to print or store.

With an eye towards the Hattie (2010) principles, O'Leary (2017) has proposed amendments to the Hambleton and Zenisky (2013) model, aimed at providing more explicit articulation to the evaluation phase of their model. The goal of those recommendations is to direct the collection of evidence concerning user comprehension of score reports. Two overarching design principles of evaluation (i.e., utility and clarity) with seven sub-domains have been promulgated (Table 8.1; O'Leary, Hattie, & Griffin, 2016b). Utility requires that score reports are designed with a clear purpose, actions, and outcomes in mind, while clarity expects score reports to be designed so that they are easily comprehensible to the target audience. These align with the proposed forms of validity evidence put forth by O'Leary, Hattie, & Griffin (2016a, 2017a). The purpose of these principles is to provide an outcomes focused lens through which score reports are considered. A rubric for evaluating the alignment of score report construction against these criteria has been developed (O'Leary, Hattie, & Griffin, 2016b) and subsequent empirical work has demonstrated

Table 8.1 Empirically derived design principles for outcomes focused evaluation of score reporting.

Utility	
Purpose	The purpose of a score report must be explicit.
Interpretation	The intended interpretations of scores must be explicit.
Actions	The intended consequences or actions of interpretation must be explicit.
Clarity	
Design Features	The design of score reports must be based upon current best practices inclusive of contemporary examples of best practices and guidelines and recommendations from within the literature.
Interpretive Guidance	Score reports must be designed to be stand-alone aiming to minimise additional work or tasks that are required to fully interpret the reported information.
Displays	Score reports must integrate multiple forms of data representation.
Language	The language used in a score report must be easily understood by the intended audience.

that the rubric is a reliable tool for obtaining evidence that 'better' designed score reports more effectively communicate their intended message (O'Leary, Hattie, & Griffin, 2017b).

These standards can be used to evaluate any test-based reports as attempts to communicate expert information to a lay end-user audience. Unsurprising then, the role of timing is not explicit. A separate argument about the importance of rapid or delayed reporting needs to be made but which could be subsumed under the notion of validity. As we have already indicated, test reports which are not available to teachers soon after test administration cannot guide instruction. Further, in light of quality management principles, providing additional insights about directions for improvement needs to be timely; and for classroom teachers timely is next Monday or tomorrow, not three months from now. Hence, well-designed test reports that arrive too late to make a difference are of little value to formative practice. The advantage of pure 'in-the-head' and 'in-the-moment' formative interactions between teachers and students (Swaffield, 2011) is that it happens immediately, while the need or opportunity is evident. Such interactions may be more error prone than tests, but they are immediate. Matching this timely facet matters to teachers and teaching; delayed reports are fundamentally purposeless.

The Audacious Example of asTTle

To illustrate the principles enunciated in this chapter, it is constructive to examine the development of New Zealand's Assessment Tools for Teaching and Learning (asTTle) test system (Hattie & Brown, 2008; Hattie, Brown, & Keegan, 2003; Hattie, Brown, Keegan, et al., 2004). The asTTle test system is an online standardised test system for reading comprehension, mathematics and writing (and the Māori language equivalents) used in New Zealand primary/elementary and secondary/high schools. The test materials have been calibrated to Levels 2 to 6 of the New Zealand national curriculum (Ministry of Education, 2007) and norms are available for students in grades 4 through to 12 (nominally ages 8 to 17). The asTTle system consists of:

1. an item bank of over 20,000 curriculum-objective and level calibrated and difficulty-calibrated multiple-choice and open-response tasks,
2. a teacher-controlled test design engine,
3. an automated test scoring engine that converted 1PL Rasch item scores to performance on achievement objectives and Curriculum Levels,
4. a reporting engine that permitted selection from a range of test reports concerning group and/or individual performance, and
5. an online catalogue of teaching resources indexed to the test reporting system.

This system was created in a policy environment that prioritised diagnostic testing for the explicit purpose of informing improved instruction and student learning outcomes (Ministry of Education, 1994). Indeed, the official policy and the rhetoric used around the research and development phases of asTTle made explicit that using the test system was a low-stakes activity; use was not required nor was reporting to government and there was no centrally determined test administration (Hattie & Brown, 2008). Furthermore, as was made clear by the Ministry of Education (2010), the system was designed to inform and support teachers by giving them access to externally-referenced norms and diagnostic curriculum-aligned reports, rather than a mechanism to be used by the Education Review Office or the media to judge or evaluate teachers and/or schools. This ensures that generation of data about student learning was done in a non-punitive manner; the goal was to inform improvement, while generating data that allowed teachers to understand how their own students compared to similar students drawn from a robust national norming (Brown & Hattie, 2012).

New Zealand primary school teachers tend to make extensive use of standardised diagnostic testing, especially at the beginning of the school year to inform within-class grouping (Crooks, 2010). It is important to note that none of the standardised tests available through the New Zealand assessment 'tool box' were compulsory or nationally administered as a national test (Brown, Irving, & Keegan, 2014); the use of all tests was completely voluntary with data retained at the school level. However, the standardised tests available before asTTle were general ability tests and reported only total score and rank order performance information. These limitations were overcome in asTTle because the system (a) allowed testing at any time, (b) allowed teachers to customise tests to classroom teaching, (c) calibration allowed different tests to be compared over time and over classes, and (d) reported performance on curriculum achievement objectives and levels, as well as normative performance.

Hence, the overall goal in designing the asTTle test report system was to give teachers a sufficiently accurate portrayal of student strengths and weaknesses so that teachers could make appropriate decisions about *who needs to be taught what's next*. This meant that the level of accuracy required in reporting a score was determined by whether the teacher would make a defensible decision about curriculum materials, pedagogical activities, or student grouping. In a sense, a principle similar to Goldilocks was used in that there were really only three options; curriculum content and material too easy, just right, or too hard. Since teachers already have a reasonably accurate sense of rank order within a class of students and have already made judgements about the curriculum level which they are teaching, a good test report system would have to go beyond this extant information. A good system would have to tell teachers something that they did not already know; teachers should be surprised rather than comforted.

Alpha Testing

To achieve this, a series of teacher interviews and focus groups were conducted early in the system development process to determine the administrative and educational goals that teachers and school leaders had for an assessment event (Meagher-Lundberg, 2000) (Note, all asTTle Technical Reports are signified * in the reference list). That research identified that information comparing their own students to national norms was desired in order to report to a range of stakeholders (e.g., parents, trustees, staff), to inform school and staff self-appraisal and professional development, to plan teaching, and to target resource commitments. Additionally, teachers wanted descriptive information relative to curriculum levels and achievement objectives. With this information, the design of test reports was initiated. This involved collaborating with a graphic artist who created simulated screen shots for a set of report templates that might achieve the various goals identified by the teachers and which were deemed to be feasible through an objectively-scored test. These report templates were iteratively presented to teacher focus groups (Meagher-Lundberg, 2001a, 2001b) to ascertain that teachers could make the intended interpretations. Initial reaction to the designs indicated a strong need for clarity as to how navigation between reports would be conducted. Teachers indicated initial designs lacked clarity as to the meaning of various communicative devices such as coloured fields depicting normative information, arrows, dials, numeric scales, labels, and the position and salience of explanatory terms (Meagher-Lundberg, 2001a).

In light of this feedback, further revisions were created taking advantage of graphical communication insights obtained from the research literature (Brown, 2001). The navigation problem was successfully addressed subsequently by placing the report images in a browser window (Meagher-Lundberg, 2001b); taking advantage of existing end-user preferences and knowledge about how software operated (Spolsky, 2001). Changes to the devices used to communicate information were generally successful according to the second focus group. This led

to the design of a report engine that included a menu system to navigate to one of the following report templates: (a) a group or cohort achievement comparison console; (b) individual and group 'kid maps'; and (c) a curriculum level achievement 'skyline' showing proportions of group performing at each level. Additional features for reporting cognitive processing against the SOLO taxonomy (Biggs & Collis, 1982) and attitudes towards tested subjects were identified for integration into either individual or group reports. Note that the sample reports shown below (Figures 8.1 to 8.4) are from the current e-asTTle version (e-asTTle Project Team, 2009). As can be seen from the following example reports, each report provided interpretive guidance on-screen, addressed a single, clear educational purpose, with the goal of supporting actual teacher decision-making.

Note that the achievement cohort comparison console (Figure 8.1) drew heavily in its design on the previously deployed CRESST Quality School Portfolio report (Baker, 1999). That report used a series of gauges and dials to capture various quality aspects of schools (e.g., safety, technology, attendance, standardised test performance, etc.) and made use of traffic light colours to indicate level of concern (i.e., red= below average; yellow=average; green=above average). This report was intended more for the cohort, subject, or school leader who needed an overview of performance relative to national normative performance. In e-asTTle, a key is used to remind the reader that the normative performance of the related comparison group is the edge between the blue field and the white space and the performance of the tested group is shown as red pointers or box plots. Any aspects of the curriculum not covered by the test are greyed out to focus interpretation on the aspects for which there was sufficient information on which to base decisions and actions. Because it may be unfair to compare 'my school' with the whole nation, especially if my school's population is drawn from either the tails or tops of the socio-economic distribution, users are able to specify the type of comparative norm by selecting either student (e.g., sex, ethnicity) or school information (i.e., school cluster). The point of this selection is not only to drill down into performance of students meriting specific attention but also to remove the obstructive claim that my students cannot achieve because they are disadvantaged; if the average for similar students or school types is higher than one's own, then it does not hold that such factors in and of themselves prevent improvement.

Likewise, the 'kidmap' reports (Figure 8.2) draw on the work of Wright and Stone (1979) in which performance is classified into one of four spatial fields or categories; that is, (a) correct and easy, (b) correct but hard, (c) incorrect but easy, and (d) incorrect and hard. This is achieved through comparison of student accuracy on the item (i.e., correct vs. incorrect) according to the difficulty of the item relative to the student's overall performance. Rather than listing items in each space, the asTTle system reports achievement objectives in each field, supplemented by item numbers in order to maximise attention on the teaching of learning outcomes rather than test items. Clearly, this report was designed for the classroom teacher or counsellor who needed to discuss with a parent or guardian the specifics of an individual child. The report uses the same conventions as the Console report to indicate overall performance relative to the same grade level norm both in terms of subject performance and motivation or attitudes. This permits partnership discussions between teacher and parent with the student to identify priorities for both work at home, as well as work in class.

To cater for the reality that teacher planning has to address groups of students (e.g., classes, grade cohorts, or special categories), the individual kidmap learning pathways report was transposed using the same colour coding to point teachers to the proportion of students having strengths, mastery, weaknesses, or gaps according to curriculum objectives (Figure 8.3). To enable priority-making decisions, teachers had only to look for objectives for which the blue space (i.e., to be achieved) were large and those in which the green space (i.e., achieved) were large. The

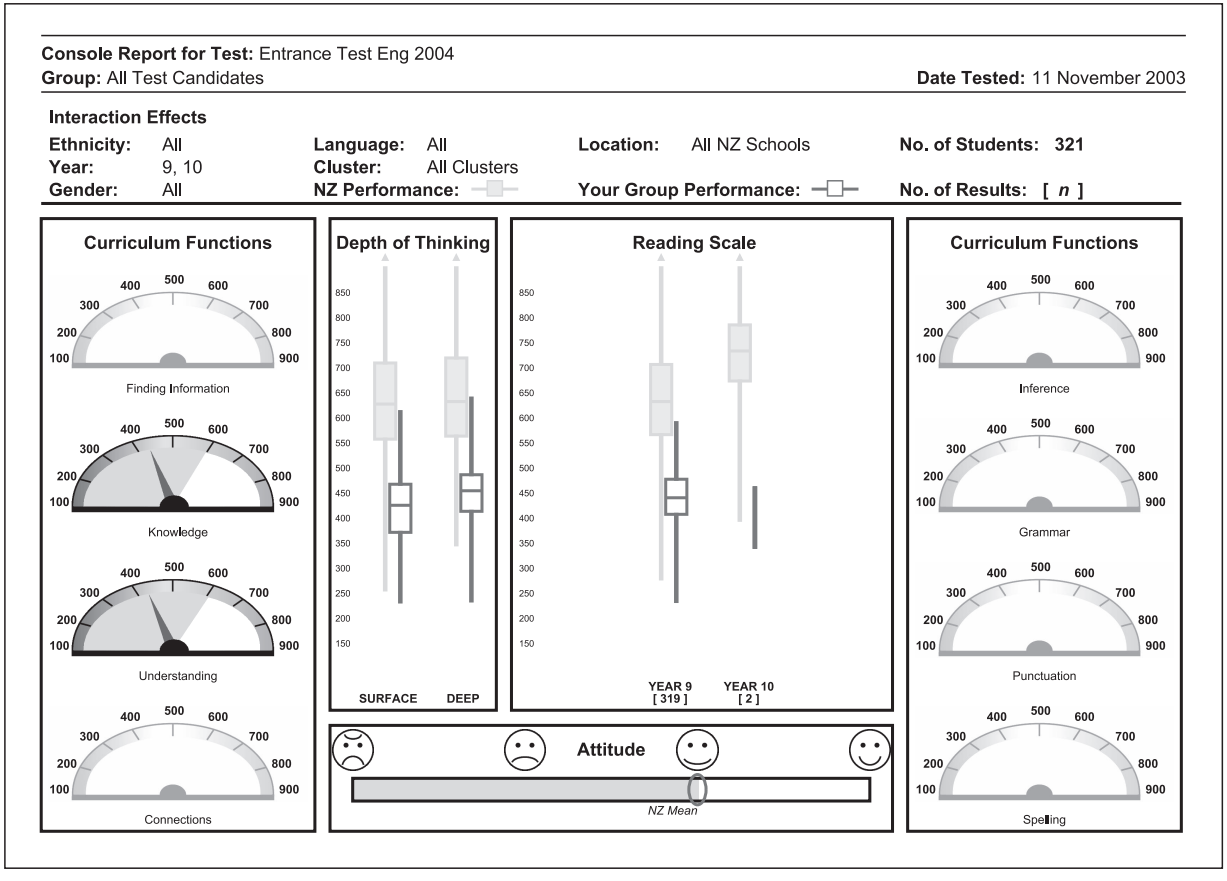


Figure 8.1 e-asTTle console report.

(e-asTTle Project Team, 2009, p. 78)

Note: Light gray is displayed as blue on-screen, the pointers and dark gray boxes are red.

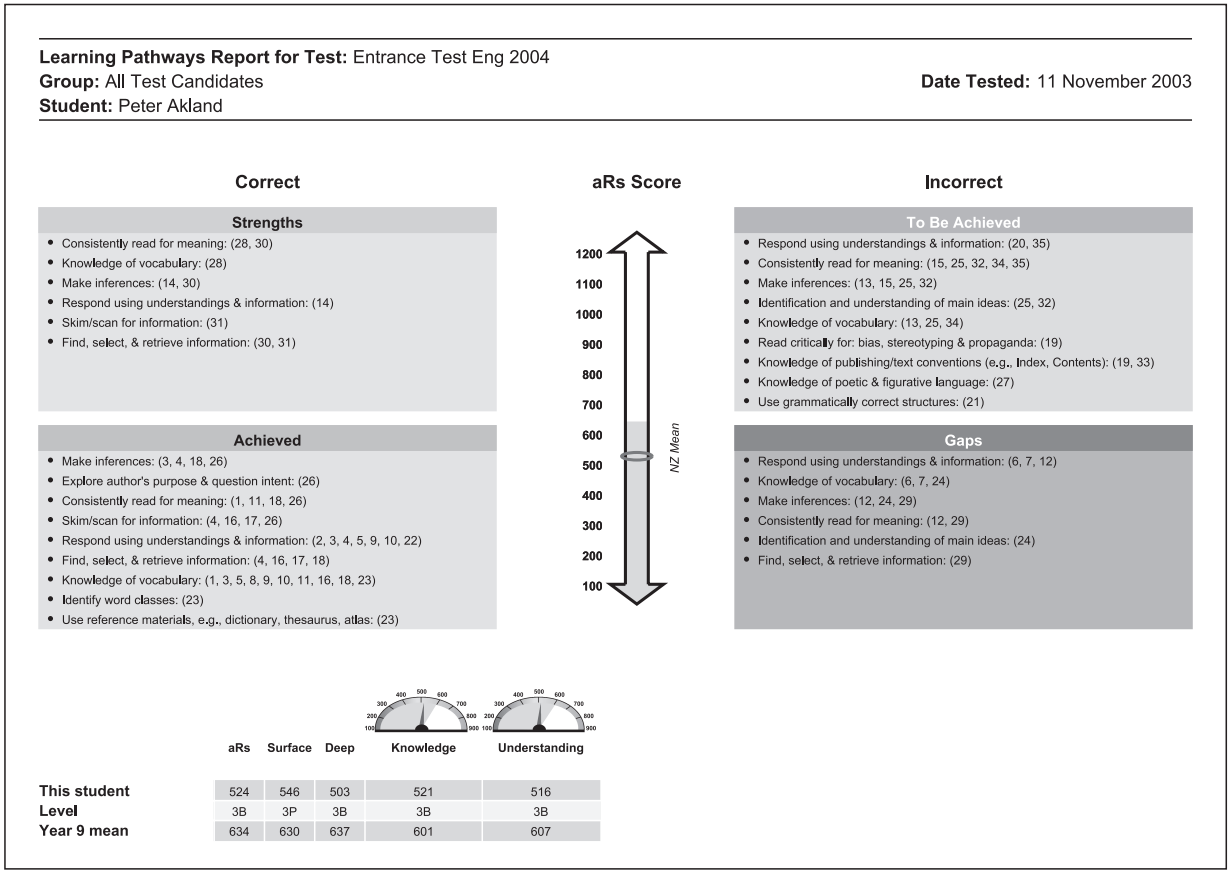


Figure 8.2 e-asTTle learning pathways report for an individual student.

(e-asTTle Project Team, 2009, p. 84)

Note: Achieved is displayed as green on-screen; Strengths is yellow; Gaps are red; and To Be Achieved is blue. The ellipse and pointers are red on a blue background.

Group Learning Pathways Report for Test: Entrance Test Eng 2004
Group: All Test Candidates
Group Size: 321
Date Tested: 11 November 2003

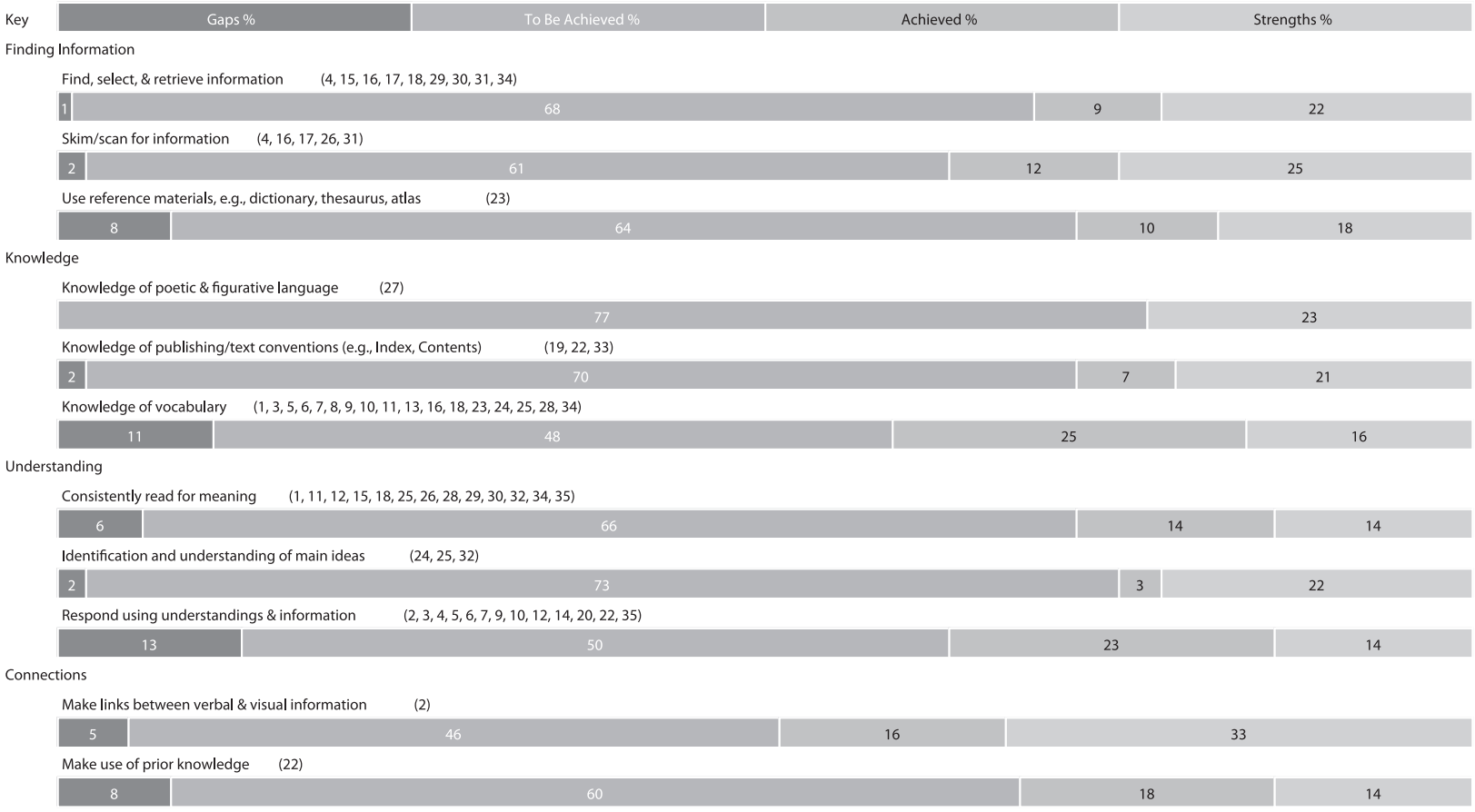


Figure 8.3 e-asTTle group learning pathways report.
(e-asTTle Project Team, 2009, p. 88)
Note: Achieved is displayed as green on-screen; Strengths is yellow; Gaps are red; and To Be Achieved is blue.

former indicated content that a high proportion of students needed to be taught, while the latter indicated material on which a high proportion needed no further instruction or practice.

Unsurprisingly, teaching to the mean will disguise the distribution of performance. Hence, the system provides a distribution of performance report (i.e., Curriculum Levels Report; Figure 8.4), which reveals both central tendency and distribution. Because New Zealand primary school teachers practice considerable within-class ability grouping, each 'skyline', when selected, displays the names of students in each performance group. This allows teachers to move children into different grouping combinations according to identified needs, rather than create persistent groups across all learning areas. In fact, this ability to differentiate for grouping was noted by early adopting teachers and their students as a positive facet of the system (Archer & Brown, 2013).

We suggest that the suite of reports and the ability to customise those for the multiple purposes of classroom teachers and school leaders meant that the system complies with the expectations of good reporting outlined in Table 8.1 and earlier.

Beta Testing

Having established through 'alpha' testing reasonably robust communicative test reports, these designs were further refined through 'beta' feedback from (a) Ministry of Education officials who were the funders and sponsors of the asTTle system, (b) the software engineering team who advised on feasibility and cost of various design options, (c) pilot testing by teachers who were exposed to a mock-up of the system, and (d) acceptance testing of asTTle version 1, containing materials for reading and writing only, which was deployed to 110 primary schools. As each stage of beta testing was conducted, formative changes were made to the asTTle system to achieve the curricular goal of helping teachers know what to teach to which students.

The evaluation of the pilot implementation of asTTle (v1) into 110 New Zealand schools (Ward, Hattie, & Brown, 2003) used a survey to ascertain, among others, the ability of teachers to accurately interpret asTTle reports. The survey included a set of report reading comprehension items, partially inspired by Hambleton and Slater (1997) and Linn and Dunbar (1992). Results indicated that in general, the Console Reports and the What Next reports had reasonably high levels of correct interpretation, whereas the means were much lower for Individual Learning Pathways and Curriculum Levels reports (Hattie, Brown, Ward, Irving, & Keegan, 2006). These results were incorporated into a structural equation model as dependent variables. The model proposed that attitudes towards computers, ICT, assessment and professional development would predict the level of involvement teachers had with the asTTle test system, which, in turn, would predict the teacher's evaluation of asTTle and their ability to answer the report reading comprehension questions. Indicating a belief that assessment is powerful for improving teaching, rather than for evaluating schools, and seeing the asTTle software as positive were clear predictors of accuracy in report interpretation (Hattie et al., 2006). The major messages were that professional development needed to be oriented most towards encouraging a positive attitude towards using ICT-based assessment as part of teaching and learning. This information was used to improve the quality and quantity of professional development resources supplied to asTTle users by the Ministry of Education. It was also used to indicate what should be in the professional development—clearly teachers needed assistance in accurately understanding and, thus, using asTTle correctly.

Based on these studies, the Ministry of Education funded for several years multiple mechanisms to support teacher learning in making use of the asTTle system. A free-phone technology-oriented help desk was deployed so that callers using asTTle and e-asTTle on their local work-stations, school-based servers, and eventually the internet could have prompt help. When

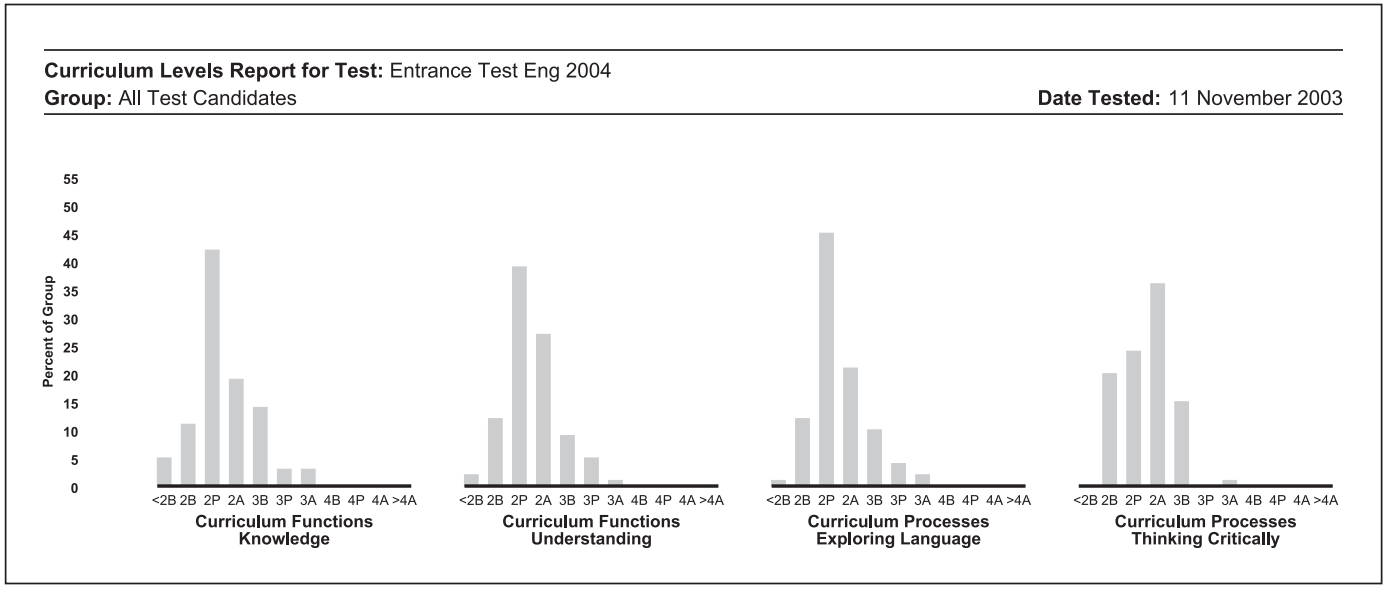


Figure 8.4 e-asTTle curriculum levels report.
(Hattie, Brown, Keegan, et al., 2004, pp. 3–18)

installed, the asTTle system provided user manuals and technical reports to support understanding and use of the system. These documents were also available online from the Ministry repository of reports and documents (http://e-asttle.tki.org.nz/). Throughout the nation, assessment-focused teacher professional development teams (Assess to Learn; AtoL) were commissioned and funded to provide within school services focused on the logic of using the asTTle reports to improve and guide instruction and reporting. The asTTle Project development team provided initial briefing to AtoL teams but was explicitly excluded from the delivery of school-based training. Nonetheless, it was apparent that the effectiveness of AtoL teams depended, in part, on the existing conceptions teachers had of the purpose of assessment—the more they considered assessment was for accountability, the less use they made of asTTle for improvement (Brown & Harris, 2009).

Extension to Secondary/High School

The original requirements brief for the asTTle system was focused solely on primary/elementary schooling; that is, Curriculum Levels 2 to 4, with norms for students in Years 5–8 only. However, given the success of asTTle v2 in primary schools, the Ministry of Education received vigorous requests from secondary/high school teachers and their union, the Post-Primary Teachers Association for extension of the system to include their students (Brown, 2013). The logic was reasonably simple: although the curriculum framework expects that Level 4 will be completed by the end of primary schooling (Year 8), empirical realities are such that many students arrive at high school still functioning at Levels 2 to 4 (Satherley, 2006). An environmental constraint in secondary schooling, not present in primary schooling, is the important role secondary schools play in preparing students for and administering formal qualifications assessments (i.e., the National Certificate of Educational Achievement-NCEA) (Crooks, 2010).

The NCEA begins with Level 1 in Year 11, culminating in Year 13 with Level 3. Nominally, NCEA Level 1 is equated to Curriculum Level 6, though some achievement objectives for Level 6 are taught in Year 12 rather than Year 11. The NCEA system evaluates student learning using a criterion-referenced, standards-based grade system (i.e., Not Achieved, Achieved, Merit, Excellence), somewhat akin to more conventional letter grade systems (i.e., D/F, C, B, A). NCEA also structures the curriculum objectives around units of work known as standards; this means that alignment of test items to NCEA standards might be of value to secondary teachers. This high-stakes evaluation system predominates educational assessment in New Zealand secondary school systems and so the possibility that the asTTle reports could be modified to accommodate this alternative system to the curriculum levels framework was explored. Additionally, within the framework of beta testing asTTle v3 in 55 secondary schools, accuracy and sufficiency of the test reports was conducted through a mixture of surveys, telephone interviews, and focus groups (Hattie, Brown, Irving, et al., 2004).

As reported in Hattie et al. (2006), secondary school teachers were positive about the asTTle reports, expressing satisfaction with the amount of detail on reports and the relevance of the reports to their needs. Teachers reported significant help from the formative and diagnostic reporting functions at both aggregated and disaggregated levels of reporting, especially in the Group Learning Pathways Report and the Individual Learning Pathways Report. In addition, they found benefit from the aggregated data in the Tabular, Curriculum Levels, and Console reports. Several enhancements based on feedback on asTTle v2 were evaluated positively. For example, instead of just reporting group means on the console report with an ellipse (i.e., *M*+/-*se*), a box-and-whisker plot showed the distribution of scores for the group being reported. The display of the national norm score as a coloured field within the dials instead of as a number below was also seen as an enhancement.

Secondary teachers indicated value in two new types of report. Focus group participants indicated value in longitudinal reports that showed how individuals or cohorts had been progressing over time. While the asTTle system had already included the ability to compare scores to similar students (i.e., schools like mine; Hattie, 2002), the ability to compare performance to different rather than similar categories (e.g., higher performing clusters or ethnicities) was seen as valuable.

Nonetheless, secondary teachers indicated significant concern about the correct or accurate interpretation of the asTTle reports. These concerns were obtained both from the Ministry Telephone Helpdesk as well as directly from the evaluation study. Confidence that reports were being understood and acted upon appropriately mattered to the teachers and needed to be addressed through modifications to the Ministry's professional development support services and asTTle documentation. Although most of the information sought by asTTle V3 users about report interpretation was available through the PDF manuals included with the asTTle V3 software, it was decided to develop an online tutorial system on understanding asTTle reports that could be used by individuals or schools as a supplement or alternative to professional development (Hattie, Brown, Irving, MacKay, & Campbell, 2005). Unlike later online tutorials that used video (Zapata-Rivera, Zwick, & Vezzu, 2016), these tutorials were slide presentations with voice over scripted dialogue that could be controlled by the user.

Lack of alignment to the NCEA system beginning in Year 11 meant that most secondary teachers had implemented asTTle V3 with students only in Years 9 and 10. However, when shown the possible reports and tests that asTTle might be able to generate for them as indicators of NCEA performance, teachers were quite enthusiastic. Teachers indicated that it was important or very important to know how the curriculum-level indexed items in an asTTle test related to the NCEA system. Of special interest to the participants would be the ability to create a test aligned to the various standards of the NCEA system, rather than to the achievement objectives of the curriculum. Despite the strong endorsement of the sampled teachers for adjustments of the asTTle reports to align with the official qualifications framework, the Ministry of Education sponsors declined to fund such research or developments. Perhaps, because the NCEA system is administered by a separate quasi-autonomous body (i.e., New Zealand Qualifications Authority; NZQA), such a development funded by the Ministry may have been seen as a breach of NZQA's autonomy and responsibility.

Aside from any systemic 'turf' issues, this last point raises some interesting challenges around alignment of formative and summative purposes. It may be that refusal to adapt asTTle to align with high-stakes qualifications system was that this may constitute a threat to the intention that asTTle serve goals related to diagnostic formative improvement of teaching and learning (Brown, 2004). When teachers perceive that assessments are for accountability purposes, our own studies have found that it is a rare teacher who can balance the tension between improvement of my teaching and evaluation of school quality (Brown & Harris, 2009). Indeed, this tension between the purposes or goals of assessment for improvement and assessment for accountability seems to remain more or less unresolved (Barnes, Fives, & Dacey, 2015; Bonner, 2016). As long as test systems are used by interested stakeholders to evaluate the work of teachers and school leaders, we can expect that there will be greater attention and effort paid to raising scores than improving instruction (Nichols & Harris, 2016). Hence, while the technical capacity exists to align formative and summative systems and information into a single test reporting framework, the real obstacles lie in political factors that may subvert well-meaning integration. Unless policy makers are willing to partner with the teachers and respect their legitimate concerns by attaching low-stakes to tests, it seems highly implausible that improvement and accountability can be effective bed-partners. Indeed, as long as high-stakes testing or examination or school accountability testing dominate the educational landscape, policies to support or require formative assessment are unlikely to be seen as 'the real thing' (Kennedy, Chan, & Fok, 2011).

Conclusion

This chapter has outlined the major challenges that face developing and validating test reports for teachers. The field has developed a reasonably robust understanding of why this has to be done and how it can be done. However, few test systems have conducted such time and resource-consuming programmes of formative evaluation and documented them as has the New Zealand asTTle system. This system is an exemplar of how accuracy in interpretation of reports and subsequent actions can be established. Clearly, the field needs more such studies that establish the validity of tests for use by communities of educators, each of which share different standards and approaches to assessment, ICT, and schooling in general.

References

*All asTTle reports are retrieved from https://e-asttle.tki.org.nz/Reports-and-research/asTTle-technical-reports

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (4th ed.). Washington, DC: American Educational Research Association.

Archer, E., & Brown, G. T. L. (2013). Beyond rhetoric: Leveraging learning from New Zealand's assessment tools for teaching and learning for South Africa. *Education as Change*, *17*(1), 131–147. doi:10.1080/16823206.2013.773932

Aschbacher, P. R., & Herman, J. L. (1991). *Guidelines for effective score reporting* (CSE Technical Report 326). Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Baker, E. L. (1999). *Technology: Something's coming-something good* (CRESST Policy Brief 2). Los Angeles: UCLA Graduate School of Education & Information Studies, National Center for Research on Evaluation, Standards, and Student Testing.

Barnes, N., Fives, H., & Dacey, C. M. (2015). Teachers' beliefs about assessment. In H. Fives & M. Gregoire Gill (Eds.), *International handbook of research on teacher beliefs* (pp. 284–300). New York, NY: Routledge.

Bertlin, J. (1983). *Semiology of graphics*. Madison, WI: The University of Wisconsin Press.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. New York, NY: Academic Press.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3–29). Dordrecht, The Netherlands: Springer.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Bloom, B., Hastings, J., & Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw Hill.

Bonner, S. M. (2016). Teachers' perceptions about assessment: Competing narratives. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 21–39). New York, NY: Routledge.

*Brown, G. T. L. (2001). *Reporting assessment information to teachers: Report of project asTTle outputs design* (asTTle Tech. Rep. #15). Auckland, NZ: University of Auckland, Project asTTle.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy and Practice*, *11*(3), 301–318. doi:10.1080/0969594042000304609

Brown, G. T. L. (2013). asTTle—A national testing system for formative assessment: How the national testing policy ended up helping schools and teachers. In M. Lai & S. Kushner (Eds.), *A national developmental and negotiated approach to school and curriculum evaluation* (pp. 39–56). London: Emerald Group Publishing. doi:10.1108/S1474-7863(2013)0000014003

Brown, G. T. L., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of MultiDisciplinary Evaluation*, *6*(12), 68–91.

Brown, G. T. L., & Hattie, J. A. (2012). The benefits of regular standardized assessment in childhood education: Guiding improved instruction and learning. In S. Suggate & E. Reese (Eds.), *Contemporary debates in childhood education and development* (pp. 287–292). London: Routledge.

Brown, G. T. L., Irving, S. E., & Keegan, P. J. (2014). *An introduction to educational assessment, measurement, and evaluation: Improving the quality of teacher-based assessment* (3rd ed.). Auckland, NZ: Dunmore Publishing.

Cleveland, W. S. (1994). *The elements of graphing data*. Monterey, CA: Wadsworth.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(4), 438–481.

Crooks, T. J. (2010). Classroom assessment in policy context (New Zealand). In B. McGraw, P. Peterson, & E. L. Baker (Eds.), *The international encyclopedia of education* (3rd ed., pp. 443–448). Oxford: Elsevier.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, *64*(1), 5–31. doi:10.17763/haer.64.1.j57n353226536276

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Dhaliwal, T., & Dicerbo, K. E. (2015, April). *Presenting assessment data to inform instructional decisions*. Paper presented at the Annual meeting of the American Educational Research Association, Chicago, IL.

e-asTTle Project Team. (2009). *Generation 2: e-asTTle year three educator manual*. Auckland, NZ: Visible Learning Lab, Auckland UniSevices, Ltd.

Few, S. (2012). *Show me the numbers: Designing tables and graphs to* en*lighten*. Oakland, MA: Analytic Press.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145–220.

Griffin, P. (2014). *Assessment for teaching*. Port Melbourne: Cambridge University Press.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, *11*(1–2), 1–18. doi:10.1080/15366367.2013.783752

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policymakers and educators?* (Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

*Hattie, J. A. (2002). *Schools like mine: Cluster analysis of New Zealand schools*. (Tech. Rep. No. 14). Auckland, NZ: University of Auckland, Project asTTle.

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 479–494). Washington, DC: American Psychological Association.

Hattie, J. A. (2010). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*, 1–15. Retrieved from www.oerj.org/View?action=viewPaper&paper=6.

Hattie, J. A., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, *36*(2), 189–201. doi:10.2190/ET.36.2.g

Hattie, J. A., & Brown, G. T. L. (2010). Assessment and evaluation. In C. Rubie-Davies (Ed.), *Educational psychology: Concepts, research and challenges* (pp. 102–117). Abingdon: Routledge.

*Hattie, J. A. C., Brown, G. T. L., Irving, S. E., Keegan, P. J., Sussex, K., Cutforth, S., . . . MacKay, A. J. (2004, September). *Use of asTTle in secondary schools: Evaluation of the pilot release of asTTle V3* (asTTle Tech. Rep. #47), Auckland, NZ: University of Auckland/Ministry of Education.

*Hattie, J. A., Brown, G. T. L., Irving, S. E., MacKay, A. J., & Campbell, A. (2005). *Using asTTle: A teachers' guide* (asTTle Tutorial). Retrieved from www.breezeserver.co.nz/p85512844

Hattie, J. A., Brown, G. T. L., & Keegan, P. J. (2003). A national teacher-managed, curriculum-based assessment system: Assessment tools for teaching & learning (asTTle). *International Journal of Learning*, *10*, 771–778.

*Hattie, J. A., Brown, G. T. L., Keegan, P. J., MacKay, A. J., Irving, S. E., Cutforth, S., . . . Yu, J. (2004, December). *Assessment tools for teaching and learning (asTTle) manual* (Version 4, 2005). Wellington, NZ: University of Auckland/Ministry of Education/ Learning Media.

Hattie, J. A., Brown, G. T. L., Ward, L., Irving, S. E., & Keegan, P. J. (2006). Formative evaluation of an educational assessment technology innovation: Developers' insights into assessment tools for teaching and learning (asTTle). *Journal of MultiDisciplinary Evaluation*, *5*(3), 1–54.

Hattie, J. A. C. (2014). The last of the 20th century test standards. *Educational Measurement: Issues and Practice*, *33*(4), 34–35.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). Formative use of test results: A user perspective. *Studies in Educational Evaluation*, *52*, 12–23.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice*, *10*(4), 16–18.

Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress* (NVS NAEP Validity Studies). Washington, DC: American Institutes for Research.

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

Kennedy, K. J., Chan, J. K. S., & Fok, P. K. (2011). Holding policy-makers to account: Exploring 'soft' and 'hard' policy and the implications for curriculum reform. *London Review of Education*, *9*(1), 41–54. doi:10.1080/14748460.2011.550433

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. New York, NY: Oxford University Press.

Lai, M. K., & Schildkamp, K. (2016). In-service teacher professional learning: Use of assessment in data-based decision-making. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 77–94). New York, NY: Routledge.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.) (2015). *Handbook of test development*. New York, NY: Routledge.

Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*(2), 177–194.

MacIver, R., Anderson, N., Costa, A-C., & Evers, A. (2014). Validity of interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment*, *22*(2), 149–164. doi:10.1111/ijsa.12065

*Meagher-Lundberg, P. (2000). *Report on comparison groups/variable for use in analysing assessment results* (asTTle Tech. Rep. #1). Auckland, NZ: University of Auckland, Project asTTle.

*Meagher-Lundberg, P. (2001a). *Report on output reporting design: Focus group 1* (asTTle Tech. Rep. #9). Auckland, NZ: University of Auckland, Project asTTle.

*Meagher-Lundberg, P. (2001b). *Report output reporting design: Focus group 2* (asTTle Tech. Rep. #10). Auckland, NZ: University of Auckland, Project asTTle.

Ministry of Education. (1994). *Assessment: Policy to practice*. Wellington, NZ: Learning Media.

Ministry of Education. (2007). *The New Zealand curriculum for English-medium teaching and learning in years 1–13*. Wellington, NZ: Learning Media.

Ministry of Education. (2010). *OECD review on evaluation and assessment frameworks for improving school outcomes: New Zealand country background report 2010*. Wellington, NZ: Ministry of Education.

Nichols, S. L., & Harris, L. R. (2016). Accountability assessment's effects on teachers and schools. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 40–56). New York, NY: Routledge.

O'Leary, T. M. (2017). *Effective score reporting: Establishing evidence informed design principles for outcomes focused score reports*. Unpublished Ph.D. thesis, The University of Melbourne, Melbourne, Australia.

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017a). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice*, *36*(2), 16–23. doi:10.1111/emip.12141

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017b, April). *Evaluating the effectiveness of score reports: Do better designed score reports result in better interpretation?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

O'Leary, T. M., Hattie, J., & Griffin, P. (2016a, April). *Reconceptualising validity evidence including evidence of user interpretation*. Paper presented to Annual Conference of the NCME, Washington, DC.

O'Leary, T. M., Hattie, J., & Griffin, P. (2016b, July). *Design principles for action and outcome focused score report design*. Presentation at the biennial meeting of the International Test Commission, Vancouver, BC.

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

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (6th ed.). Boston, MA: Allyn & Bacon.

Rankin, J. G. (2016). *Standards for reporting data to educators: What educational leaders should know and demand*. New York, NY: Routledge.

Satherley, P. (2006). *Student outcome overview 2001–2005: Research findings on student achievement in reading, writing and mathematics in New Zealand schools*. Wellington, NZ: Ministry of Education, Research Division.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation & education: At quarter century* (Vol. 90, Part II, pp. 19–64). Chicago, IL: NSSE.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1066–1101). Washington, DC: American Educational Research Association.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Westport, CT: Praeger.

Spolsky, J. (2001). *User interface design for programmers*. Berkeley, CA: APress LP.

Swaffield, S. (2011). Getting to the heart of authentic Assessment for Learning. *Assessment in Education: Principles, Policy & Practice*, *18*(4), 433–449. doi:10.1080/0969594X.2011.582838

Torrance, H. (1986). Expanding school-based assessment: Issues, problems and future possibilities. *Research Papers in Education*, *1*(1), 48–59. doi:10.1080/0267152860010104

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretations of the score reports from the computer program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation*, *39*, 144–152.

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the computer program LOVS: A redesign study. *Studies in Educational Evaluation*, *43*, 24–39. doi:10.1016/j.stueduc.2014.04.004

Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. New York, NY: Copernicus Books.

*Ward, L., Hattie, J. A., & Brown, G. T. (2003, June). *The evaluation of asTTle in schools: The power of professional development* (asTTle Tech. Rep. #35). Auckland, NZ: University of Auckland/Ministry of Education.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Zapata-Rivera, D., & VanWinkle, W. (2010). *A research-based approach to designing and evaluating score reports for teachers* (Research Memorandum 10–01). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*(3), 215–229. doi:10.1080/10627197.2016.1202110

Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, *21*, 442–463. doi:10.1080/0969594X.2014.936357

Zenisky, A., & Hambleton, R. K. (2015). Good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 585–602). New York, NY: Routledge.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. Educational Measurement: Issues and Practice, *31*(2), 21–26.

Zumbo, B. D. (2009). Validity as a contextualised and pragmatic explanation, and its implications for validation practice. In R. W. Lizzitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). Charlotte, NC: IAP-Information Age Publishing, Inc.

9

Applying Learning Analytics to Support Instruction

Mingyu Feng, Andrew Krumm, and Shuchi Grover

This chapter highlights the ways in which *learning analytics* can be used to better understand and improve learning environments, instruction, and assessment (Siemens & Long, 2011). As a set of approaches for engaging in educational research, learning analytics and educational data mining represent relatively new modes of inquiry. The growth of these approaches maps closely to the availability of new forms of data being collected and stored in digital learning environments, administrative data systems, as well as sensors and recording devices. Moreover, the growth of these fields maps closely onto what the National Science Foundation refers to as "data-intensive research," which encompasses more than learning analytics and educational data mining to include a broad range of social and physical sciences. As new forms of data have emerged (i.e., transaction level data from digital learning environments as well as digital forms of audio, video, and text) and been collected at ever increasing scales, there has been an explosion of efforts to make use of these data for the purposes of research. By and large, most early work beginning in the mid-2000s was directed at exploring research questions that were tractable within highly structured, well-designed digital learning environments like intelligent tutoring systems (ITS; e.g., Koedinger, Anderson, Hadley, & Mark, 1997; VanLehn et al., 2005). The tight alignment between the learning tasks students were expected to engage in and the data that were collected in these environments made them ideal for exploring not just the outcomes of learning but the various ways in which students engaged in learning activities. A basic insight from these early researchers continues to fuel research and efforts to improve instruction—data on students' learning processes is as useful and sometimes more so than data on students' learning outcomes.

In this chapter, we expand upon this insight and highlight the ways in which data from digital learning environments, administrative data systems, and sensors as well as recording devices can be used to support instruction in real classrooms by reporting on students' learning activities through various data products (e.g., dashboards). We do so across four cases that represent varying degrees of proximity to instruction. By highlighting these varying degrees of proximity, we intend to demonstrate the multiple ways in which learning analytics can be used to support instruction. Cases 1 and 2 describe efforts to use learning analytics to support instruction

through partnerships that bring researchers closer to practice and practitioners closer to the work of analytics. Case 3 describes how process data from digital learning environments can be used to develop better assessments of learning that can be used to organize better learning opportunities for students. Case 4 describes how providing practitioners with access to carefully designed data products and dashboards can help them in making more timely and targeted decisions. The data produced in each of these cases are shared with stakeholders in different ways including online report systems or dashboards (Corrin, this volume).

Overview of Cases

Case 1 with Summit Public Schools and case 2 with the Carnegie Math Pathways are based in an approach to using analytics that Krumm and colleagues refer to as collaborative data-intensive improvement (CDI; Krumm, Means, & Bienkowski, 2018). Collaborative data-intensive improvement is an approach that combines tools and routines from improvement science, data-driven decision-making, as well as learning analytics and educational data mining. The overarching goal of this approach is to provide a structured way for researchers and practitioners to work together around identifying a question to pursue, analyzing complex datasets, developing change ideas, and testing change ideas in local learning environments. Across multiple partnerships, Krumm and colleagues (2018) identified a series of phases and supporting conditions for using data from digital learning environments and administrative data systems to improve learning environments and support instruction. Phase I of a collaborative data-intensive improvement project involves setting up a partnership, which includes identifying participants and jointly defining the aim of the partnership (Bryk, Gomez, Grunow, & LeMahieu, 2015). The second phase entails developing a practical theory for how the partnership will reach its aim (Bennett & Provost, 2015; Yeager, Bryk, Muhich, Hausman, & Morales, 2013). Phase III centers on data wrangling, exploration, and modeling (Wickham & Grolemund, 2017). Phase IV builds on insights from data-intensive analyses in the form of co-developed change ideas, and lastly, Phase V is where members of a partnership iteratively refine change ideas in real classrooms over time. Cases 1 and 2 describe the partnerships from which many of these phases were identified (Krumm, 2017).

Case 3 is situated in the context of introductory programming and computational thinking (CT), a new skill that seeing rapid adoption at all levels of school curricula as part of nationwide efforts to support "Computer Science for All" (The White House, 2016). There is a growing need to measure students' learning of computational thinking in the context of the complex problem-solving processes inherent in programming, and also support all learners through this process of learning computational problem solving. Given that there are few examples of using learning analytics to measure students' learning in open-ended programming environments that are popularly used in K-12 classroom, Grover and colleagues push into the emerging realm of *computational psychometrics* (von Davier, 2017) for detection of student behavior for formative assessment (Black & William, 2009; Heritage & Popham, 2013). They explored how principled, top-down, approaches of measuring complex skills can be combined with bottom-up, data-driven learning analytics approaches for better interpretation (of data logs from such programming environments), and consequently better measurement of computational thinking practices and programming processes (Zapata-Rivera, Liu, Chen, Hao, & von Davier, 2016). Based on learnings from analyzing data logs from ~300 students using the Alice programming environment, they developed a framework (Grover et al., 2017) that formalizes a process where a hypothesis-driven approach informed by Evidence-Centered Design effectively complements data-driven learning analytics in interpreting students' programming process and assessing computational thinking in block-based programming environments. The framework is shared here, as well as a brief description of the application of the framework on an ongoing research project.

Case 4 is based in recent instructional reforms that address the importance of administrator and teacher making use of student assessment data to inform decisions about curriculum and instruction (Means, Padilla, DeBarger, & Bakia, 2009) and thus making instructional practices more effective (Mandinach & Gummer, 2016). The advances in technology and its popularity in schools have made it easier to collect student performance data. Learning analysts build dashboards and a variety of reports to incorporate information from such data, together with other possible sources of data, and present them to teachers. Although there have been many different types of dashboards built, few studies have shown evidence that teachers make use of information presented on the dashboards and adjust instructions accordingly. Case 4 describes an online homework support tool that was implemented in 44 schools for two years during a large-scale efficacy trial. Data collected during the study suggested that teachers implementing the intervention made substantial shifts in their approach to homework review and instructional practice more broadly.

Case 1: Data-Intensive Research-Practice Partnership

The partnership with Summit Public Schools (Summit) began in the fall of 2014 with the goal of developing a research-practice partnership around data collected and stored in multiple online learning systems used throughout the charter management organization. The partnership included researchers and practitioners from multiple organizational levels at Summit. The partnership built on the ideas of (a) learning directly from practitioners about the problems they experience in their day-to-day work; (b) jointly analyzing and interpreting data generated by students in digital learning environments to solve practitioner-identified problems; and (c) co-developing ideas for changes informed by multiple data-intensive analyses.

Around the same time as the fields of learning analytics and educational data mining were coalescing, new partnership models for engaging in educational research were emerging under the banner of research-practice partnerships (e.g., Coburn, Penuel, & Geil, 2013). Newer forms of data combined with newly developing models of research served as the primary building blocks for the partnership. The research goals for the project included (a) using Summit's increasingly diverse and sizable datasets to answer their own research questions, and through engaging in these analysis activities, (b) develop a generalizable set of tools and routines for engaging in collaborative data-intensive research that other partnerships could use. To accomplish these research goals, we (i.e., Krumm and colleagues) used a design-based research approach (e.g., Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). A central feature of design-research is that it represents a mode of inquiry that seeks to build theory through directly intervening on learning environments (e.g., Barab & Squire, 2004; Bell, 2004). To inform our design, development, and intervention activities, we used theory and prior research from data-driven decision-making (e.g., Boudett, City, & Murnane, 2013), research-practice partnerships (e.g., Coburn, Penuel, & Geil, 2013), and learning analytics as well as educational data mining (e.g., Baker & Siemens, 2014).

Students in Summit consistently interact with digital learning environments in all grades and subject areas. This level of interaction results in a large volume of structured data on both what students are doing as they engage in learning tasks and how well they perform on those tasks. Summit believes that every student is capable of being college and career ready and that personalized learning opportunities can help students build necessary knowledge, habits, and skills. To accomplish this, Summit developed a whole-school approach where students engage in (a) project-based learning, (b) personalized learning time, and (c) one-on-one mentoring with teachers. Across these three learning opportunities, which span all grades and subjects, students interact with a common learning management system called the Summit Learning Platform (SLP). The

platform houses teacher-curated digital learning resources, such as online videos, and two types of assessments based on "playlists." One of the two types of assessments referred to as a diagnostic assessment is used to identify gaps in students' knowledge and a summative assessment, referred to as a content assessment, pulls randomly selected items from a large item-bank and is used to identify whether students have achieved mastery for a focal content area. Students can access resources and take assessments at their own discretion and as many times as necessary. Students spend 30% of their instructional time engaged in self-directed learning as they work to complete multiple summative assessments for a course and spend the remaining percentage of instructional time (i.e., 70%) engaged in project-based learning, which is an approach to instruction where students gain knowledge, skills, and productive dispositions by developing authentic products that are organized around broad and motivating driving questions (Larmer, Mergendoller, & Boss, 2015).

A challenge for any partnership is developing a focus for the partnership's work (Penuel & Gallagher, 2017). To set the research direction for the partnership, we engaged in a multi-meeting, iterative process of having practitioners from Summit brainstorm topics and questions and having researchers reflect back and react to each question. Based on this process, the first data-intensive analyses that the partnership engaged in addressed the ways in which students attempted and completed content assessments. As noted previously, students have discretion in terms of when they attempt content assessments and how they prepare for them. Using data from students' interaction with the Summit Learning Platform, we initially examined relationships among students' standardized test performances on the NWEA MAP, their use of teacher-curated resources, and their content assessment taking in relation to course grades. Based on these analyses and within math courses, we observed that students who had lower incoming MAP math test scores tended to attempt content assessments more frequently. Beyond this, we also observed that students with higher MAP math scores used the Summit Learning Platform in different ways than their peers with lower incoming scores. For example, students with higher incoming test scores, on average, used *more* unique learning resources. For example, on one math playlist called "linear functions," students in the lowest quintile on the MAP Math test used approximately 17 unique teacher-curated resources whereas students in the highest quintile used approximately 21. These same students also, as compared to their peers, used more resources *prior* to taking their first content assessment and overall attempted content assessments many fewer times. Overall, these early analyses hinted at the potential for using data from the platform to inform instruction—it provided a window into the processes that students were engaging in that held potential for explaining students' eventual performances on individual playlists and for the course overall.

A key element of the overall partnership was working with practitioners at multiple levels of Summit—from organizational leaders to teachers—to make sense of data generated by the platform. Over time, we came to view opportunities to work directly with practitioners as *learning events* (Cobb & Jackson, 2012). These learning events proved to be the primary locations for going from a data product developed by researchers to a set of implications that would kick off the development of concrete instructional change ideas. At a general level, learning events involved structured activities where members of the partnership developed new understandings by engaging in joint work. Types of learning events included simple meetings where members of the partnership jointly interpreted data products, but learning events also included structured co-design session (Penuel, Roschelle, & Shechtman, 2007), opportunities where teachers could work first hand with data products that researchers developed, and workshops where researchers provided explicit instruction on data analysis software. Learning events provided opportunities for both researchers and practitioners to use their respective expertise to make the data useful for instructional improvement.

At a multi-day learning event referred to as a *data sprint*, Summit staff worked directly with data and engaged in data wrangling, exploration, and modeling tasks in collaboration with researchers. One data product that the partnership built upon out of this event involved an analysis that identified students who scored low on an assessment and followed it up with another assessment—and often another low score. This cycle of repeated, negative assessment taking was thought to stall students' progress and lead to, in some cases, students falling further behind their peers. Using math courses once again, the partnership operationalized these patterns as conditional probabilities (i.e., conditional on a student not succeeding on an assessment, what is he or she likely to do next based on prior use of the platform?), and then scaled these analyses to include all grades and courses taught at Summit. These follow up analyses revealed that patterns referred to as *adverse transitions* were correlated with poorer performances across a range of courses, and also that these patterns declined in frequency over time, which demonstrated that students gradually stopped making these transitions.

Along with the types of transitions that students made following a low score on an assessment, we also explored students' use of learning resources across playlists. Recall that each playlist in the Summit Learning Platform is comprised of both assessments and resources, and that students are expected to use resources in order to help them pass assessments. We used an unsupervised machine learning approach referred to as hierarchical cluster analyses, combined with a heat map visualization, to explore patterns in students' resource use (e.g., Bowers, 2010). Much as with the conditional probability analyses coming out of the data sprint, an important next step following the resource-use heat map analyses was scaling the visualization to include all grades and subjects. Key to making analyses useful to practitioners and avoiding over generalizing a finding, we explored within-courses patterns of resource use in order to control for variations in content and the developmental differences of students across grades. Taking an analysis to scale meant examining whether a pattern identified in a handful of courses appeared in other courses. The ability to run analyses on one course and then on all courses proved to be an important value that the research team brought to the overall partnership.

Our partnership with Summit highlighted the ways in which researchers and practitioners can come together to jointly analyze and take action on data from digital learning environments. The design-based nature of the project, which was organized around bringing researchers closer to practice and practitioners closer to research, surfaced multiple factors associated with data-intensive partnerships and helped in clarifying the multiple steps that can be involved in using large, complex datasets to improve instruction.

Case 2: Measuring Productive Persistence to Help Faculty and Students

The second of two cases described in this chapter that was central to the development of collaborative data-intensive improvement as an approach phases and conditions was with the Carnegie Foundation for the Advancement of Teaching (Carnegie) and the Carnegie Math Pathways. At the start of our work together, Carnegie was well into launching and supporting the Pathways, which is a national effort focused on improving developmental, or remedial, math courses in two- and four-year colleges. At many colleges, these courses are significant barriers to a student's college completion (Bailey, Jeong & Cho, 2010). To help more students get past the hurdle of developmental mathematics, Carnegie brought together researchers and practitioners around the tools and routines of improvement science (Langley, Moen, Nolan, Nolan, Norman, & Provost, 2009). As the "hub" of a developing group researchers and community colleges, Carnegie formed a networked improvement community (NIC) in an effort to accelerate learning and improvement among network members (Bryk, Gomez, Grunow, & LeMahieu, 2015). Members of the Carnegie Math Pathways NIC designed two different course sequences geared toward

helping students fulfill their developmental math requirements as well as earn college credit in either statistics (i.e., "Statway") or quantitative reasoning (i.e., "Quantway").

The success of both Statway and Quantway are well documented (e.g., Yamada, 2017; Yamada, Bohannon, & Grunow, 2016; Yamada & Bryk, 2016). Key to the success of the Carnegie Math Pathways NIC is a systemic approach supported by the use of improvement tools and routines. One component of Carnegie's systemic approach is a focus on "noncognitive" factors that affect student success (see Yeager & Walton, 2011; Zimmerman, 2002). Many of these factors center on students persisting through failure and using good learning strategies, which Carnegie defines as "productive persistence." At the beginning of our partnership, Carnegie wanted to explore how data from various online learning systems used in the Pathways could be leveraged in measuring and supporting students' academic tenacity and use of effective learning strategies (see Krumm et al., 2016).

We began working with data from Statway's online learning system at the time, which was the Online Learning Initiative (OLI) platform. The platform collected information on each page that a student viewed as part of the Statway curriculum, when the page was viewed as well as information on a variety of assessments housed within the system. Through the platform, Statway provided students with practice assessments that students could use to test their own knowledge embedded within the material that they were reading. Each item that was attempted on an assessment, when it was attempted, and whether an item was answered correctly or not were collected and stored by the Online Learning Initiative platform. Along with page-views and practice assessments, the platform also captured time- and item-level data from assessments referred to as "Checkpoints," which are quiz-like assessments that come at the end of "topics" and "modules" that make up the Statway curriculum.

In the fall of 2014, we started to explore the ways in which the online system was used across individual Pathways courses. One of the benefits of looking at data stemming from the Online Learning Initiative platform was that these data were collected unobtrusively at the scale of entire Pathways NIC. These data were unobtrusive in that they were gathered directly from students as they engaged in learning activities based on what was programmed to be captured by the online learning system. While large volumes of data could be collected, that did not mean that all of it would prove to be useful for understanding students' learning behaviors, strategies, or outcomes. One step involved in identifying useful data involved becoming familiar with students' experience of using the Online Learning Initiative platform, such as the ways in which students could read pages, practice material, and take formal assessments—along with the ways in which these data were collected and stored by the system.

One the first exploratory analyses that we conducted focused on the dates with which students submitted Checkpoints. We were interested in identifying how much variation there was among students within a course for when they turned in a Checkpoint as well as between courses for when, on average or modally, students completed a Checkpoint. These analyses revealed that approximately half of all Statway courses had a modal pattern that followed the intended order of Checkpoints and that individuals as well as courses that followed the intended order tended to perform better in terms of end-of-course grades. We followed up course-level analyses by further exploring students' use of the online system by focusing on the "session" as the level of analysis. A session was defined by the online environment as the time between logging into the system and logging out or being timed out of the system. We explored patterns in what students did before and after a low score on a Checkpoint (i.e., a score below 60%). A key component of productive persistence is persisting with a task after experiencing challenge or difficulty. Low scores on Checkpoints offered a unique opportunity to measure these behaviors (Krumm et al., 2016). We also explored the number of sessions a student logged per week, the number of days between each session, and the types of sessions that students logged, such as *assessment only*

sessions where students only worked on Checkpoints or *robust* sessions where students engaged in reading, practicing, and assessment activities within the same session. All of these different operationalizations helped in understanding the ways in which productive persistence played out and could be measured using data from the OLI platform.

One way in which we sought to expand the use of these measures was to put them in front of Statway faculty to better understand how well they captured students' learning strategies and behaviors as well as whether they could be used to help faculty more effectively intervene with students. In working directly with faculty, we organized *design workshops* that were geared toward jointly interpreting data products that the research team provided and co-developing data products and follow up actions, such as change ideas that faculty could implement using a data product. Design workshops were structured activities where researchers would present evidence on students' use of the online learning system and instructors would co-develop additional data products, follow-actions, or both. Over time, the partnership viewed design workshops as the location where data became actionable. Despite the sophistication of any analysis, no data product proved to be actionable in and of itself; each data product required an explicit action to be developed.

One of our first workshops was organized around developing instrumental data products related to students' productive persistence within the Online Learning Initiative platform. Outcomes from this first workshop included finding new ways to operationalize students' engagement with online learning materials over time and creating data products that captured what students did alongside how well they did. For a second design workshop, evidence for the importance of attempting and succeeding at Checkpoints had been building across multiple analyses, and the data products and change ideas that were developed during this workshop led to a focused improvement project related to students completing Checkpoints. An initial *improvement sprint* following the workshop led to demonstrable increases in students completing end-of-module Checkpoints (Meyer, Krumm, & Grunow, 2017).

Across multiple iterations, the design workshops themselves as well as the data products and change ideas that were produced to support them proved to be valuable for both researchers and practitioners. For researchers, they offered venues for learning from faculty on what they found meaningful and whether certain patterns that were identified had face validity. For practitioners, they offered an efficient touch-point for engaging in data-intensive research activities. While they offered efficiencies for practitioners, they required significant pre-work on the part of the research team both in terms of data analysis and in organizing the workshops themselves. Following up with practitioners after a workshop was key to the overall success of the workshop. Overall, these workshops were a potent strategy for translating findings from data-intensive analyses into changes in instructional practices.

Case 3: Learning Analytics for Supporting Novice Programmers

The Context of Introductory Programming in K-12 Classrooms

Policy and educational leaders see computer science (CS) and computational thinking (CT) skills (Grover & Pea, 2013, 2018; Wing, 2006) as necessary for all citizens, not only computer scientists, with a view to building a strong STEM pipeline. Such problem-solving skills are seen as necessary to succeed and innovate in a world infused with—and lives shaped by—computing and digital devices.

Most K-12 computer science courses teach programming to support learning of computational thinking ***practices*** such as logical and algorithmic thinking, decomposing problems, debugging, and use of computational thinking ***concepts*** to create solutions that can be executed by a computer. However, programming has historically been difficult for novices to learn (e.g., Pea & Kurland, 1984;

Soloway & Spohrer, 1989). This is because programming is a complex activity that involves understanding a problem as a computational task, mapping a design for the program, drawing on problems previously programmed that have a similar structure, instantiating abstract program patterns, coding the program, and then testing and debugging (Pea & Kurland, 1984). It involves not only issues of syntax of the programming environment but also the semantics of putting together computational solutions as well as strategies and pragmatics such as testing and debugging the code.

These problems persist for novices despite the emergence of block-based programming environments that provide a visual programming interface that makes it easy for novices to get started with creating programs and animations without worrying about issues of programming syntax. However, these environments do not currently aid in formative assessment of the use of computational thinking practices and disciplinary concepts of computing to aid the learning process in the context of programming. Examining programming **process** using learning analytics (LA) gives a more complete picture (Baker & Siemens, 2014). Being able to support and scaffold this process requires us to have the ability to ***detect and recognize actions*** (single or sequences of multiple actions taken together) as evidence in support for or against the use of computational thinking. Thus, students' actions need to be interpreted as they work so that formative feedback can be provided to steer learning.

Recent learning analytics work in the context of programming has included analyzing students' steps to a solution using data from digital environments such as number of actions in students' programs and number of successful and unsuccessful program compilations (Blikstein et al., 2014). The use of clustering techniques (Bouchet, Harley, Trevors, & Azevedo, 2013) has led to identifying various programmer behavior profiles, and unsupervised methods have been used to derive program-state patterns and state transitions to predict success outcomes (Berland, Martin, Benton, Petrick Smith, & Davis, 2013). Most of these techniques have involved looking for patterns in data largely from the "bottom-up" (Winne & Baker, 2013).

New *hybrid* or *blended LA* have begun to assess students' learning processes in digital learning environments for science & math that combine top-down and bottom-up approaches to better understand students' knowledge and skills. Examples include Gobert, Sao Pedro, Raziuddin, and Baker (2013), Shute and Ventura (2013), and Zapata-Rivera, Liu, et al. (2016). Many of these by combining bottom-up LA with Evidence Centered Design (ECD; Mislevy, Almond, & Lukas, 2003), a principled approach to guide assessment design for top-down, hypothesis-driven generation of *a priori* patterns about learner actions. Evidence Centered Design focuses on three related models: student (what are targeted cognitive constructs?), task (what activities allow students to demonstrate cognitive constructs?), and evidence (what data provide evidence of cognitive constructs?). It helps connect important constructs that we want to measure with observable behaviors (including patterns of learner actions). Also, importantly, evidence is obtained by deliberately putting students in situations or tasks that will elicit the needed evidence. Once semantically meaningful patterns are defined *a priori*, data mining and learning analytics techniques can be used to analyze the patterns further.

We present a theoretical framework that researchers can use to design measurement systems for programming environments for research or application. We are using this framework currently as part of a broader effort to study and detect patterns of learner behavior during programming, as a first step toward being able to provide feedback to the learner and instructor about student learning in real-time.

Exploratory Work as a Backdrop to the Evolution of a Framework

We analyzed a dataset from an assessment task designed and used in prior research (Werner, Denner, Campe, & Kawamoto, 2012). 118 females and 202 males aged 10 to 14 years completed

the 30-minute task which involved modifying existing code in the Alice programming environment (Dann, Cooper, & Pausch, 2009). Students' computer programs and Alice data logs were collected, and the programs were scored manually using a rubric for computational thinking including algorithmic thinking and abstraction. We applied Evidence Centered Design to "reverse engineer" this task into specific computational thinking concepts and skills and give evidence of what those might look like in log files. We also compared action sequences between students who scored high and low (relative to the median) to determine commonality of sequences for each group. We found sequences that were significantly more common among students with high grades and one sequence that occurred significantly more frequently for students with low grades. Our analysis showed ***positive correlations among higher grades, number of code edit actions, and number of testing events.***

Through our exploratory work, we gained insights into interpreting student actions from logs. However, we also discovered that tasks need to be complex enough to yield rich process data logs as students apply more strategic computational thinking skills for *better coverage of focal constructs*. Measuring learning through automated means requires evidence of appropriate as well as repeated use of constructs (Koedinger, Corbett, & Perfetti, 2012). Lastly, it became apparent that without additional measures for ground-truthing or mapping the sequences back to specific instances in students' programming progressions that we have evidence for, one cannot validly interpret such sequences.

A Framework for Blending Hypothesis- and Data-Driven Learning Analytics

Building on the learning from our exploratory work, we designed a framework, or process, that employs Evidence Centered Design in its typical forward-design application, beginning from important focal knowledge and skills, and proceeding to task implementation. This approach yields an overall methodology for combining top-down Evidence Centered Design-like approaches to assessment development and delivery with bottom-up, data-driven LA approaches.

The framework (Figure 9.1) describes an iterative process that begins with identifying important computational thinking concepts and practices that we would like to measure. Careful design of tasks put students in situations that evoke behaviors to provide potential observables of these concepts and practices. Detailed analysis of program code from different solutions reveals students' use of constructs (correct or otherwise) and varied approaches to solutions. Similarly, analyzing data from screen recording and/or in-person "over-the-shoulder" observations reveals aspects of students' actions that are never seen in the final program. These can reveal student misunderstanding of concepts even if the final solution seemingly demonstrates appropriate usage of constructs. Combined qualitative analyses of the program solutions along with data-driven examination of programming process of a designed task together provide a deeper understanding of students' actions than is possible from data-driven analytics alone, including potential code sequences that map to practices that are identified through data logs. These *a priori* patterns lay the foundation for detectors for these patterns and provide a richer interpretation of student process in programming environments.

Applying the Framework

Guided by the framework, we applied Evidence Centered Design for the design of programming tasks to generate richer ***process*** data to observe repeated use of constructs and computational thinking practices. Two such tasks were piloted in two high school introductory computer science classrooms with 27 and 28 students. Data included final Alice files and log data for all

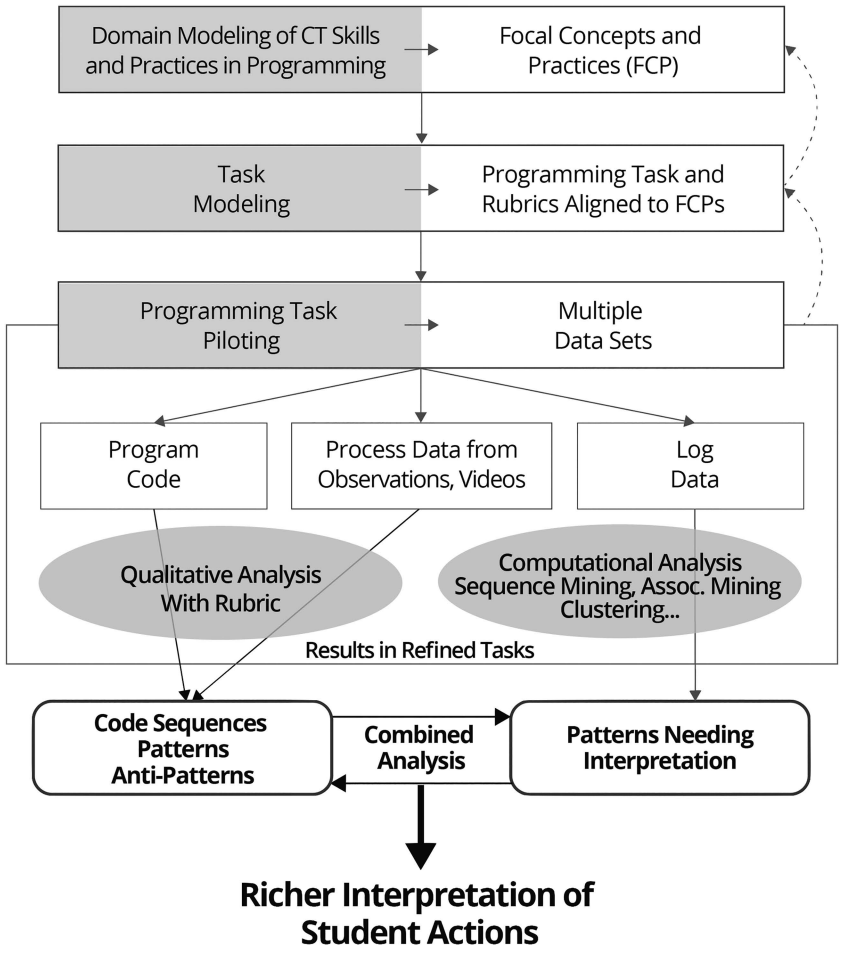


Figure 9.1 A Framework for hypothesis-driven analyses to support data-driven analytics. (Grover at al., 2017)

students and screen recordings for six students. Analysis of logs revealed similar issues that students struggled with in both tasks, for example, hard-wired vs. general solutions, improper termination conditions, decisions pertaining to parallel vs. sequential execution, effective solution decomposition, and (in-)appropriate random number use.

Analyses of screen captures from the six students using a "process over product" lens to assess computational thinking practices suggested that some students demonstrated abstraction, modularization, and testing in parts while others did not. Such observations will serve as useful patterns to search for in students' log data as evidence for computational thinking skills. In addition, we noticed certain phases during students' programming process when a student was unable to progress. Such situations can easily lead to frustration and loss of engagement and can thus serve as good candidates for potential patterns to be detected as students work on their assessment tasks. Detection of such flailing behaviors in real-time can help a teacher identify when to help a student.

Task piloting and analysis led to more refined tasks that were then used in three high school classrooms in the Western US with a total of close to 100 students. Data collection also included

screen recordings and interviews with three students in each classroom. The screen recordings are being used to validate program snapshots created from log data, in addition to aiding student recall of process during one-on-one interviews with each student. These interviews are also being used to ascertain the nature and timing of help that students may have liked to support their work. This will help us understand the nature of formative feedback and supports that can scaffold learning for students during programming.

Case 4: Learning Analytics Enabled Formative Assessment and Changes in Teacher's Instructional Practices

Homework is already required in schools and a meaningful amount of instructional time is allocated to homework (Fairman, Porter, & Fisher, 2015; Loveless, 2014). But it is also controversial and perceived as needing improvement (Kohn, 2006; Bennett & Kalish, 2006; Trautwein & Koller, 2003). Online homework tools can provide immediate feedback to students and real-time information for teachers to monitor student progress. In this section, we focus on how teachers' homework review practices change when they have access to data on student homework performance and the role of such formative assessment data for informing teachers' instructional decisions and adaptations.

Formative Assessment and Data Use in School

The concept of formative assessment has received much attention in K-12 research and practitioner communities (Black & William, 1998a, 1998b; Boston, 2002; Heritage & Popham, 2013; Roediger & Karpicke, 2006). Researchers and practitioners characterize formative assessment as a process that uses student data to inform adaptive changes in instruction (Bennett, 2011; Brookhart, 2007; Guskey, 2007; Heritage & Popham, 2013). The growing interest in formative assessment is, in part, an outcome of the general dissatisfaction with the quality of information obtained from summative assessments that generally do not provide sufficiently fine-grained or timely feedback on student learning (McMillan, 2007; Wiliam, 2016). Research documents modest to medium effect sizes of formative assessment on student learning (Black & Wiliam, 2009; Brookhart, 2007; Guskey, 2007; Hattie, 2009; Kingston & Nash, 2012; Shavelson, 2008; Speece, Molloy, & Case, 2003; Thum, Tarasawa, Hegedus, Yun, & Bowe, 2015) for a variety of different modes, grade levels, content areas, and cultural settings. Frequent use of formative assessments can improve achievement, particularly when the results are used to adjust instruction (Bergan, Sladeczek, Schwarz, & Smith, 1991; Speece, Molloy, & Case, 2003).

In recent years, administrator and teacher use of student assessment data to inform instructional decisions has been at the forefront of instructional reforms (Means, Padilla, DeBarger, & Bakia, 2009). Advocates of these reforms emphasize that teaching should be responsive to student needs and assessment data is essential to enable teachers to adjust instruction to better support individual learners. The expectation is that teachers skilled in data use will develop more effective classroom and instructional practices (Mandinach & Gummer, 2016).

The ASSISTments Online Homework Support Tool

ASSISTments is a web-based platform that provides support to students as they solve mathematics problems and provides detailed student-level and class-level formative assessment data to teachers to help inform adjustments in classroom instruction and pacing (Heffernan & Heffernan, 2014). Prior small-scaled studies showing the promise of ASSISTments have been synthesized (Rittle-Johnson & Jordan, 2016). In a recent efficacy trial funded by the IES, SRI

recruited 46 middle schools from Maine, including 87 teachers and over 2,800 seventh-grade students. In the study schools were randomly assigned to the treatment or control condition for two years. Schools assigned to the control condition continued with their homework practices as they normally would. For the schools assigned to the treatment condition, in the first year, teachers received professional development and practiced using ASSISTments with their seventh-grade classes. In the second year, these teachers continued to use ASSISTments with a new cohort of seventh-grade students, who were the student-level study population. The TerraNova Common Core mathematics assessment was administered to students to measure end-of-year outcomes. Using a hierarchical linear model (HLM), we analyzed student outcomes by condition. The adjusted mean scores on the TerraNova were 8.84 points higher in the treatment condition, and this result was statistically significant (effect size $g = .18$, $p = .007$) (Roschelle, Feng, Murphy, & Mason, 2016). According to published technical norms (CTB/McGraw-Hill, 2012) that relate TerraNova scale scores to grade level equivalents, the degree of improvement corresponds to what would be expected from .5 to one years of additional learning time.

A key component of ASSISTments is the easy-to-use online reports. The system analyses use log of all students and generates Item Report that shows data for each student on every problem and each math skill covered in the assignment, which questions and/or skills were particularly challenging, and what the common wrong answers were. The data allows teachers to make real-time, informed decisions about what to teach next, and it is ideally used to guide homework review. Figure 9.2 shows an item report with the results of six items. Teachers can see the percent correct per problem and use that data to identify weaknesses for the class. The common wrong answers support cognitive diagnosis of misconceptions. Problem numbers appear across the top row, class-level results appear in the next rows, and individual student results appear on anonymized rows below. Each cell shows what the student entered first. The cell will be yellow if the student had to be shown the answer.

Research has found that teachers do not typically know how to use data to inform instruction (Mandinach & Gummar, 2016; Means et al., 2009) or they could make errors when trying to make sense of score report results (Zapata-Rivera, Zwick, & Vezzu, 2016). Pape et al. (2013), however, found that when both professional development and formative assessment technology are provided, teachers can learn more about their students and adapt instruction with resulting improvements in student outcomes. In the ASSISTments model, teachers received a total of five days of professional development. The professional development entails discussions of the foundational instructional and learning theories behind ASSISTments as well as practical issues associated with the use of ASSISTments. Teachers learn how to use the system and how to interpret ASSISTments reports. They also receive advice on practical instructional strategies for responding to students with different needs. The later sessions focused on helping teachers sharpen their ability to adjust instruction in response to information in the reports and refine their class routines.

Impact of Using ASSISTments Reports on Teacher Practices

During the Maine efficacy study, extensive data was collected through system use log, interviews with teachers and school principals, teacher observations, surveys, and logs, and system use data generated within ASSISTments. All the data focus on the implementation process leading to outcomes. We analyzed and triangulated data from different sources to see whether teachers used data from ASSISTments reports as a part of the implementation of the intervention and whether teacher's instructional practices changed given the availability of the formative assessment data.

Based on the teacher use log data, we measured the proportion of ASSISTments reports that a teacher opened at least once. Opening a report is an important indicator of whether

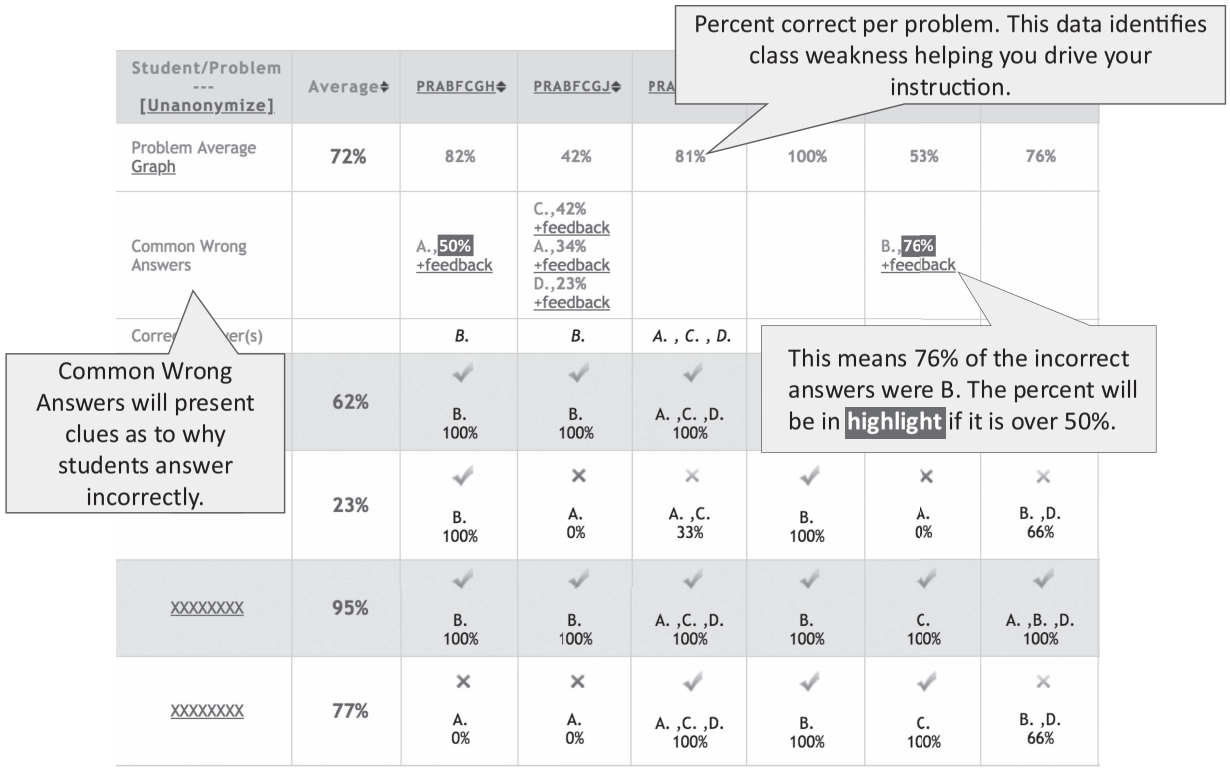


Figure 9.2 An item report for teachers showing the results of six items.
(Neil & Cristina Heffernan)

the teacher is using ASSISTments to review student work and is a precursor to using ASSISTments to adapt instruction. Across classrooms, the median for report-opening was 64%, which is above the expected opening rate (50%).

In the instructional logs and surveys, teachers were asked whether they reviewed all homework problems, they asked students which problems to review, or they reviewed selected problems based on student's performance (aka. data-driven targeted review). We found that the intervention had statistically significant effects on homework review practices. When the continuous variable (based on teacher logs) was used as the outcome measure, the effect size was 1.23, p=0.005 and for the dichotomous variable (based on the teacher survey), the odds-ratio was 45.8,[1] p=0.001. Among all 38 treatment teachers who responded to the survey, 37 reported doing targeted homework review. While in the control condition, 12 of the 36 teachers reported that they didn't do targeted homework review.

Analysis of interview transcripts and classroom observations data also provided convergent evidence of shifts in teaching practices. These shifts centered around three areas:

1. Targeted in-class review of homework problems and concepts based on needs of students. Compared to those in the control condition, treatment teachers were more likely to focus their homework review ($p < .01$) to cover fewer number of homework problems but in more depth. Teachers stated that the item report provided a starting point for their instructional planning; they reviewed the item report to quickly identify problems where a majority of students struggled and the common wrong answers and purposefully select which concepts they needed to review during the class. In contrary, control teachers relied on students' willingness to ask for help on certain problems, random or sequential selection of homework problems for review, recitation of correct answers of all homework problems but not demonstrating or discussing solution procedures, or projection of answers for students to self-correct.
2. Use of data from homework to initiate and motivate homework discussion. Presenting reports engages students directly with data on the homework results and reduces students' reluctance to ask for help on problems, as they could see that other students struggled with some of the same problems. This helped to create a safe classroom environment where students were more willing to speak up and engaged in the discussion of homework.
3. Use of homework data to inform instructional decisions during subsequent lessons. Treatment teachers acknowledged that they used the data from ASSISTments to inform instructional decisions broadly. These decisions included: instructional pacing, what concepts they needed to address during subsequent lessons, and/or which students to provide more instructional support. Treatment teachers viewed the ASSISTments reports as a valuable resource for understanding how students performed on the homework generally but more specifically how well students understood or struggled with certain concepts and procedures.

Conclusion and Discussion

In this chapter, we presented four cases that demonstrate how learning analytics can be used to improve learning environments across different grade levels and subjects. While data from digital learning environments, administrative data systems, as well as sensors and recording devices can be used to support instructional improvement, it is important to recognize that these improvements are as much about the supporting work of researchers and practitioners as they are about the data themselves—data is not a self-activating resource as it requires teams

of individuals to interpret, derive implications, and develop change ideas. Across the four cases described in this chapter, researchers and practitioners working in collaboration as well as the use of approaches such as Evidence Centered Design can provide structures and activities for translating data into an instructional change.

Key to the types of data addressed across the four cases is that they originated from processes that preceded a valued outcome, such as an end-of-course grade or standardized test performance. While these data can be collected from activities over time, they still need to be reliable, valid measures of those processes. One challenge to creating valid measures is that the technology from which the data are being collected may not collect all of the relevant data (Krumm, Means, & Bienkowski, 2018). A great deal of work and energy can go into analyzing these data, all the while critical instructional activities are occurring outside of the learning system. Working directly with practitioners can help in better understanding the instructional context in which technologies are used as well as in making more informed interpretations of the events that are captured by a technology. Moreover, approaches such as Evidence Centered Design provide a framework for interpreting available data and in developing an evidence-based argument around what processes are being measured.

Assessment in online learning has been studied for a number of years, but only recently, researchers have begun promoting and advocating the use of learning analytics for assessing academic progress, predicting future performance, and spotting potential problematic issues (Johnson, Smith, Willis, Levine, & Haywood, 2011, p. 28). The Gordon Commission (2013) recommends "separate responsibility for the use of data drawn from rich descriptions of these transactions for administrative and for student development purposes. Teachers would be enabled to interpret these data diagnostically and prescriptively" (p. 15). When using learning analytics for assessment, researchers are urged to differentiate assessment of learning (e.g. summative assessment) versus assessment for learning (e.g. formative assessment, diagnostic assessment). When the purpose of the assessment differs, the design of the learning task's focal knowledge, its features and timing (e.g. when learning is still happening vs. when learning has completed), its alignment with learning standards, potential observations, and inferences from the tasks shall be adjusted accordingly. Learning analytics can be a powerful tool for formative assessment, and for instructors to take corrective measures and monitor progress. Data collected through learning environments tends to be rich, multi-dimensional, longitudinal, embedded, and importantly—inexpensive. Such data can provide opportunities for assessing learners at a much finer-grained scale than a traditional exam; we can not only score an answer entered by a learner right or wrong, but also look at characteristics of how learners answer the question, such as how long it took them to answer, or whether their mouse hovered over a wrong answer for a while, to gauge the level of performance and confidence. On the other hand, such data can also be noisy as compared to data collected from more controlled testing environments. While there are promising applications as shown in case 3, strong evidences are warranted with regard to reliability and validity of the measures produced by learning analytics (Tannenbaum, this volume).

Acknowledgements

The research reported in this chapter was supported by the Institute of Educational Sciences, US Department of Education (R305A120125), National Science Foundation (SMA-1338487, DRL-1444621, and DRL-1418332), the William and Flora Hewlett Foundation, and SRI International. The opinions expressed are those of the authors and do not represent views of the funders. We would like to acknowledge the contributions of Marie Bienkowski, Barbara Means, Jeremy Roschelle, Neil Heffernan, Janet Fairman, Satabdi Basu, John Stamper, Michael Eagle,

and Nicholas Diana in the work presented here. We would also like to thank Jill Denner and Linda Werner for their Alice dataset that informed the research.

Note

1 We noticed that the odds-ratio for the dichotomous mediator is very big. This was possibly due to the lack of variability in the mediator for the treatment condition and a big contrast between conditions.

References

Bailey, T., Jeong, D. W., & Cho, S-W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, *29*(2), 255–270. Retrieved from https://doi.org/10.1016/j.econedurev.2009.09.002

Baker, R. S. J. D., & Siemens, G. (2014). Educational data mining and learning analytics. In Sawyer, K. (Ed.), *Cambridge handbook of the learning sciences* (2nd ed., pp. 253–274). New York, NY: Cambridge University Press.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences*, *13*(1), 1–14.

Bell, P. (2004). On the theoretical breadth of design-based research in education. *Educational Psychologist*, *39*(4), 243–253.

Bennett, B., & Provost, L. (2015). What's your theory? Driver diagram serves as tool for building and testing theories for improvement. *Quality Progress*, 36–43.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*, 5–25.

Bennett, S., & Kalish, N. (2006). *The case against homework: How homework is hurting our children and what we can do about it*. New York, NY: Crown Publishers.

Bergan, J. R., Sladeczek, I. E., Schwarz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on kindergarteners' cognitive development and educational programming. *American Educational Research Journal*, *28*(3), 683–714.

Berland, M., Martin, T., Benton, T., Petrick Smith, C., & Davis, D. (2013). Using learning analytics to understand the learning pathways of novice programmers. *Journal of the Learning Sciences*, *22*(4), 564–599.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*, 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–149.

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment: Educational Assessment. *Evaluation and Accountability*, *21*(1) (2009), 5–31.

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., & Koller, D. (2014). Programming pluralism: Using learning analytics to detect patterns in the learning of computer programming. *Journal of the Learning Sciences*, *23*(4), 561–599.

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, *8*(9).

Bouchet, F., Harley, J., Trevors, G., & Azevedo, R. (2013, April). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *Journal of Educational Data Mining*, *5*(1).

Boudett, K. P., City, E. A., & Murnane, R. J. (2013). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning* (5th ed.). Cambridge, MA: Harvard Education Press.

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment Research and Evaluation*, *15*, 1–18.

Brookhart, S. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43–62). New York, NY: Teachers College Press.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13.

Cobb, P., & Jackson, K. (2012). Analyzing educational policies: A learning design perspective. *Journal of the Learning Sciences*, *21*(4), 487–521.

Coburn, C. E., Penuel, W. R., & Geil, K. E. (2013). *Research-practice partnerships*. New York, NY: William T. Grant Foundation.

CTB/McGraw-Hill. (2012). *TerraNova* (3rd ed.). Spring Norms Book Spring, 2011 Norms. Monterey, CA: CTB/McGraw-Hill.

Dann, W., Cooper, S., & Pausch, R. (2009). *Learning to program with Alice* (2nd ed.) Upper Saddle River, NJ: Prentice Hall Inc.

Fairman, J., Porter, M., & Fisher, S. (2015). Principals discuss early implementation of the ASSISTments online homework tutor for mathematics: ASSISTments efficacy study report 2. Menlo Park, CA: SRI International.

Gobert, J., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics for science inquiry using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563.

The Gordon Commission. (2013). *To assess, to teach, to learn: A vision for the future of assessment*. Princeton, NJ: Author. Retrieved from www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Grover, S., Basu, S., Bienkowski, M., Eagle, M., Diana, N., & Stamper, J. (2017). A framework for using hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming environments. *ACM Transactions on Computing Education (TOCE)*, *17*(3), 14.

Grover, S., & Pea, R. D. (2013). Computational thinking in K—12: A review of the state of the field. *Educational Researcher*, *42*(1), 38–43.

Grover, S., & Pea, R. (2018). Computational thinking: A competency whose time has come. In S. Sentance, Carsten, S., & Barendsen, E. (Eds), *Computer Science Education: Perspectives on teaching and learning*. London: Bloomsbury.

Guskey, T. R. (2007). Formative classroom assessment and Benjamin S. Bloom: Theory, research, and practice. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 63–78). New York, NY: Teachers College Press.

Hattie, J. (2009). *Visible thinking: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Heffernan, N., & Heffernan, C. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497.

Heritage, M., & Popham, W. J. (2013). *Formative assessment in practice: A process of inquiry and action*. Cambridge, MA: Harvard Education Press.

Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K. (2011). *The 2011 horizon report*. Austin, TX: New Media Consortium.

Kingston, N., & Nash, B. (2012). How many formative assessment angels can dance on the head of a meta-analytic pin: .2. *Educational Measurement: Issues and Practice*, *31*(4), 18–19.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30–43.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, *36*(5), 757–798.

Kohn, A. (2006). *The homework myth: Why our kids get too much of a bad thing*. Cambridge, MA: Da Capo Press.

Krumm, A. E. (2017). *Collaborative data-intensive improvement: Cases and how-to guide*. Menlo Park, CA: The William and Flora Hewlett Foundation.

Krumm, A. E., Beattie, R., Takahashi, S., D'Angelo, C., Feng, M., & Cheng, B. (2016). Practical measurement and productive persistence: Strategies for using digital learning system data to drive improvement. *Journal of Learning Analytics*, *3*(2), 116–138.

Krumm, A. E., Means, B., & Bienkowski, M. (2018). *Learning analytics goes to school: A collaborative approach to improving education*. New York, NY: Routledge.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance*. New York, NY: Jossey-Bass.

Larmer, J., Mergendoller, J., & Boss, S. (2015). *Setting the standard for project based learning: A proven approach to rigorous classroom instruction*. Alexandria, VA: Association for Supervision & Curriculum Development.

Loveless, T. (2014). The 2014 Brown Center Report on American Education: How well are American students learning? The Brown Center on Education Policy. Washington, DC: The Brookings Institution. Retrieved on April 17, 2016, from http://www.brookings.edu/~/media/Research/Files/Reports/2014/03/18-Brown-Center-Report/2014-Brown-Center-Report_FINAL.pdf?la=en

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*, 366–376. Retrieved from https://doi.org/10.1016/j.tate.2016.07.011

McMillan, J. (2007). Formative classroom assessment: The key to improving student achievement. In J. McMillan (Ed.), *Formative classroom assessment: Theory into Practice* (Chpt. 1, pp. 1–7). New York, NY: Teachers College Press.

Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). *Implementing data-informed decision making in schools: Teacher access, supports and use*. Report prepared for U.S. Department of Education, Office of Planning, Evaluation and Policy Development. Menlo Park, CA: SRI International.

Meyer, A., Krumm, A. E., & Grunow, A. (2017, April). *Are these changes an improvement? Using data to inform the improvement of homework practices*. Paper presented at the Annual Meeting of the American Education Research Association. San Antonio, TX.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report 03–16). Princeton, NJ: Educational Testing Service.

Pape, S. J., Irving, K. E., Owens, D. T., Boscardin, C. K., Sanalan, V. A., Abrahamson, L., Kaya, S., Shin, H.S., Silver, D. (2013). Classroom connectivity in Algebra I classrooms: Results of a randomized control trial. *Effective Education*, *4*(2), 1–21.

Pea, R., & Kurland, D. (1984). On the cognitive effects of learning computer programming. *New Ideas in Psychology*, *2*, 137–168.

Penuel, W. R., & Gallagher, D. (2017). *Creating research-practice partnerships in education*. Cambridge, MA: Harvard Education Press.

Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). Designing formative assessment software with teachers: An analysis of the co-design process. *Research and Practice in Technology Enhanced Learning*, *2*(1), 51–74.

Rittle-Johnson, B., & Jordan, N. C. (2016). *Synthesis of IES-funded research on mathematics: 2002–2013* (NCER 2016–2003). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved July 26, 2016 from http://ies.ed.gov/ncer/pubs/20162003/pdf/20162003.pdf

Roediger, H. III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Roschelle, J., Feng, M., Murphy, R., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, *2*(4), 1–12. doi:10.1177/2332858416673968

Shavelson, R. J. (2008). Guest editor's introduction. *Applied Measurement in Education*, *21*(4), 293–294.

Shute, V. J. and Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press.

Siemens, G., & Long, P. (2011). *Penetrating the fog*. Retrieved from www.educause.edu/ero/article/penetrating-fog-analytics-learning-and-education

Soloway, E., & Spohrer, J. C. (1989). *Studying the novice programmer*. Hillsdale, NJ: L. Erlbaum Associates.

Speece, D. L., Molloy, D. E., & Case, L. P. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities: Research & Practice*, *18*(3), 147–156.

Thum, Y. M., Tarasawa, B., Hegedus, A., Yun, X., & Bowe, B. (2015). *Keeping learning on track: A case-study of formative assessment practice and its impact on learning in Meridian school district* (Research Report). Portland, OR: Northwest Evaluation Association.

Trautwein, U., & Koller, O. (2003). The relationship between homework and achievement – still much of a mystery. *Education Psychology Review*, *15*(2), 115–145.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, *15*, 1–47.

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, *54*(1), 3–11.

Werner, L., Denner, J., Campe, S., & Kawamoto, D. C. (2012). The fairy performance assessment: Measuring computational thinking in middle school. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education* (pp. 215–220). Raleigh, NC, February 29–March 3, 2012. New York, NY: ACM.

The White House. (2016). *Computer science for all*. Retrieved from http://bit.ly/2tcPrAj

Wickham, H., & Grolemund, G. (2017). *R for data science*. Sebastopol, CA: O'Reilly Media.

Wiliam, D. (2016). *Leadership for teacher learning: Creating a culture where all teachers improve so students succeed*. West Palm Beach, FL: Learning Sciences International.

Wing, J. A. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35.

Winne, P. H., & Baker, R. S. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *JEDM-Journal of Educational Data Mining*, *5*(1), 1–8.

Yamada, H. (2017). *Do effects of Quantway® persist in the following year? A multilevel propensity score approach to assessing student college mathematics achievement*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Yamada, H., Bohannon, A., & Grunow, A. (2016). *Assessing the effectiveness of Quantway®: A multilevel model with propensity score matching*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Yamada, H., & Bryk, A. S. (2016). Assessing the first two years' effectiveness of Statway®: A multilevel model with propensity score matching. *Community College Review*, *44*, 179–204.

Yeager, D., Bryk, A., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical measurement*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301.

Zapata-Rivera, D., Liu, L., Chen, L., Hao, J., & von Davier, A. (2016). Assessing science inquiry skills in immersive, conversation-based systems. In B. K. Daniel (Ed.), *Big data and learning analytics in higher education* (pp. 237–252). Cham, Germany: Springer International. doi:10.1007/978-3-319-06520-5_14.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*(3), 215–229. doi:10.1080/10627197.2016.1202110

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*, 64–70.

10

Evaluating Students' Interpretation of Feedback in Interactive Dashboards

Linda Corrin

Dashboards have long been used in business and engineering fields to provide users with a consolidated view of data to inform decision making. These decision makers are most often experts in their profession (for example, sales managers in business or pilots in engineering), who bring their expertise into the process of interpreting the data provided through the dashboard view. Dashboards are designed to use data to communicate information about areas that may need attention and action (Few, 2013). The rise of 'big data' across many industries has prompted new and innovative approaches to bringing together and displaying this data in ways that are meaningful and informative. With increasing amounts of data being collected about students' behaviour in learning environments, it is therefore not surprising that the idea of building dashboards to provide an overview of student progress and performance has also become popular in education, sparking a range of dashboard development for students across all stages of education.

In the educational context, learning dashboards have been defined as: 'a single display that aggregates different indicators about learner(s), learning process(es) and/or learning context(s) into one or multiple visualisations' (Schwendimann et al., 2017). While the majority of dashboards developed in education initially focused on providing information to teachers and administrators, an increasing number of student-facing dashboards are starting to emerge. For students, dashboards provide an opportunity to gain feedback on their learning activities and assessments, providing evidence to inform decisions around how they approach their study. Many universities, schools, learning management system vendors, and other educational technology companies are currently exploring innovative ways to deliver interactive dashboards to students which incorporate useful information displayed in ways that are easily interpretable by students.

However, there is an emerging concern about students' ability to interpret the data provided in dashboards in a way that is beneficial to their learning (Clow, 2013; Corrin & de Barba, 2014; Teasley, 2017). Research into student dashboards, to date, has tended to focus on measuring an increase in grade or a decrease in attrition in cohorts of students who have had access to a dashboard (Arnold & Pistilli, 2012). Other studies have sought students' opinions about what

they would like to see in a dashboard prior to design and development (Roberts, Howell, Seaman, & Gibson, 2016), or evaluated student satisfaction with dashboards once they have been implemented (Govaerts, Verbert, Duval, & Pardo, 2012). However, fewer studies have examined students' interpretation of dashboards and the actions they take as a result of exposure to this feedback in detail. Understanding the ways that students interact with and interpret data provided by interactive dashboards is vital in order to design effective dashboards that can support student learning. Consequently, the development of more sophisticated ways of evaluating students' interpretation of feedback delivered via dashboards is required. In establishing ways to improve evaluation it is wise to draw on tested and established practices from fields such as score reporting to inform the ways such evaluation is undertaken.

This chapter explores the role of interactive dashboards in educational environments and ways in which students' interpretation of feedback delivered through dashboards can be evaluated. This investigation is guided by the following questions:

1. What are the key design considerations and evaluation approaches used when developing learning analytics dashboards to provide feedback to students?
2. What lessons can be learnt from the score reporting literature that can guide the design and evaluation of learning analytics dashboards for students?

The chapter will also include two case studies of student-facing dashboards and the ways that students' interpretation of these dashboards have been evaluated. The chapter concludes with a discussion of the importance of considered design of dashboards that link representations of feedback to educational theories and the design of learning and assessment activities. The ways that support for student interpretation of dashboards can be delivered will also be discussed, including how approaches and principles from the score reporting literature can be used to guide the development of such support mechanisms.

Background

A decade ago the field of learning analytics emerged as the use of technology in education became more widespread and researchers began to recognise the value in the data automatically generated and collected by such technologies. The Society for Learning Analytics Research (SoLAR) was subsequently established in 2012 and define learning analytics as: 'the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs'. Research and development in the field has grown exponentially over the past few years to encompass a wide range of contexts, tools, frameworks, and issues. A key strength of the learning analytics community is that it brings together researchers and developers from across multiple disciplines including education, learning sciences, computer science, and psychology. This wealth of perspectives and knowledge offers great potential for the development of powerful tools and approaches to support and enhance student learning.

Amongst the wide range of learning analytics tools and techniques that have emerged, the idea of creating dashboards of data has featured prominently. This idea has appealed strongly, not only to learning analytics researchers, but also to educational institutions and educational technology vendors. The utilisation of dashboards is seen as a way to harness the huge amounts of data available from learning technologies and make this data accessible to those who can make best use of it. The majority of learning analytics dashboards currently in use primarily focus on providing data to teachers and educational administrators. A recent study of 55 learning dashboard projects found that 75% of dashboards were aimed at teachers, while 51%

were aimed at students (25% provided data for both students and teachers) (Schwendimann et al., 2017). A predominant focus of these systems has been to identify individuals or groups of students who are 'at risk' of either low performance or failure. Many of the student-facing dashboards available focus on providing students ways of seeing whether they are at risk in relation to a single task or across a course of study.

One of the earliest and most well-known examples of this form of retention-focused student dashboard is the Course Signals system which was implemented in 2007 at Purdue University (Arnold & Pistilli, 2012). This system's dashboard used a traffic light visualisation scheme to indicate a level of risk for students at different points throughout the semester (red = high risk, yellow = moderate risk, green = low risk). The colours of the traffic lights are determined by a predictive algorithm which incorporates data on student marks, interaction with the learning management system, prior academic history, and student demographics (e.g., age, residency, enrolled credits). In addition to the dashboard, the Course Signals system was designed to allow teachers to implement appropriate intervention strategies for students such as sending emails/text messages or scheduling face-to-face meetings to discuss the risk to a student's performance. Students also had the ability to click on their traffic light colour and receive a list of resources that can help them in their course. Early evaluation of the Course Signals system through surveys and focus groups found that most students (88%) reported a positive experience of interacting with the system. The evaluation of the dashboard focused on measuring changes in performance, retention and students' self-reports of motivation taken at the end of the semester (Pistilli, Arnold & Bethune, 2012). However, subsequent analysis of the data has raised concerns about some of the findings of this evaluation due to the reverse-causality effect (Caulfield, 2013). This example highlights the importance of careful evaluation design in measuring impacts of learning analytics-based systems, including dashboards.

Over time, many different forms of student-facing dashboard have been developed to address a range of different purposes. From helping students to monitor their activity and performance to providing evidence to promote self-reflection, dashboards have been built around a desire to allow students to view their own data in order to promote sense-making. Recently a number of systematic reviews of learning analytics dashboard design have been conducted on both teacher- and student-facing dashboards (Verbert et al., 2014; Yoo, Lee, Jo, & Park, 2015; Bodily & Verbert, 2017; Jivet, Scheffel, Drachsler, & Specht, 2017; Schwendimann et al., 2017). These reviews have explored the purpose, design, and evaluation of dashboards in order to provide guidance on how dashboards can be designed effectively and used to support student learning. The next section of this chapter will explore the outcomes of these reviews in relation to the first research question: What are the key design considerations and evaluation approaches used when developing learning analytics dashboards to provide feedback to students?

Designing and Evaluating Learning Dashboards

The Verbert et al. (2014) review examined 24 papers on dashboards with 14 focused on student dashboards. The review profiled the type of user actions that were represented in the dashboards including artefacts produced, social interactions, time spent on tasks, resource use and activity/assessment results. Of the 14 student-focused dashboards examined in the study, only 10 reported details about the evaluation undertaken. These evaluations focused on the perceived usefulness, usability and effectiveness (including student satisfaction) of the dashboards. It was observed that the results of these evaluations were mixed depending on the dashboard purpose and the data included in the dashboard design. While some studies had reported an increase in grades, retention and self-assessment, others showed no significant difference in these areas. It was concluded that there was limited consensus across the studies as to the most relevant data to

be included in dashboards and that more research is needed to consider what other data about learners and the learning process could be useful. It is also suggested that more longitudinal approaches to evaluating the impact of dashboards on student learning are required.

The review conducted by Yoo et al. (2015) focused on educational dashboards based on data from learning management systems. 10 dashboards were included in the review with seven of these having a student-facing component. The information presented in these dashboards included login trends, performance results, content usage, message analysis, online social networks, and at-risk student prediction. The review applied Kirkpatrick and Kirkpatrick's (2006) four level evaluation model (reaction, learning, behaviour and result) to each of these 10 dashboards to assess the evaluation conducted on each. Only six of the 10 dashboard studies were found to have addressed any of the four levels, with only one dashboard study (Upton & Kay, 2009) fully evaluating each of the four levels. Yoo et al. (2015) then went on to propose an evaluation framework for educational dashboards which brings together Kirkpatrick's four level model (Kirkpatrick & Kirkpatrick, 2006), Verbert et al.'s (2013) learning analytics process model (impact, sense-making, reflection and awareness), and Few's (2009) blocks of information visualisation (see Table 10.1).

The systematic review conducted by Schwendimann et al. (2017) incorporated studies of 55 dashboards, of which 28 were student-facing. They identified six forms of data included in the dashboard designs (activity logs, learning artefacts, self-report data, institutional databases, physical activity and external systems) and 200 individual indicators, which they categorised by how each related to the learner, action, content, results, social or context. Over half (58%) of the studies didn't contain any evaluation, but of those studies that did contain evaluation 65% involved a mixed methods approach which combined quantitative and qualitative techniques. The evaluations tended to focus on usability, usefulness and user satisfaction, with very few exploring the impact of the dashboards on learning. The review authors observed that most evaluation methods appeared to be 'low-effort, low-detail' (p. 38) and called for more comparative studies of dashboards, indicators, visualisations and impact.

The review by Bodily and Verbert (2017) broadened out the inclusion criteria to look not just at student dashboards, but any form of student-facing analytics output. The review included 94 articles covering student-facing learning analytics reporting systems designed for the purposes of awareness/reflection, recommendation of resources, improvement of retention or engagement, increasing social behaviour online, and recommendation of courses. Of the 94 systems reviewed only 29 provided any interactive elements for students and very few (12) provided

Table 10.1 Summary of the evaluation framework for educational dashboards (adapted from Yoo et al., 2015).

Criteria	Sub-categories
Reaction	Goal-orientation
	Information usefulness
	Visual effectiveness
	Appropriation of visual representation
	User friendliness
Learning	Understanding
	Reflection
Behaviour	Learning motivation
	Behavioural change
Result	Performance improvement
	Competency development

justifications for the design choices made in relation to the dashboard design. In terms of evaluation, 10 articles included some form of usability testing, 32 sought student perceptions of usability, 34 looked at usefulness and 35 asked students if they perceived a change in behaviour, achievement or skills. The review concluded with a list of recommendations for implementing reporting systems which can also be used to provide a good structure for evaluating student-facing dashboards. These include questions on intended goals, visualisation techniques, information selection, needs assessment, usability testing, visual design, student perceptions, actual effects, and student use.

The most recent published review of learning analytics dashboards was conducted by Jivet et al. (2017) and focused on how theories and models from the learning sciences have been used to inform the design of dashboards. From an initial sample of 95 papers that reported on student-facing dashboards, widgets or visualisations, the authors identified 26 of these studies that met the further criteria of being empirical and relying on educational concepts in their design. Across the included dashboards six educational concepts were identified: cognitivism, constructivism, humanism, descriptive models, instructional design, and psychology. The most common goal of student-facing dashboards was to support awareness and reflection which, along with improving metacognitive skills, monitoring progress and supporting planning, were classified as relating to metacognitive competence. The other three competences identified were cognitive, behavioural and emotional. Each of these competencies relate to the core theory of self-regulated learning and the review authors suggest that dashboards should be complemented with tools that can help students who are struggling with their self-regulation to develop their skills. This review also raised the concern of the common use of student comparisons in dashboards and suggests that using goal achievement as a standard for comparison could be a more pedagogically-sound approach than creating competition among students. Unlike the previous reviews, the Jivet et al. (2017) review didn't specifically investigate evaluation of dashboards, but did make a recommendation that evaluation should be linked to the educational concepts that inform the dashboard design.

While each of the reviews included here had a slight different focus or sample, a number of consistent themes emerged. The reviews showed that there are many different ways of designing dashboards and many different purposes for which dashboards can be used. However, the details in the literature about theoretical foundations, design considerations and data specifications were varied or, in some cases, missing. This makes it difficult to determine whether these elements were considered and just not reported, or whether they weren't part of the dashboard design process. In their review, Schwendimann et al. (2017) provide a checklist of elements that they recommend be included when reporting on dashboard project. In practice, this list can also be used as a checklist for the process of dashboard design. The checklist includes having a clear definition of learning dashboards, outlining the technologies used, the educational context, the evaluation approach (including how learning impacts were evaluated), and the resulting learner/teacher practices (Schwendimann et al., 2017). Two additions to this list can be made from suggestions from the Bodily and Verbert (2017) review, including a needs assessment prior to design and development, and a justification of the visual techniques chosen to represent the data in the dashboard.

It should also be noted that two critical challenges were raised in the dashboard reviews that need to be addressed by dashboard designers in order for dashboards to be effective in educational environments. The first of these are the ethical and privacy considerations around the use of student data to populate learning dashboards. Much has been written about the need for strong ethical frameworks to guide educational institutions on the protection of students' privacy when developing learning analytics systems (Slade & Prinsloo, 2013; Sclater, 2014). It is also imperative for the student voice to be included in discussions around the use of their data.

Recently several studies have emerged that have involved students in discussions around their willingness to share their data for the purposes of building learning analytics tool such as dashboards (Brooker, Corrin, Mirriahi, & Fisher, 2017; Roberts et al., 2016). While many schools and universities have started to implement processes to protect and make ethical use of student data, there is still some way to go in ensuring this protection is universal.

The second challenge focuses on the emphasis some dashboard designs place on allowing students to compare their engagement and/or performance with their peers. The inclusion of comparative elements, such as a class average, is common in learning dashboard design, especially in dashboard products developed by educational technology vendors. However, the literature on social comparison theory (Festinger, 1954), motivation (Pintrich, 2004), and self-regulation (Butler & Winne, 1995) suggest that these comparative elements can have different effects on different students. This is an area that requires more research to determine how this can best be approached in terms of dashboard design. This should include studies in real educational environments over time to see, not only the short term effects on students' engagement with a single task or subject, but also how this impacts the ways students approach their study going forward.

Approaches to the evaluation of the impact dashboards vary across the literature. Perhaps the most surprising outcome from the dashboard reviews was the fact that many studies either do not undertake evaluation or, if they do, do not report the evaluation outcomes in their work. Of those papers that did report evaluation findings, the main areas of evaluation focused on usability, usefulness, satisfaction, and effectiveness. While the issues of usability, usefulness, and satisfaction are all very important in ensuring that dashboards are designed well, the issue of effectiveness is key to determining whether dashboards are a good mechanism for delivering feedback to students. In the Verbert et al. (2014) study, measures of effectiveness were said to include higher levels of engagement, higher performance in assessment, increased student retention, and improvements in self-assessment. Interestingly, no clear pattern of increase in these measures was seen across the studies in the dashboard review. In fact, several studies that measured changes in engagement or performance found no significant change (e.g., Morris, Piper, Cassanego, & Winograd, 2005). So, despite the fact that students are generally happy to receive data and feedback via dashboards, there is inconsistent evidence about the impact these dashboards are having on student learning.

Evaluation of Students' Interpretation of Dashboard Feedback

Evaluation of students' interpretation of the feedback delivered through learning dashboards remains limited in the learning analytics literature. The ability to conduct this form of evaluation faces two main challenges. The first is to be able to gather data from students at the moment they interact with the dashboard, to understand how they translate what they are seeing through the data visualisations into some form of action. The second challenge is to be able to track whether the student follows through with this action and what impact this has on their learning. The ability to measure these two things often goes beyond what is currently captured in online systems and requires additional data collection, such as student self-report data.

To evaluate students' interpretation of feedback delivered through dashboards it is important to underpin the evaluation process with a strong theoretical framework. This can help to guide the evaluation design and identify the evidence required to determine how feedback was interpreted by students, and also what actions resulted from this interpretation. When evaluating dashboards, it is not always possible to measure the impact on student learning directly, as changes in learning performance can be influenced by many other factors in the educational environment. However, evaluation can be targeted at investigating the extent to which the purpose of the dashboard has been achieved, for example, the impact that a dashboard has had on

students' ability to self-regulate their learning. The following two case studies demonstrate different approaches that have been taken to evaluating students' interpretations of dashboard visualisations and the impact these interpretations have had on students' approaches to their study.

Case Study 1: Learning Analytics Visualisations for a Single Task

The first study by Beheshitha, Hatala, Gašević, and Joksimović (2016) was designed to investigate the effect of students' access to learning analytics visualisations on learning activity while controlling for the motivational construct of achievement goal orientation. Situated in an authentic learning context in higher education, students were part of an experimental condition where they were shown one of three dashboard-style visualisations of their activity in an asynchronous online discussion. The three visualisations showed the student their activity in relation to either the class average for posts, the top five contributors, or the number of key concepts included in posts. Log data was collected on the visualisation views and posted messages in the learning management system. This was supplemented with a self-report survey using Elliot, Murayama, and Pekrun's (2011) 3 x 2 achievement goal model to measure students' goal orientation. The analysis in this study involved discourse analysis of the discussion posts (using Coh-Metrix) and hierarchical linear mixed models for the statistical analysis of the variables from across the data sources.

The results of the study showed different impacts on students' activity as a result of viewing different visualisations. For example, students with interpersonal achievement goals (i.e., a preference to compare their work to the standard of others) who viewed the top contributors or key concepts visualisations subsequently went on to post more. Whereas, those who had viewed the class average visualisation posted less. In relation to the content of discussion posts, students with a self-avoidance goal orientation (i.e., a motivation to avoid doing the task worse than their own previous work) had higher levels of narrativity, deep cohesion, syntactic simplicity, and referential cohesion in their posts after viewing the key concepts visualisations, but not when viewing the class average or top contributors' visualisations.

This study highlights the complexity of designing learning dashboards for students who may have different motivations and goals for their study. Methodologically, this study demonstrated a quantitative approach to measuring changes in students' approaches to a learning task. The ability to control for different achievement goal orientations provided a more sophisticated view of the impact of learning analytics visualisations on student behaviour and engagement with a task. The authors of the study suggest that further research involving other theoretically-informed constructs could help to build a more complete picture of the impact of dashboards and visualisations on student approaches to learning.

Case Study 2: Learning Analytics Dashboard for Multiple Assessments and Tasks

In contrast to the first study, the second case study employed a more qualitative approach to exploring students' interpretation of feedback delivered via a learning analytics dashboard (Corrin & de Barba, 2014, 2015). The study used a semester-long, multi-phase, mixed methods approach to explore higher education students' interpretation of dashboard data and the actions they took in response, with reference to their self-regulated learning. The dashboard that was used as part of this study was built to replicate the common ways of displaying student data used by learning management system vendors. Although the dashboard wasn't live, it contained real data of participants' activities and performance. The 24 student participants were recruited from across two discipline areas (science and languages). At the beginning of the semester the participants were asked to fill in a survey about their personal learning goals and motivations. This was followed

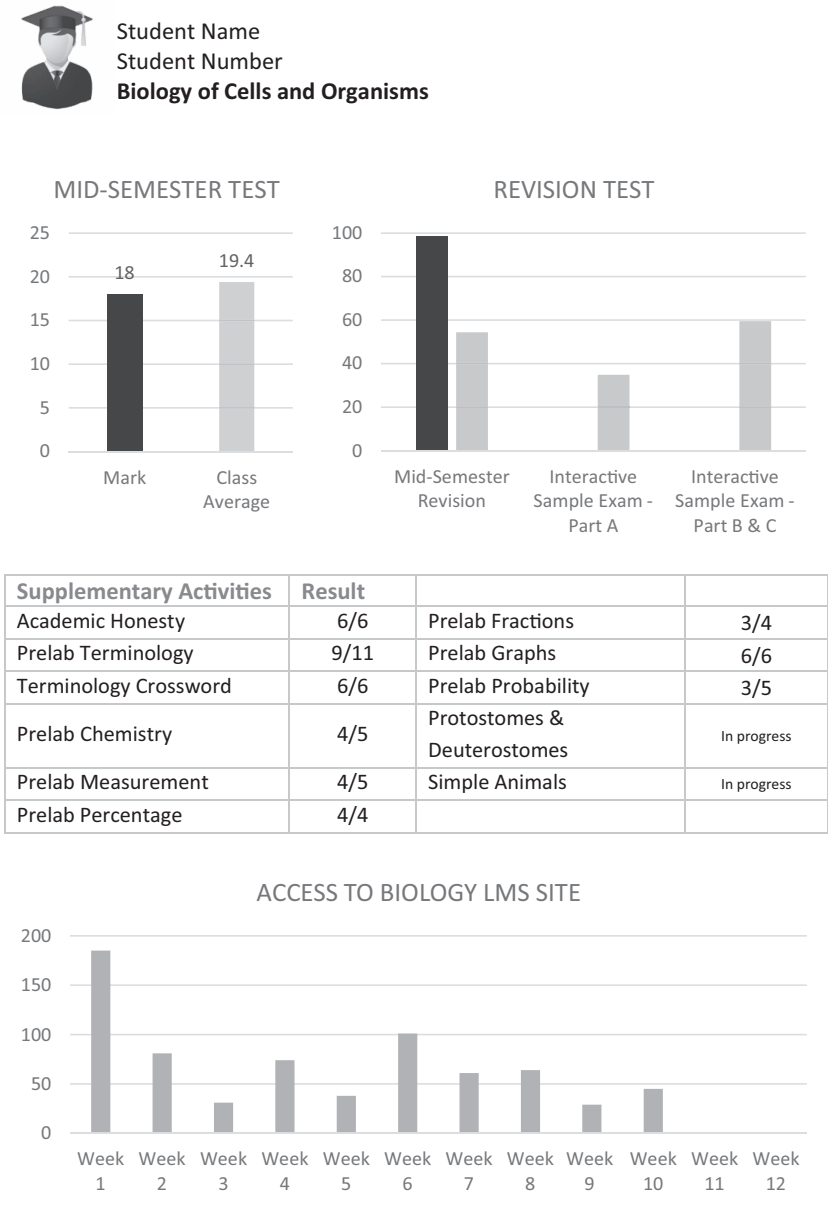
by an interview in week six of semester where students were shown their dashboard and, using a think-aloud method, asked to explain their interpretation of the data visualisations. The students were also asked to outline any actions they might take as a result of seeing this data. A similar interview then took place in week 11 where students were first asked to describe how seeing the data in the previous interview had impacted their approaches to study, before going through the same think-aloud process after seeing their updated dashboard. At the end of the semester, once they had received their final grade for the subject, students were asked to fill in another survey to reflect on the impact that having this feedback had on their study throughout the semester. The survey also included questions about the usefulness of the visualisations in the dashboard.

The outcomes of the research showed a diversity in how students interpreted the dashboard data and their ability to determine suitable actions to take in response. While the dashboard designs incorporated visualisations of both activities in the LMS and results from assessment tasks (see Figure 10.1), students focused primarily on the representation of summative assessment items. Some students were able to associate the effort they put into a learning activity or assessment with the results shown in the dashboard, while others struggled to work out what they needed to do in order to improve their performance. An aspect of the dashboard design that participants found useful was the fact that all the assessment and online learning activities were shown in one consolidated view. Effectively the dashboard acted as a map of the activities and assessments that students needed to complete throughout the semester. Students were able to use the dashboard layout and feedback provided to plan their study schedules and identify tasks that they may have missed.

For each of the assessment tasks and the learning management system access statistics a class average was given as a standard for comparison. The ability to see their activity in relation to the average had a range of impacts on students depending on their different motivations and goals. Those whose performance sat below the class average tended to feel it wasn't useful to compare their work with others. Those substantially higher than the class average reported that it was good to know that their work was of high standard, but that the average did not influence their actions going forward. Those students whose performance was close to the class average either were happy that their work was comparable with others or saw this as a motivation to try harder. Interestingly, some of these students who were happy with their average-level performance and didn't see a need to change their study strategies, had expressed a higher performance goal at the beginning of the semester. What this meant was that by seeing their average performance on the dashboard they had been distracted from their original goal. While this is only a small study with a small sample, this particular finding indicates that more research is required to determine the broad influence of these comparative standards on students' interpretation of dashboard feedback and motivation.

Both these case studies highlight that there are many individual differences in how students approach their studies that can impact the interpretation they make of feedback delivered via learning analytics dashboards or visualisations. Another strong theme that emerged from these two studies was the importance of the pedagogical design of learning tasks and assessments in how students interpret the visualised feedback. While the students may know the process, they followed to complete a learning or assessment task, this doesn't always translate into a strong understanding of the pedagogical intent behind the task, which is important in helping students to identify appropriate actions to take to improve their performance. These studies also demonstrate that in order to gain a fuller understanding of the impact of dashboards on student interpretation research needs to go beyond a single data source (e.g., the log data from a learning management system) to incorporate multiple methods of data collection.

Emerging research into student dashboards has begun to uncover many issues that dashboard designers and teachers need to take into consideration when designing and implementing



Supplementary Activities	Result		
Academic Honesty	6/6	Prelab Fractions	3/4
Prelab Terminology	9/11	Prelab Graphs	6/6
Terminology Crossword	6/6	Prelab Probability	3/5
Prelab Chemistry	4/5	Protostomes & Deuterostomes	In progress
Prelab Measurement	4/5	Simple Animals	In progress
Prelab Percentage	4/4		

Figure 10.1 Science subject dashboard from the Corrin and de Barba (2015) study.

dashboards in a way that can enhance the learning experience for students. Like any new field, there is still lots to learn and the popularity of this form of feedback provision will hopefully inspire more research in this area. It is also important to look to other fields where the use of dashboards and reporting of educational data are more established. Once such area is that of score reporting. Consideration of existing research on score reporting is rare in the learning analytics literature. The next section of this chapter explores what lessons can be learnt from the score reporting literature to guide the design and evaluation of learning analytics dashboards for students.

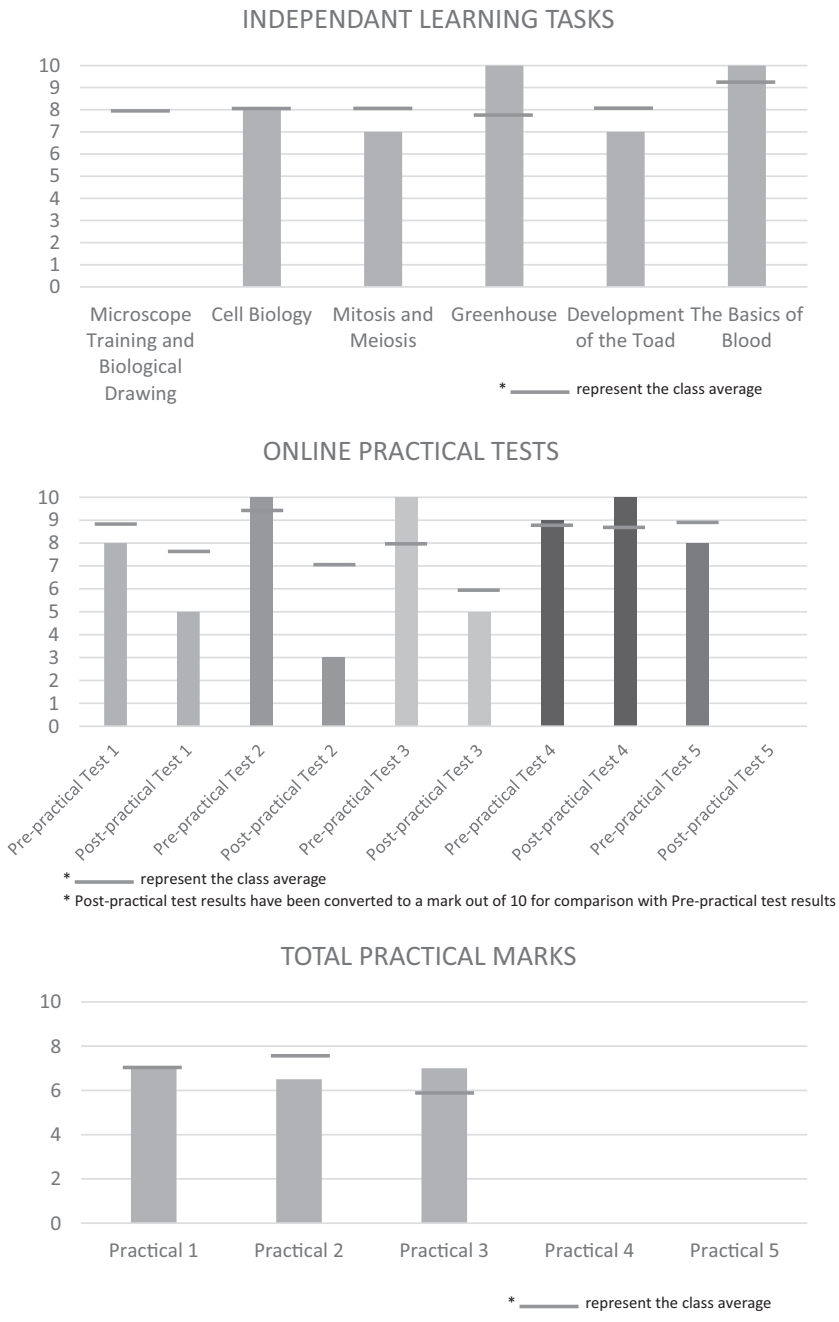


Figure 10.1 (Continued)

Lessons From the Field of Score Reporting

The field of score reporting has a long history of exploring effective ways to communicate information about student learning and the impact of the curriculum to students, teachers, parents, and educational administrators (Ryan, 2006). The design and evaluation of score reports are

often guided by national standards. For example, in the United States the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education work in partnership to produce the *Standards of Educational and Psychological Testing*. These standards address issues such as the validity, reliability, and fairness of testing and the reporting of testing scores.

Particularly relevant to the area of student dashboards is the requirement in the standards for score reports to be accompanied by supporting documentation that help the report audience to interpret the contents of the report. Across the reviews of learning analytics dashboards very little mention was made of supporting resources for interpretation of dashboard visualisations. Some dashboards were designed with the explicit purpose to be used in conversation between student and teachers or academic advisors (e.g., Aguilar, Lonn & Teasley, 2014), but many were designed to promote student self-reflection. The fact that these dashboards are commonly delivered online and accessible at any time would suggest that support for interpretation should be built into the dashboard itself (or available in a linked resource) rather than being reliant on conversations with teaching staff. While some support for interpretation can be built into the visualisation itself, information about the design of the learning and assessment activities should also be provided. As seen in the two case studies above, an understanding the pedagogical context of the data presented in dashboards is vital to the process of making an interpretation.

In relation to the validity and design of score reports, several large reviews of practice have been conducted which set forward recommendations for the design of score reports. Hattie (2009) proposes 15 principles to maximise the ability of the reader to make appropriate interpretations. These principles address the validity of score reports by suggesting that there should be minimal use of numbers and an effort made not to make the interface too cluttered. It is suggested that each report should have a theme and should be designed to answer specific questions. Among the principles are suggestions for support materials that provide a justification for the assessment design. Hattie also calls for evidence to demonstrate how audiences interpret the reports, in particular, an exploration of what the audience sees and what action they will take next. Similar themes were observed in a review by Goodman and Hambleton (2004) who also provide more specific recommendations on the visual design elements of score reports, such as the grouping of data in meaningful ways and the highlighting of main findings using boxes and graphics. The value of piloting reports is also included as a recommendation of this review. In this volume, Tannenbaum adds to the discussion around validity by setting forward strategies for alignment between assessment and score reports, design decision-points, and steps for report development. The useful perspective provided around setting the criteria for inclusion of data in a report can be used to inform how criteria could be set for the inclusion of data in a dashboard, with particular reference to whether the purpose of the dashboard is summative or formative in nature.

Further recommendations for the delivery of online static and interactive scores reports are provided by Zenisky and Hambleton (2012b). The online format provides opportunities to allow users to take more control over their exploration and interpretation of score reports through the use of links and subpages to provide more detail about the data presented. The online and interactive nature of student dashboards can also provide this opportunity, yet to date few dashboards have been built in a way that provides this functionality to students. In a recent study of student perspectives on technology-supported feedback several participants expressed a desire to be able to drill down into more detail about the concepts, competencies or skills they need to focus on to improve their assessment performance (Corrin & de Barba, 2018). In the way that subscores can provide a more detailed picture of areas in need of development in score reports, the ability to classify assessment questions in terms of areas of knowledge or skill could help students to better interpret the marks and grades they can view through dashboards so they can

determine appropriate actions to take. Yet the challenge, particularly in the more generic dashboards being delivered through learning management systems, is to get teachers to perform this classification and for adequate testing of subscores to be undertaken (Sinharay, 2010). It is also necessary for the dashboard to support the functionality of classification and the interactivity of drilling down to see this greater level of detail.

The models of evaluation used to test interpretations of score reports prior to and post release have the potential to provide useful guidance to designers of student dashboards. A research-based model outlined by Zapata-Rivera and VanWinkle (2010) involves the gathering of assessment information needs, the reconciliation of these needs with the score reporting needs, the design of a score report prototype, an internal evaluation (with experts in subject matter, usability, measurement and accessibility), and an external evaluation (with representatives of the intended audience). Iterations of these steps can be taken as many times as needed to develop the most useful score report. Similarly, Zenisky and Hambleton's (2012a) score report development model advocates field testing with potential user groups in controlled studies as well as the development of programmes of ongoing monitoring and maintenance. It is also important that the evaluation extends beyond the score reports themselves to the interpretation support materials created. For example, Zapata-Rivera, Zwick and Vezzu (2016) conducted an evaluation of the usefulness of a tutorial designed to help teachers to understand representations of measurement error in score reports. In addition to questions about usability, the participants were asked to complete a comprehension questionnaire to assess their understanding of the concepts covered. The development of similar instruments to investigate students' interpretation of dashboard visualisations, and any supporting materials could be useful for dashboard design and delivery.

Conclusion

Research into student-facing analytics feedback is still in its early days and the literature on student dashboards is currently not mature enough to be able to provide authoritative guidance on the most effective visual elements to assist student learning. However, major themes are emerging around the importance of the theoretical foundation behind the purpose of dashboards and pedagogical design of the learning activities included in the dashboard in providing feedback to students. The research has also shown that student characteristics, such as goal orientation and motivation can have a considerable influence on how dashboards are interpreted by students. Designing dashboards to address these issues presents a particular challenge to educational technology vendors who often seek to provide a 'one-size-fits-all' dashboard product to institutions. The emerging research would indicate that this approach, as Teasley (2017, p. 6) suggests, 'may be unwise'.

So, if we return to where we started, to the history of dashboards and their role in supporting decision-making in business and engineering, we see that the audience for this form of feedback are experts in their field. However, while students have experience at being students, they are not experts in education. They often lack sufficient knowledge about the pedagogical intent of learning activities, the design of assessment, and an understand of how this all fits within the broader curriculum. Additionally, they may not have adequate levels of data literacy to be able to understand the statistics and visualisations presented in dashboards and reports (MacNeill, Campbell, & Hawksey, 2014; Vezzu, VanWinkle, & Zapata-Rivera, 2012).

These challenges highlight the importance of the provision of support for the interpretation of feedback given through learning analytics dashboards. This can be done in a number of ways including the incorporation of visual elements and descriptions to accompany data representations, the provision of supporting materials, or the provision of face-to-face support from

teachers and/or academic advisors. Tactics such as the use of evidence-based or stealth assessment designs could also help to strengthen students' understanding of learning objectives and expectations of performance (Shute & Kim, 2014).

Important lessons can also be learnt from the work on creating reports for 'open learner models' which looks at ways to provide information to students about the content and skills associated with educational systems such as intelligent tutoring systems (Bull & Kay, 2016; Bull, Wasson, Johnson, Petters, & Hansen, 2012). This work demonstrates how visualisations of data can be built within the context of a learning activity design and adapted for students in ways that allow them to explore different levels of granularity, effectively creating an 'active report' (Zapata-Rivera, Hansen, Shute, Underwood, & Bauer, 2007). The value in these approaches is that visualisations are grounded within a model that clearly represents the design of the learning activity, helping to support and raise students' metacognitive awareness (Vatrapu, Teplovs, Fujita, & Bull, 2011).

In designing such support and evaluating the interpretations that students make of feedback delivered through dashboards the field of learning analytics would be wise to look to work already done in other disciplines. As has been outlined in this chapter, the field of score reporting provides useful information on ways that feedback can be visualised as well as models for the evaluation of student interpretations of this feedback. This literature and the emerging literature on evaluation of dashboards in the field of learning analytics suggests a move beyond simple measures of student satisfaction and usability towards methods which can capture students' understanding of the feedback as well as the ways that they translate these understandings into action. The proposed stages of score report development models presented here (Zapata-Rivera & VanWinkle, 2010; Zenisky & Hambleton's, 2012a) are also useful to ensure that evaluation is built in across the whole dashboard development process. There is lots of work still to be done in determining the most effective ways of designing student dashboards, but by drawing on these evaluation models as well as educational theories and learning designs, institutions and teachers will be better placed to design dashboards that provide useful feedback to support student learning.

References

Aguilar, S., Lonn, S., & Teasley, S. D. (2014). Perceptions and use of an early warning system during a higher education transition program. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 113–117). New York, NY: ACM.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics & knowledge* (pp. 267–270). New York, NY: ACM.

Beheshitha, S. S., Hatala, M., Gašević, D., & Joksimović, S. (2016). The role of achievement goal orientations when studying effect of learning analytics visualizations. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 54–63). New York, NY: ACM.

Bodily, R., & Verbert, K. (2017). Trends and issues in student-facing learning analytics reporting systems research. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 309–318). New York, NY: ACM.

Brooker, A., Corrin, L., Mirriahi, N., & Fisher, J. (2017). Defining 'data' in conversations with students about the ethical use of learning analytics. In H. Partridge, K. Davis, & J. Thomas (Eds.), *Me, us, IT! Proceedings ASCILITE2017: 34th international conference on innovation, practice and research in the use of educational technologies in tertiary education* (pp. 27–31). Toowoomba: ASCILITE.

Bull, S., & Kay, J. (2016). SMILI☺: A framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, *26*(1), 293–331.

Bull, S., Wasson, B., Johnson, M. D., Petters, D., & Hansen, C. (2012). Helping teachers effectively support group learning. In J. Kim & R. Kumar (Eds.), *Proceedings of workshop on intelligent support for learning in groups*, *ITS2012*. Berlin Heidelberg: Springer.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Education Research*, *65*(3), 245–281.

Caulfield, M. (2013, November 12). Purdue course signals data issues explainer. *e-Literate blog*. Retrieved from https://mfeldstein.com/purdue-course-signals-data-issue-explainer/

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683–695.

Corrin, L., & de Barba, P. (2014). Exploring students' interpretation of feedback delivered through learning analytics dashboards. In B. Hegarty, J. McDonald, & S-K. Loke (Eds.), *Rhetoric and reality: Critical perspectives on educational technology. Proceedings ascilite Dunedin 2014* (pp. 629–633). Dunedin: ASCILITE.

Corrin, L., & de Barba, P. (2015). How do students interpret feedback delivered via dashboards? Poster abstract available in P. Blikstein, A. Merceron, & G. Siemens (Eds.), *Proceedings of the 5th international conference on learning analytics and knowledge* (pp. 430–431). New York, NY: ACM.

Corrin, L., & de Barba, P. (2018). *Determining students' assessment feedback preferences for personal analytics solutions*. Sydney: Department of Education and Training.

Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3× 2 achievement goal model. *Journal of Educational Psychology*, *103*(3), 632.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140.

Few, S. (2009). *Now you see it: Simple visualisation techniques for quantitative analysis*. Burlingame, CA: Analytics Press.

Few, S. (2013). *Information dashboardesign: Displaying data for at-a-glance monitoring* (2nd ed.). Burlingame, CA: Analytics Press.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145–220.

Govaerts, S., Verbert, K., Duval, E., & Pardo, A. (2012). The student activity meter for awareness and self-reflection. In *CHI'12 extended abstracts on human factors in computing systems* (pp. 869–884). New York, NY: ACM.

Hattie, J. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*, 1–15. Retrieved from www.oerj.org/View?action=viewPaper&paper=6

Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice. In *European conference on technology enhanced learning* (pp. 82–96). Cham, Switzerland: Springer.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). San Francisco: Berrett-Koehler.

MacNeill, S., Campbell, L., & Hawksey, M. (2014). Analytics for education. *Journal of Interactive Media in Education*. Retrieved from www-jime.open.ac.uk/jime/article/view/2014-07

Morris, M. R., Piper, A. M., Cassanego, A., & Winograd, T. (2005). *Supporting cooperative language learning: Issues in interface design for an interactive table* (Research Report, 8). Retrieved from http://inclusive.northwestern.edu/languagelearning.pdf

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385–407.

Pistilli, M. D., Arnold, K., & Bethune, M. (2012). *Signals: Using academic analytics to promote student success*. EDUCAUSE Review Online. Retrieved from https://er.educause.edu/articles/2012/7/signals-using-academic-analytics-to-promote-student-success

Roberts, L. D., Howell, J. A., Seaman, K., & Gibson, D. C. (2016). Student attitudes toward learning analytics in higher education: 'The Fitbit Version of the Learning World'. *Frontiers in Psychology*, *7*, 1959. Retrieved from http://doi.org/10.3389/fpsyg.2016.01959

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. *Handbook of Test Development*, 677–710.

Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., . . . Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41.

Sclater, N. (2014). *Code of practice for learning analytics: A literature review of the ethical and legal issues* (pp. 1–64). Jisc. Retrieved from http://repository.jisc.ac.uk/5661/1/Learning_Analytics_A-_Literature_Review.pdf.

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In *Handbook of research on educational communications and technology* (pp. 311–321). New York, NY: Springer.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150–174.

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1510–1529.

Teasley, S. D. (2017). Student facing dashboards: One size fits all? *Technology, Knowledge and Learning*, *22*(3), 377–384.

Upton, K., & Kay, J. (2009). Narcissus: Group and individual models to support small group work. *User Modeling, Adaptation, and Personalization*, 54–65.

Vatrapu, R., Teplovs, C., Fujita, N., & Bull, S. (2011, February). Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 93–98). New York, NY: ACM.

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, *57*(10), 1500–1509.

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing*, *18*(6), 1499–1514.

Vezzu, M., VanWinkle, W., & Zapata-Rivera, D. (2012). *Designing and evaluating an interactive score report for students* (Research Memorandum No. RM—12–01). Princeton, NJ: Educational Testing Service.

Yoo, Y., Lee, H., Jo, I. H., & Park, Y. (2015). Educational dashboards for smart learning: Review of case studies. In *Emerging issues in smart learning* (pp. 145–155). Berlin, Heidelberg: Springer.

Zapata-Rivera, D., Hansen, E., Shute, V. J., Underwood, J. S., & Bauer, M. (2007). Evidence-based approach to interacting with open student models. *International Journal of Artificial Intelligence in Education*, *17*(3), 273–303.

Zapata-Rivera, D., & VanWinkle, W. (2010). *A research-based approach to designing and evaluating score reports for teachers* (Research Memorandum 10–01). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*(3), 215–229.

Zenisky, A. L., & Hambleton, R. K. (2012a). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26.

Zenisky, A. L., & Hambleton, R. K. (2012b). From 'Here's the Story' to 'You're in Charge': Development and maintaining large-scale online test and score reporting resources. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Handbook of large-scale assessment* (pp. 175–184). London: Taylor & Francis.

Index

Note: Page numbers in *italics* indicate figures and in **bold** indicate tables on the corresponding pages.