

Stephen N. Elliott
Ryan J. Kettler
Peter A. Beddow
Alexander Kurz
Editors

Handbook of
**Accessible Achievement
Tests for All Students**

Bridging the Gaps Between Research, Practice, and Policy

Handbook of Accessible Achievement Tests for All Students

Stephen N. Elliott · Ryan J. Kettler ·
Peter A. Beddow · Alexander Kurz
Editors

Handbook of Accessible Achievement Tests for All Students

Bridging the Gaps Between
Research, Practice, and Policy

 Springer

Editors

Stephen N. Elliott
Learning Sciences Institute
Arizona State University
Tempe, AZ, USA
steve_elliott@asu.edu

Ryan J. Kettler
Department of Special Education
Vanderbilt University
Nashville, TN, USA
ryan.j.kettler@vanderbilt.edu

Peter A. Beddow
Department of Special Education
Vanderbilt University
Nashville, TN, USA
peterbeddow@gmail.com

Alexander Kurz
Department of Special Education
Vanderbilt University
Nashville, TN, USA
alexander.kurz@vanderbilt.edu

ISBN 978-1-4419-9355-7
DOI 10.1007/978-1-4419-9356-4

e-ISBN 978-1-4419-9356-4

Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011925233

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

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

I dedicate this book to state assessment leaders like Sandra Berndt, Elizabeth Compton, Leola Sueoka, and Dawn McGrath who have dedicated their careers to supporting teachers and making sure all students have access to quality instruction and meaningful assessments. Without them and colleagues like them in other states, this book would be irrelevant. With them, there is hope.

- Steve

I dedicate this book to Kelly and Austin, who are always supportive of my career.

- Ryan

Special thanks to Darin Gordon, who showed me the way out of the mire, and to Austin Cagle, who helped set my feet on the rock.

- Peter

Wir stehen am Anfang. Der Dank hierfür geht an Maria, Heinrich, Rebecca, Gavin, Zachary und Doktorvater Steve. Together we can try to turn the tide.

- Alex

Preface

For several decades, the concept of access has been emphasized by educators, researchers, and policy makers in reference to the critical need for educational equity for all students. Recently, the term *accessibility* has been used to describe the degree to which achievement tests permit the full range of test-takers to demonstrate what they know and can do, regardless of disability status or other individual characteristics. This handbook contains the perspectives of experts in policy, research, and practice who share the common goal of defining and operationalizing accessibility to advance the development and use of tests that yield valid inferences about achievement for all students.

Tempe, Arizona
Nashville, Tennessee
Nashville, Tennessee
Nashville, Tennessee

Stephen N. Elliott
Ryan J. Kettler
Peter A. Beddow
Alexander Kurz

Contents

1	Creating Access to Instruction and Tests of Achievement: Challenges and Solutions	1
	Stephen N. Elliott, Peter A. Beddow, Alexander Kurz, and Ryan J. Kettler	

Part I Government Policies and Legal Considerations

2	U.S. Policies Supporting Inclusive Assessments for Students with Disabilities	19
	Susan C. Weigert	
3	U.S. Legal Issues in Educational Testing of Special Populations	33
	S.E. Phillips	
4	IEP Team Decision-Making for More Inclusive Assessments: Policies, Percentages, and Personal Decisions	69
	Naomi Zigmond, Amanda Kloo, and Christopher J. Lemons	
5	Australian Policies to Support Inclusive Assessments	83
	Michael Davies and Ian Dempsey	

Part II Classroom Connections

6	Access to What Should Be Taught and Will Be Tested: Students' Opportunity to Learn the Intended Curriculum	99
	Alexander Kurz	
7	Instructional Adaptations: Accommodations and Modifications That Support Accessible Instruction	131
	Leanne R. Ketterlin-Geller and Elisa M. Jamgochian	
8	Test-Taking Skills and Their Impact on Accessibility for All Students	147
	Ryan J. Kettler, Jeffery P. Braden, and Peter A. Beddow	

Part III Test Design and Innovative Practices

- 9 Accessibility Theory: Guiding the Science and Practice of Test Item Design with the Test-Taker in Mind . . . 163**
Peter A. Beddow, Alexander Kurz, and Jennifer R. Frey
- 10 Validity Evidence for Making Decisions About Accommodated and Modified Large-Scale Tests 183**
Gerald Tindal and Daniel Anderson
- 11 Item-Writing Practice and Evidence 201**
Michael C. Rodriguez
- 12 Language Issues in the Design of Accessible Items 217**
Jamal Abedi
- 13 Effects of Modification Packages to Improve Test and Item Accessibility: Less Is More 231**
Ryan J. Kettler
- 14 Including Student Voices in the Design of More Inclusive Assessments 243**
Andrew T. Roach and Peter A. Beddow
- 15 Computerized Tests Sensitive to Individual Needs 255**
Michael Russell
- 16 The 6D Framework: A Validity Framework for Defining Proficient Performance and Setting Cut Scores for Accessible Tests 275**
Karla L. Egan, M. Christina Schneider, and Steve Ferrara

Part IV Conclusions

- 17 Implementing Modified Achievement Tests: Questions, Challenges, Pretending, and Potential Negative Consequences 295**
Christopher J. Lemons, Amanda Kloof, and Naomi Zigmond
- 18 Accessible Tests of Student Achievement: Access and Innovations for Excellence 319**
Stephen N. Elliott, Ryan J. Kettler, Peter A. Beddow, and Alexander Kurz
- Subject Index 329**

Contributors

Jamal Abedi Graduate School of Education, University of California, Davis, CA 95616, USA, jabedi@ucdavis.edu

Daniel Anderson University of Oregon, Eugene, OR 97403, USA, daniela@uoregon.edu

Peter A. Beddow Department of Special Education, Peabody College of Vanderbilt University, Nashville, TN 37067, USA, peterbeddow@gmail.com

Jeffery P. Braden College of Humanities and Social Sciences, North Carolina State University, Raleigh, NC 27695-8101, USA, jeff_braden@ncsu.edu

Michael Davies School of Education and Professional Studies, Faculty of Education, Griffith University, Brisbane, QLD 4122, Australia, m.davies@griffith.edu.au

Ian Dempsey School of Education, University of Newcastle, Callaghan, NSW 2308, Australia, ian.dempsey@newcastle.edu.au

Karla L. Egan CTB/McGraw-Hill, Monterey, CA 93940, USA, karla_egan@ctb.com

Stephen N. Elliott Learning Sciences Institute, Arizona State University, Tempe, AZ 85287, USA, steve_elliott@asu.edu

Steve Ferrara CTB/McGraw-Hill, Washington, DC 20005, USA, steve_ferrara@ctb.com

Jennifer R. Frey Peabody College of Vanderbilt University, Nashville, TN 37067, USA, jennifer.frey@vanderbilt.edu

Elisa M. Jamgochian University of Oregon, Eugene, OR 97403, USA, ejamgoch@uoregon.edu

Ryan J. Kettler Department of Special Education, Peabody College of Vanderbilt University, Nashville, TN 37067, USA, ryan.j.kettler@vanderbilt.edu

Leanne R. Ketterlin-Geller Southern Methodist University, Dallas, TX 75252, USA, lkgeller@smu.edu

Amanda Kloo Department of Instruction and Learning, School of Education, University of Pittsburgh, Pittsburgh, PA 15260, USA, dramandakloo@gmail.com

Alexander Kurz Department of Special Education, Peabody College of Vanderbilt University, Nashville, TN 37067, USA, alexander.kurz@vanderbilt.edu

Christopher J. Lemons University of Pittsburgh, Pittsburgh, PA 15260, USA, lemons@pitt.edu

S.E. Phillips Consultant, Mesa, AZ 85205, USA, sephillips2@aol.com

Andrew T. Roach Department of Counseling and Psychological Services, Georgia State University, Atlanta, GA 30302, USA, cpsatr@langate.gsu.edu

Michael C. Rodriguez University of Minnesota, Minneapolis, MN 55455, USA, mcrdz@umn.edu

Michael Russell Boston College, Chestnut Hill, MA, USA; Nimble Innovation Lab, Measured Progress, Newton, MA 02458, USA, mike@nimbletools.com

M. Christina Schneider CTB/McGraw-Hill, Columbia, SC 29205, USA, christina_schneider@ctb.com

Gerald Tindal University of Oregon, Eugene, OR 97403, USA, gerald.tindal@mac.com

Susan C. Weigert U.S. Department of Education, Office of Special Education Programs, Alexandria, VA, USA, susan.weigert@ed.gov

Naomi Zigmond University of Pittsburgh, Pittsburgh, PA 15260, USA, naomi@pitt.edu

About the Editors

Stephen N. Elliott, PhD, is the founding Director of the Learning Sciences Institute, a trans-university research enterprise at Arizona State University, and is the Mickelson Foundation Professor of Education. He received his doctorate at Arizona State University in 1980 and has been on the faculty at several major research universities, including the University of Wisconsin-Madison and Vanderbilt University. At Wisconsin (1987–2004), Steve was a professor of educational psychology and served as the Associate Director of the Wisconsin Center for Education Research. At Vanderbilt (2004–2010), he was the Dunn Family Professor of Educational and Psychological Assessment in the Special Education Department and directed the Learning Sciences Institute and Dunn Family Scholars Program. His research focuses on scale development and educational assessment practices. In particular, he has published articles on (a) the assessment of children’s social skills and academic competence, (b) the use of testing accommodations and alternate assessment methods for evaluating the academic performance of students with disabilities for educational accountability, and (c) students’ opportunities to learn the intended curriculum. Steve’s scholarly and professional contributions have been recognized by his colleagues in education and psychology research as evidenced by being selected as an American Psychological Association Senior Scientist in 2009. Steve consults with state assessment leaders on the assessment and instruction of Pre-K-12 students, and serves on ETS’s Visiting Research Panel, and is the Director of Research and Scientific Practice for the Society of the Study of School Psychology.

Ryan J. Kettler, PhD, is a Research Assistant Professor in Special Education at Peabody College of Vanderbilt University. He received his doctorate in Educational Psychology, with a specialization in School Psychology, from the University of Wisconsin-Madison in 2005. Ryan’s dissertation, *Identifying students who need help early: Validation of the Brief Academic Competence Evaluation Screening System*, won the 2006 Outstanding Dissertation award from the Wisconsin School Psychologists Association. In 2007, he was named an Early Career Scholar by the Society for the Study of School Psychology. Prior to joining Vanderbilt University, Ryan was an assistant professor at California State University, Los Angeles, and completed an APA-accredited internship at Ethan Allen School in Wales, Wisconsin. He has worked on multiple federally funded grants examining the effectiveness of alternate assessments, academic and behavioral screening systems, and testing

accommodations. Ryan is the author of peer-reviewed publications and presentations within the broader area of data-based assessment for intervention, representing specific interests in academic and behavioral screening, inclusive assessment, reliability and validity issues, and rating scale technology. He currently serves as a consultant to College Board and to the Wisconsin Center for Education Research, providing expertise in the area of inclusive assessment.

Peter A. Beddow, PhD, received his doctorate in Special Education and Educational Psychology at Vanderbilt University in 2011. His research focuses on test accessibility and item writing for assessments of student achievement. He is the senior author of the *Test Accessibility and Modification Inventory (TAMI)* and the *Accessibility Rating Matrix*, a set of tools for evaluating the accessibility of test items for learners with a broad range of abilities and needs. Based on his work on accessibility theory, Peter was awarded the Bonsal Education Research Entrepreneurship Award in 2009 and the Melvyn R. Semmel Dissertation Research Award in 2010. Prior to beginning his academic career, Peter taught for 7 years in Los Angeles County, including 5 years teaching Special Education for students with emotional and behavior problems at Five Acres School, part of a residential treatment facility for children who are wards-of-the-court for reasons of abuse and neglect. Peter's primary goal is to help children realize their infinite value and achieve their ultimate potential. Peter lives in Nashville, Tennessee.

Alexander Kurz, MEd, is a doctoral student in Special Education and the Interdisciplinary Program in Educational Psychology at Vanderbilt University. He has studied in Germany and the United States, earning degrees in Special Education and Philosophy. Upon moving to the United States, he worked as a special education teacher in Tennessee and California, designed and implemented curricula for reading intervention classes, and participated in school reform activities through the Bill and Melinda Gates Foundation. Prior to beginning his doctoral studies, Alex worked as behavior analyst for children with autism and as an educational consultant to Discovery Education Assessment. During his graduate work at Vanderbilt, he collaborated with the Wisconsin Center for Education Research and Discovery Education Assessment, leading research efforts to examine curricular alignment and its relation to student achievement for students with and without disabilities. Alex has coauthored several peer-reviewed publications on alignment and alternate assessment. His latest scholarly contributions have reexamined the concepts of opportunity-to-learn (OTL), alignment, and access to the general curriculum in the context of curricular frameworks for general and special education. Alex is the senior author of *My Instructional Learning Opportunities Guidance System*, a teacher-oriented OTL measurement tool. His current interest in educational technology and innovation is aimed at identifying and creating pragmatic solutions to the problems of practice.

Creating Access to Instruction and Tests of Achievement: Challenges and Solutions

1

Stephen N. Elliott, Peter A. Beddow,
Alexander Kurz, and Ryan J. Kettler

Access is a central issue in instruction and testing of all students. Most of us can remember a testing experience, whether for low or high stakes, where the test questions covered content that we had not been taught. Many of us also have had testing experiences where the test items seemed “tricky” or poorly written, thus making it difficult to show what we had learned about the content the test was intended to measure. In both types of testing situations, one’s test performance is often negatively influenced because access to the intended knowledge and skills is limited by incomplete instruction or poor test items. In addition, it is likely that these testing situations were considered unfair and engendered negative attitudes about tests. These are not desired outcomes of instruction or testing. By improving access to instruction and tests, the results of testing can be more meaningful, and potentially positive, for all users.

For many students with disabilities, testing situations like these continue to occur. Access has been affected, if not denied, as a result of limited opportunities to learn valued knowledge and skills as well as by test items that feature extraneous content and designs insensitive to persons with various sensory and cognitive disabilities.

Access to education, and in particular the grade-level curriculum, lies at the heart of virtually all federal legislation for students with disabilities. This access to instruction is a prerequisite and necessary condition for validity claims about test scores from statewide assessments of academic achievement. Ideally, all students are provided high-quality instruction that offers the opportunity to learn the knowledge and skills in a state’s intended curriculum and assessed on the state’s achievement test. Ideally, eligible students also are provided needed testing accommodations to reduce the effects of disability-related characteristics, and thus facilitate access to tested content. Tests used to measure student achievement should be designed to provide all students optimal access to the targeted constructs without introducing variance due to extraneous test features.

Unfortunately, this ideal scenario of an unobstructed access pathway to learning and demonstrating the knowledge and skills expressed in the general curriculum is not verified by recent research or our observations of educational practices in numerous states. Too often it seems that students have not had opportunities to learn essential knowledge and skills nor have the tests used to evaluate their achievement been highly accessible or well aligned with their instruction. When access is limited to high-quality instruction and the tests that measure it, the test results are misleading at best and the side effects on students are potentially demoralizing.

S.N. Elliott (✉)
Learning Sciences Institute, Arizona State University,
Tempe, AZ 85287, USA
e-mail: steve_elliott@asu.edu

This chapter is intended to lay the foundation for subsequent chapters by researchers with experience in teaching, testing, and advancing equitable educational accountability. Here we examine access challenges related to instruction and testing often experienced by students with special needs, and survey-relevant research related to three areas: (a) opportunity to learn (OTL), (b) testing accommodations, and (c) the accessibility of achievement tests. In addition, we advance actions for overcoming barriers to access for all students. The context for this focus on access has been strongly influenced by federal legislation, inclusive assessment research, test score validity theory, and the desire to improve education for students with special needs.

Legislative Context and Key Concepts

The education of students with disabilities is strongly influenced by federal legislation that mandates physical and intellectual access to curriculum, instruction, and assessment. Key federal legislation on access for students with disabilities includes the Rehabilitation Act of 1973 and the Individuals with Disabilities Education Act (IDEA, 1975) and its subsequent reauthorizations. (See [Chapter 3](#) by Phillips for a comprehensive review of legal issues of access for students.) These legal mandates serve as the foundation for the inclusion of students with disabilities in standards-based reform and test-based accountability under the No Child Left Behind Act (NCLB) of 2001. The reauthorization of IDEA in 1997 included the access to the general curriculum mandates, which were intended to (a) provide all students with disabilities access to a challenging curriculum; (b) yield high expectations for all students with disabilities; and (c) ensure that all students with disabilities were included in test-based accountability mechanisms such as large-scale testing, progress monitoring, and public performance reporting. The universal accountability provisions of NCLB continued to underscore and expand access for students with disabilities by mandating academic content that is aligned with the local and statewide grade-level

standards of students without disabilities (Kurz & Elliott, in press).

Current federal legislation requires the application of universal design principles to the development of all state and district-wide achievement tests. Universal design (UD), as defined in the Assistive Technology Act (P.L. 105-394, 1998), is “a concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly usable (without requiring assistive technologies) and products and services that are made usable with assistive technologies” (§3(17)). This legislation provides the rationale for the use of UD principles as follows:

The use of universal design principles reduces the need for many specific kinds of assistive technology devices and assistive technology services by incorporating accommodations for individuals with disabilities before rather than after production. The use of universal design principles also increases the likelihood that products (including services) will be compatible with existing assistive technologies. These principles are increasingly important to enhance access to information technology, telecommunications, transportation, physical structures, and consumer products (PL105-394(§3(10))).

Prior to the 2007 amendments to NCLB (U.S. Department of Education, 2007a, 2007b), access barriers in testing were addressed primarily by the use of testing accommodations, which are typically defined as changes in the administration procedures of a test to address the special needs of individual test-takers (Hollenbeck, 2002). More recently, test developers have begun examining tests and items with the goal of modifying them to reduce the influence of intrinsic access barriers on subsequent test scores (e.g., Kettler, Elliott, & Beddow, 2009). This process has led to the development of what Beddow (2010) has called accessibility theory. Accessibility is defined as “the degree to which a test and its constituent item set permit the test-taker to demonstrate his or her knowledge of the target construct of the test” (Beddow, 2010, see Chapter 9). Accessibility is conceptualized as

the sum of interactions between features of the test and individual test-taker characteristics.

Although the term *accessibility* is not used in the definition of UD, the principles as applied to assessment technology clearly are intended to address issues of access. To the extent a test contains access barriers for a portion of the tested population, the validity of inferences made from test scores is affected. The validity of subsequent norms and comparisons across the population is also likely affected. In summary, the validity of test score inferences is dependent on the accessibility of the test for the entirety of the target test-taker population.

One of the final amendments to NCLB (U.S. Department of Education, 2007a, 2007b) indicates that a small group of students with disabilities is allowed to show proficiency through an alternate assessment based on modified academic achievement standards (AA-MAS). These students can take a version of the regular assessment test with modifications and may constitute up to 2% of all who are reported proficient within a school. According to a recent report, 14 states are developing AA-MASs (Lazarus, Hodgson, & Thurlow, 2010). *Modifications* are defined as changes to a test's content or item format. The regulations strongly emphasize that although modifications may make a test easier, out-of-level (i.e., below grade level) testing is not acceptable, leaving developers and users of these AA-MASs to determine at which point a test is no longer within the intended level. Modifications to large-scale achievement tests, like testing accommodations, are intended to facilitate access to the assessment for students with special needs, so that their scores can be meaningfully compared with the scores of students who take the standard test. If this can be accomplished, better assessment and accountability for students with disabilities will be the result (Kettler et al., 2009).

The legislative push to develop AA-MASs for students identified with disabilities has inspired the examination of accessibility and its relation to test score validity. The resulting theory and evidence supporting its importance has indicated accessibility affects students' test performance across the range of the test-taker population.

Indeed, the differential boost observed across the testing accommodations literature is similarly evident in the results of research on item modifications (Kettler et al., in press; Sireci, Scarpati, & Li, 2005). Much more will be reported about the effects of item modifications in Chapter 9 by Beddow.

The terms *access*, *accommodations*, and *modifications* all have been used for decades when discussing educational testing and the validity of resulting test scores. These terms represent key concepts in the world of testing and federal assessment and accountability policies. Thus, these terms demand attention to ensure they are understood in the context of emerging issues around tests and testing for students with disabilities.

For instruction, *access* is the opportunity for a student to learn the content of the intended and assessed curricula. In the current education reform framework, this means students have meaningful opportunities to acquire the knowledge and skills featured in the content standards of their state and ultimately assessed on the state's end-of-year achievement test. Teachers are encouraged to teach to the standards, not the test, and create engaging instruction for all students to increase the chances that learning occurs.

For educational testing, *access* is the opportunity for test-takers to demonstrate proficiency on the target construct of a test (e.g., language arts, mathematics, or science) or item (e.g., synonyms, homonyms, and homographs). In essence, complete access is manifest when a test-taker is fully able to show the degree to which he or she knows the tested content. Access, therefore, must be understood as an interaction between individual test-taker characteristics and features of the test itself.

The purpose of both testing accommodations and modifications is to increase individuals' access to tests. The definitions of these access-enabling strategies, however, have been the subject of debate, in part because of their inconsistent use in the *Standards for Educational and Psychological Testing* (Standards; American Educational Research Association, American Psychological Association, & National Council

for Measurement in Education, 1999) and some states' testing guidelines. An examination of the *Standards for Testing* finds the term *modification* cited nearly a dozen times in the index. A number of the standards refer to a modification as an accommodation. For example, in the section entitled "Testing Individuals with Disabilities" in the *Standards for Testing*, it is stated:

The terms *accommodation* and *modification* have varying connotations in different subfields. Here accommodation is used as the general term for any action taken in response to a determination that an individual's disability requires a departure from established testing protocol. Depending on circumstances, such accommodation may include modification of test administration processes or modification of test content. No connotation that modification implies a change in construct(s) being measured is intended. (AERA et al., 1999, p. 101)

The implementation of testing accommodations for students with disabilities is a policy endorsed in all states; some states even allow testing accommodations for all students if they have been provided the same accommodations during instruction. *Accommodations* are widely recognized in state testing guidelines as changes to the setting, scheduling, presentation format, or response format of an assessment (Kettler & Elliott, 2010). The modification of test content, however, is inconsistent with the definition of a testing accommodation in the majority of state testing accommodation guidelines (Lazarus, Thurlow, Lail, Eisenbraun, & Kato, 2006). Accommodations are made to increase the validity of inferences that can be made from a student's scores, so that those scores can be meaningfully compared to scores of students for whom testing accommodations are not needed.

The modification of test content is inconsistent with the definition of a testing accommodation in the majority of state testing accommodation guidelines (Lazarus et al., 2006). The AA-MAS policy, however, extends the notion of access and the spirit of individualized accommodations to changes made to item content. Such changes are defined as modifications by most test developers and administrators. As we have previously observed, when item and test alterations are made (e.g., by changing the layout, reducing the length

of the reading passage, adding graphic support), it is not always clear without test results whether the changes affect only access to the test or, in fact, also affect the construct being measured and the subsequent inferences that are drawn from scores (Kettler et al., 2009). If the content of an item or test has been changed and evidence is not available to show that scores remain comparable to the original construct to be measured, it has been customary to consider the alteration a modification (Phillips & Camara, 2006; Koretz & Hamilton, 2006). To ensure the modifications are acceptable under AA-MAS policy, research is needed to confirm the modified test measures the same construct(s) as the original test. Indeed, the policy assumes AA-MASs for eligible students measure the same grade-level content standards and performance objectives (constructs) as general assessments for students without disabilities. It should be noted that some modifications may result in an AA-MAS that yields scores that are not comparable to scores obtained from the original test. These modifications could still be permissible for an AA-MAS, assuming the content is commensurate with the intended grade-level of the test (Kettler et al., 2009).

Given the approach we recommend for developing an AA-MAS, we defined the term *modification* to refer "to a process by which the test developer starts with a pool of existing test items with known psychometric properties, and makes changes to the items, creating a new test with enhanced accessibility for the target population. When analyses indicate inferences made from the resulting test scores are valid indicators of grade-level achievement, the modifications are considered appropriate. Conversely, if analytic evidence suggests the inferences made from resulting scores are invalid indicators of grade-level achievement, the modifications are inappropriate" (Kettler et al., 2009, p. 531). Thus, like individualized testing accommodations, modifications must be studied to determine their appropriateness. Unlike accommodations, modifications are intended to afford access to an entire group of students, resulting in better measurement of their achieved knowledge.

Providing Access to Overcome Barriers

We believe three critical points exist for improving access for students within a test-based accountability system. These points occur during instruction, arrangement of the testing situation, and the design of tests and include providing students meaningful opportunities to learn the intended curriculum and to show what they have learned on highly accessible test with appropriate accommodations. The validity of some test score inferences depends on students' opportunity to learn the intended curriculum as well as on students' access to tests that are well aligned with the intended curriculum. When access at these points is not optimal, it is difficult to draw inferences about students' learning in school. Figure 1.1 illustrates how the failure to provide students access to curriculum and tests creates barriers to success.

The state of research differs for each access barrier. OTL features a substantial body of theory-driven research that supports its importance for student achievement, especially in the context of test-based accountability (Airasian & Madaus, 1983; Hermann, Klein, & Abedi, 2000; McDonnell, 1995; Porter, 1995). However, challenges remain regarding the conceptualization and measurement of OTL using practical tools (Kurz, Elliott, Wehby, & Smithson, 2010;

Pullin & Haertel, 2008; Roach, Niebling, & Kurz, 2008). Testing accommodations research has grown significantly over the past decade with reports focusing on their appropriateness (Hollenbeck, 2002; Phillips, 1994), assignment and delivery (Elliott, 2007; Fuchs & Fuchs, 2001; Ketterlin-Geller, Alonzo, Braun-Monegan, & Tindal, 2007), and effects on student achievement (Elliott, Kratochwill, & McKeivitt, 2001; Elliott & Marquart, 2004; Sireci, Scarpatti, & Li, 2005). In short, strategies for increasing access via testing accommodations and its typical effects (if delivered with integrity) are available. Research concerned with designing accessible tests is relatively new and largely based on UD principles (Ketterlin-Geller, 2005; Kettler et al., 2009; Thurlow et al., 2009). More recently, researchers have applied principles of UD, cognitive load theory, and item development research to modify items and tests in an effort to increase access for students with disabilities (Elliott, Kurz, Beddow, & Frey, 2009; Kettler et al., 2009). In the remainder of this chapter, we examine critical research related to each access barrier and discuss available strategies for increasing access.

Access via Opportunity to Learn

Although the primacy of OTL in the context of accessibility may be apparent – after all,

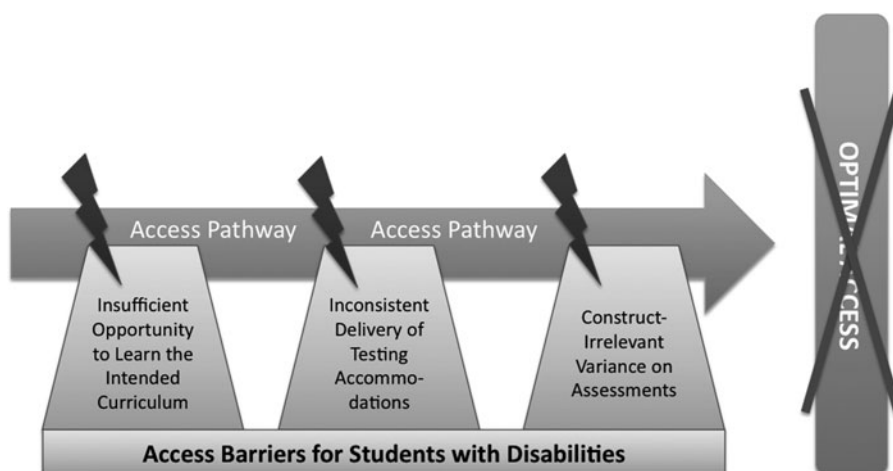


Fig. 1.1 Access barriers to the general curriculum and tests for students with disabilities

one fundamental assumption of schooling is that student learning occurs as a function of formal instruction – OTL has not yet captured a foothold in test-based educational accountability. The conceptual and methodological challenges surrounding OTL are at least partly responsible for its slow ascendancy in the realm of curriculum, instruction, and assessment (Porter, 1995). In its most general definition, OTL refers to the opportunities that schools afford their students to learn what is expected of them (Herman et al., 2000). This definition highlights two important conceptual issues: the “who” and “what” of OTL. With regard to the latter issue, standards-based reforms influenced by minimum competency testing, performance assessment, and large-scale assessment for educational accountability led to the establishment of performance expectations for students via subject-specific content standards across the grade spectrum. The content of these standards is typically referred to as the intended curriculum (Porter, 2006). Consequently, the definition by Herman et al. can be revised to OTL as referring to *students’* opportunity to learn the *intended curriculum* (Kurz & Elliott, in press).

Numerous factors can affect OTL at the school (e.g., class size, instructional resources), teacher (e.g., instructional content, subject matter knowledge, instructional time), and student level (e.g., engagement, extracurricular activities), which has contributed to the proliferation of the OTL acronym under disparate conceptual definitions including OTL as *teachers’* opportunity to learn subject matter knowledge (Schmidt et al., 2008). These various factors notwithstanding, the most proximal variable to the instructional lives of students and their opportunity to learn the intended curriculum is the *teachers’ instruction* (e.g., Kurz & Elliott, in press; Rowan, Camburn, & Correnti, 2004). In the case of special education teachers, there even exists a legal precedence to provide students with disabilities access to intended curriculum under IDEA’s access to the general curriculum mandates (Cushing, Clark, Carter, & Kennedy, 2005; Roach et al., 2008). Conceptual and methodological challenges regarding OTL, however, persist in the decomposition of classroom instruction:

What aspects of classroom instruction are significant contributors to student achievement? How can these aspects be measured?

Over the last four decades, three broad strands of OTL research have emerged related to teacher instruction focused primarily on the *content* of instruction (e.g., Husén, 1967; Rowan & Correnti, 2009), the *time* on instruction (e.g., Carroll, 1963; Vannest & Hagan-Burke, 2009), and the *quality* of instruction (e.g., Brophy & Good, 1986; Pianta, Belsky, Houts, Morrison, & NICHD, 2009). A merger of the various OTL conceptualizations was first articulated by Anderson (1986) and explicitly conceptualized by Stevens (1993), although some researchers have combined some of these OTL variables prior to Stevens (e.g., Cooley & Leinhardt, 1980). Researchers have further provided empirical support for the relation between each of those variables and student achievement (e.g., Gamoran, Porter, Smithson, & White, 1997; Rowan, 1998; Thurlow, Ysseldyke, Graden, & Algozzine, 1984). However, few research studies have examined all three aspects of OTL to determine their combined and unique contributions (e.g., Wang, 1998).

The measurement of OTL related to the content of instruction was initially motivated by concerns about the validity of test score inferences (e.g., Husen, 1967; Airasian & Madaus, 1983; Haertel & Calfee, 1983). Two popular approaches were *item-based* OTL measures and *taxonomic* OTL measures. To determine students’ opportunity to learn tested content, researchers adopted item-based OTL measures that required teachers to indicate the extent to which they covered the content measured by different test items using some type of rating scale (e.g., Comber & Keeves, 1973). To determine students’ opportunity to learn important content objectives, researchers developed content taxonomies that could be used to judge whether different tests covered the same objectives delineated in the taxonomy (e.g., Porter et al., 1978). The latter approach permitted content comparisons across different tests and other media such as textbooks. The flexibility of this approach contributed to its popularity. Porter and colleagues,

for example, continued their work on taxonomic OTL measures and eventually developed the *Surveys of the Enacted Curriculum* (SEC), one of two alignment methodologies presently used by states to document the content overlap between assessments and state standards (CCSSO, 2009). Besides teacher self-report, other measurement techniques used to measure the content of instruction include direct observation and review of permanent products (see Porter, Kirst, Osthoff, Smithson, & Schneider, 1993).

The concept of OTL as time on instruction was introduced as part of John Carroll's (1963) model of school learning. He defined OTL as the amount of time allocated for learning at the school or program level. Subsequent research examined the relation between allocated time and student achievement, continuously refining the "quantity of instruction" from allocated time (e.g., 45-min lessons) to time actually engaged in learning (e.g., Harnischfeger & Wiley, 1976; Denham & Lieberman, 1980; Gettinger & Seibert, 2002). The measurement of time on instruction initially involved the straightforward summing of allocated class time across the school year, while later measures relied on teacher report or direct observation (e.g., Ysseldyke et al., 1984; O'Sullivan, Ysseldyke, Christenson, & Thurlow, 1990). The amount of time dedicated to instruction has received substantial empirical support in predicting student achievement (e.g., Berliner, 1979; Brophy & Good, 1986; Fisher & Berliner, 1985; Scheerens & Bosker, 1997; Walberg, 1988). Vannest and Parker (2009) examined time usage related to instruction and concluded that time on instruction represents the single best documented predictor of student achievement across schools, classes, student abilities, grade levels, and subject areas.

School and classroom indicators of OTL related to the quality of instruction were first introduced as part of several models of school learning (e.g., Bloom, 1976; Carroll, 1963; Gagné, 1977; Harnischfeger & Wiley, 1976). Walberg (1986) reviewed 91 studies that examined the effect of putative quality indicators on student achievement such as frequency of praise statements, frequency of corrective

feedback, availability of instructional resources, and instructional grouping and reported the highest mean effect sizes for praise and corrective feedback with 1.17 and 0.97, respectively. Brophy (1986) in a review of his meta-analysis reported active teaching, effective classroom management, and teacher expectations related to the content of instruction as key quality variables of OTL with strong empirical support (e.g., Brophy & Everston, 1976; Coker, Medley, & Soar, 1980; Fisher et al., 1980). More recently, OTL research on the quality of instruction also considered teacher expectations for the enacted curriculum (i.e., cognitive demands) and instructional resources such as access to textbooks, calculators, and computers (e.g., Boscardin, Aguirre-Munoz, Chinen, Leon, & Shin, 2004; Herman & Klein, 1997; Porter, 1991, 1993; Wang, 1998). Teacher self-report and direct observation by trained observers represent the typical measurement techniques for determining quality aspects of instruction such as expectations for student learning, instructional practices, and/or instructional resources (e.g., Rowan & Correnti, 2009; Pianta & Hamre, 2009).

In summary, the concept of OTL is concerned with students' opportunity to learn the intended curriculum. Students access the intended curriculum indirectly via the teacher's instruction. In the context of test-based educational accountability, access to the intended curriculum – OTL – is critical, because end-of-year testing programs sample across the content domains of the intended curriculum for purposes of measuring student achievement and the extent to which schools and teachers have contributed to achievement. Although the various strands of OTL research have converged on the teacher's enacted curriculum, they did so by focusing on different aspects of a teacher's instruction: its time, content, and quality. Empirical data support each OTL variable as a correlate of student achievement, yet few researchers have investigated all three OTL variables – time, content, and quality – as a set to determine their unique and combined contributions. This gap in the researcher literature is unfortunate, because neither aspect of OTL can occur in isolation for

all practical purposes. That is, instructional content enacted by a teacher always has to unfold along (at least) two additional dimensions: time and quality. For example, a teacher's instruction is not adequately captured by referring solely to the content of instruction such as *solving algebraic equations*. In actuality, a teacher decides to teach students the *application of solving an algebraic equation to a context outside of mathematics for 35 min* through *guided practice*. The different sets of italicized words refer to various aspects of OTL – time, content, and quality of instruction – that have to occur in conjunction with one another whenever instruction is enacted by a teacher. Two important challenges thus have to be met before OTL can begin to play an integral part in curriculum, instruction, and assessment. First, the measurement of OTL has to be refined to allow for the efficient and reliable data collection of the enacted curriculum along time, content, and quality aspects of instruction. Second, the measurement of OTL has to move beyond measurement for measurement's sake. Besides measuring and verifying comprehensive OTL as a significant contributor to student achievement, we envision a shift in OTL research and measurement that provides practitioners with tools that yield additional *formative benefits* that can enhance instruction and increase access for all students. Kurz addresses this shift in detail in [Chapter 6](#).

Access via Testing Accommodations

Testing accommodations historically have been used with the aim of reducing construct-irrelevant variance due to a variety of access skill deficits exhibited by students with disabilities. Testing accommodations are individualized depending on each student's access needs. Typically, accommodations involve changes in the *presentation format* of a test (e.g., oral delivery, paraphrasing, Braille, sign language, encouragement, permitting the use of manipulatives), the *timing or scheduling* of a test (e.g., extended time, delivering the test across multiple days), the *recording or response format* (e.g., permitting test-takers to respond in the test booklet instead of on the

answer sheet, transcription), or the *assessment environment* (e.g., separate room, elimination of distractions) (Elliott, Kratochwill, Gilbertson-Schulte, 1999; Kettler & Elliott, 2010).

Appropriate testing accommodations, while applied individually based on specific student needs, should not interfere with the test's measurement of the target construct and should permit the same validity of inferences from the results of the test as those from students not receiving accommodations (Hollenbeck, Rozek-Tedesco, & Finzel, 2000). The application of accommodations should also differentially affect test results of students for whom accommodations are intended, compared to those for whom testing accommodations are not needed. Most accommodation researchers subscribe to the concept of *differential boost* (Fuchs & Fuchs, 2001; Fuchs et al., 2000), which is used to identify valid accommodations as those that “will lead to greater score improvements for students with disabilities than for students without disabilities” (Sireci, Scarpati, & Li, 2005, p. 481). Sireci et al. differentiated the concept of differential boost from the traditional definition of the *interaction hypothesis*, which states that “(a) when test accommodations are given to the SWD [students with disabilities] who need them, their test scores will improve, related to the scores they would attain when taking the test under standard conditions; and (b) students without disabilities will *not* exhibit higher scores when taking the test with those accommodations” (p. 458).

A review of the research literature reveals a number of studies examining the differential boost of testing accommodations and the validity of score interpretations (e.g., Elliott, McKeivitt, Kettler, 2002; Feldman, Kim, & Elliott, 2009; Fuchs, Fuchs, & Cappizzi, 2005; Kettler et al., 2005; Kosciolik & Ysseldyke, 2000; Lang, Elliott, Bolt, & Kratochwill, 2008; McKeivitt & Elliott, 2003; Pitoniak & Royer, 2001; Sireci et al., 2005). The National Research Council's commissioned review of research on testing accommodations by Sireci et al. (2005) is one of the most comprehensive reviews of the evidence for effects on test scores of testing accommodations. Specifically, Sireci et al. (2005)

reviewed 28 experimental, quasi-experimental, and non-experimental empirical studies on the effects of testing accommodations over nearly two decades. They found the most common accommodations were reading support (39%) and extra time (24%). Aggregate results of studies on reading support (usually in the form of verbatim presentation of directions and test items) were mixed. The interaction hypothesis was upheld for five of the six studies examining the effect of the accommodation on scores from mathematics tests. For two studies on reading and two studies across multiple content areas, the interaction hypothesis was not upheld. The authors concluded that reading support, while likely increasing the validity of inferences for mathematics tests, may not have the desired effect when used with tests of other content domains. Results of five out of eight studies on extended time indicated students identified with disabilities exhibit higher score gains than students not identified with disabilities. The results of one study rejected the interaction hypotheses, and the results of two other studies that did not indicate extra time resulted in gains for either group. Based on these findings, Sireci and colleagues concluded that while the interaction hypothesis was not strictly upheld, “evidence . . . is tilted in that direction” (p. 469). Sireci and colleagues also reviewed several studies on the effects of multiple accommodations (i.e., accommodations packages). The findings of the four studies that used experimental designs supported the interaction hypothesis.

Reported effect sizes of most testing accommodations studies appear small, but there is evidence they are practically meaningful. In a survey of the accommodations literature, Kettler and Elliott (2010) reported in some studies, effect sizes from accommodations for students with IEPs (Individualized Education Program) were twice as large as those for students without IEPs. In one study, effect sizes ranged from 0.13 for students without IEPs to 0.42 for students with IEPs. While conventional interpretations of effect sizes (e.g., Cohen, 1992) would suggest these effects are small, a meta-analysis conducted by Bloom, Hill, Black, and Lipsey (2008) provides evidence that they are meaningful within the context of

changes in achievement test scores. Bloom et al. found that mean effect sizes across six standardized achievement tests from the spring semester of one school year to the next ranged from 0.06 to 1.52 in reading and 0.01 to 1.14 in mathematics, with larger effect sizes consistently observed for lower grades and steadily decreasing until grade 12. Further, the data suggest a steep increase in effect sizes from grade K until grade 5, after which no effect sizes above 0.41 are observed for either reading or mathematics through grade 12. This indicates that effect sizes of 0.40 or higher for students with disabilities may denote that the impact of accommodations is meaningful. Indeed, the differential boost reported by Kettler and Elliott provides evidence of an interaction that may heretofore have been underestimated. As applied to the accommodations literature, these results suggest for some students, appropriate accommodations may indeed reduce barriers and yield more accurate measures of achievement.

Although testing accommodations can be helpful for many students with disabilities, there are a number of challenges associated with implementing them. First, many students are averse to testing accommodations for different reasons including the fact that the accommodations often draw attention to student challenges (Feldman et al., 2009). Additionally, there are logistical challenges associated with their appropriate implementation including time, personnel, and cost, which often result in poor integrity. Another challenge is the difficulty in identifying which students should receive specific accommodations, and which combination of accommodations may be appropriate. Further, little is known about the extent to which accommodations interact with each other differentially across students or packages, notwithstanding the breadth of the research based on their use. Finally, each time accommodations are used, general and comparative validity are threatened. Not only is an additional variable introduced into the test event with each accommodation, but it is also introduced for some students and not for other students who may need some of the same accommodations but are not eligible for them (Kettler & Elliott, 2010).

Access via Well-Designed Test Items

Systematically attending to the accessibility of tests and their constituent items can reduce the need for individualized accommodations, thereby reducing potential threats to validity across the test-taker population. Results of research on item accessibility suggest many achievement test items can be improved to reduce access barriers for more test-takers (see Elliott et al., 2010; Kettler et al., in press). Cognitive interviews and test-taker survey data suggest when presented with two similar items, students respond favorably to the item developed with a focus on accessibility (Roach et al., 2010).

Universal design (UD; Mace, 1991), originally conceptualized as a set of guiding principles for ensuring buildings permit equal access for all individuals, provides a useful framework for understanding the variety of perspectives that must be taken into consideration when undertaking the effort to ensure assessment instruments are equally accessible across the range of the test-taker population. The theory does not, however, contain guidelines that apply specifically to test and item development. In 2002, Thompson, Johnstone, and Thurlow applied the principles of UD to large-scale assessment and distilled them into seven recommendations, as follows: (a) inclusive assessment population; (b) precisely defined constructs; (c) accessible, non-biased items; (d) amendable to accommodations; (e) simple, clear, and intuitive instructions and procedures; (f) maximum readability and comprehensibility; and (g) maximum legibility. This report was a step forward in that it extended the use of UD beyond its architectural origins, but it contained few specific guidelines that informed the development of accessible tests or items.

Accessibility theory (Beddow, 2010) operationalizes test accessibility as the sum of interactions between characteristics of the individual test-taker and features of the test and its constituent item set (see also Ketterlin-Geller, 2008). Accessibility, therefore, is an attribute of the test event. Optimally accessible tests generate test events that minimize the influence of

construct-irrelevant interactions on subsequent test scores. While in theory, this objective is shared with testing accommodations, it must be noted that test items typically are delivered universally across the test-taker population. As such, item writers should focus on maximizing the universality of all test items, focusing particularly on item features that may reduce their accessibility for some test-takers.

Locating accessibility in the test event as opposed to defining it as an attribute of the test itself is an important distinction insofar as a test or test item may contain features that pose access barriers for one test-taker while permitting unfettered access to the target construct for another. Accessible items, therefore, must contain little or no content that compels the test-taker to demonstrate skills that are irrelevant to the construct intended for measurement. This is of particular importance when skills that are required in addition to the target construct (referred to as prerequisite skills) are challenging for the test-taker. Of these prerequisite skills, the clearest example is found in the vast number of mathematics items in which the target problem is situated in text. For a test-taker with low reading ability, complex text in a mathematics test likely represents an access barrier that may preclude him or her from fully demonstrating knowledge, skills, and/or abilities in math.

The inclusion of extraneous and/or construct-irrelevant demands, therefore, must be addressed at both the test and item levels to ensure that the resulting scores represent, to the extent possible, a measure of the target construct that is free from the influence of ancillary interactions due to access barriers. To this end, cognitive load theory (CLT; Chandler & Sweller, 1991), a model for understanding the effects of various features of instructional task demands on learning outcomes, offers a useful lens through which to evaluate the accessibility of tests and items. Based on Miller's (1956) notion of the limitations of working memory, CLT disaggregates task demands into three types of cognitive demand. For optimal learning efficiency, proponents of the theory recommend designers of instructional materials to eliminate

extraneous load while maximizing intrinsic load. This ensures the learner is permitted to allocate his or her cognitive resources to the primary objectives of the task. Beddow, Kurz, and Frey detail the application of CLT to the design of accessible test items in [Chapter 15](#).

In general, there are two primary options for ensuring tests are accessible for all test-takers. The first is to use existing research and theory to guide the process of developing tests and items, with the goal of maximizing accessibility for all test-takers. The second option is to evaluate existing tests and items with the purpose of identifying potential access barriers. Evaluation data are then used to guide test and item modifications based on specific access concerns to enhance their accessibility for more test-takers.

The goals of modification are twofold. First, modifications should preserve the target construct of the original test or test item. This includes preserving essential content, depth of knowledge, grade-levelness, and intended knowledge, skills, and abilities targeted for measurement. Second, modifications should isolate the target construct by reducing, to the degree possible, the influence of ancillary interactions on the test score. These modifications may include eliminating extraneous material, reorganizing item graphics and page layouts, and highlighting essential content.

Kettler et al. (2009) described a paradigm to guide the process of developing accessible tests and items that yield scores from which inferences are equally valid for all test-takers. The model comprises five stages. The first stage consists of a systematic accessibility review of an existing pool of test items. To facilitate the comprehensive and systematic analysis of tests and items based on the principles of accessibility theory, the authors developed a set of tools called the *Test Accessibility and Modification Inventory (TAMI)*; Beddow, Kettler, & Elliott, (2008) and *TAMI Accessibility Rating Matrix (ARM)*; Beddow, Elliott, & Kettler, (2009). The second stage involves collaborative modification, including content area experts, assessment specialists, educators, and/or item writers with expertise in issues of accessibility. The third

stage involves documenting all changes to the test and items. In the fourth stage, the new items are pilot tested with a small sample of respondents to gather information on the appropriateness and feasibility of the modifications. The final stage involves a large field test to evaluate item characteristics such as difficulty, discrimination, and reliability. Beddow and colleagues provide a detailed examination of theory and tools for designing highly accessible items in [Chapter 9](#).

Actions and Innovations Needed to Keep Moving Forward

To ensure and expand access to instruction and the tests designed to measure what all students have learned from instruction, we believe there are a number of actions and innovations needed. The theoretical and empirical foundations for these actions and innovations for improving access have been introduced in this chapter and are thoroughly examined throughout the rest of this book.

Key actions and innovations needed to improve access to instruction for all students include the following:

- *Providing teachers more support to implement curricula that are highly aligned with content standards and testing blueprints, and subsequently providing these teachers tools for getting periodic feedback about their efforts to increase students' opportunities to learn the intended content.* To provide the support teachers need is certain to involve more and innovative professional development activities that focus on the intended curriculum. With regard to feedback on their efforts to increase opportunities to learn, teachers will need new tools for measuring and monitoring instructional content, time, and quality.
- *Providing teachers more information about what students are learning as part of instruction.* To accomplish this action, instructionally sensitive, standards aligned, and accessible interim measures of achievement are needed.

These measures will be most effective if delivered online and scored immediately so that feedback about the relation between what has been taught and what has been learned is timely and salient.

Key actions and innovations needed to improve access to achievement tests for all students include the following:

- *The provision of testing accommodations for students who demonstrate a need and documentation that such accommodations were implemented with integrity.* To accomplish this action, it is likely that teachers will need support from colleagues during testing events and tools for documenting what accommodations are actually delivered will be needed. Another possible innovation needed to accomplish this action is to use computer-delivered tests with software that makes the accommodations possible and also records their implementation.
- *The development of test accessibility review panels, similar to fairness review panels, to increase the likelihood that all test items meet high standards of accessibility.* To accomplish this action, state assessment leaders simply need to recognize that the accessibility of many of the items they currently use can be enhanced. Innovative tools designed to measure test item accessibility and provide diagnostic information for both guiding the development and evaluation of items are needed.
- *The support of ongoing research and evaluation of test results to ensure they are valid indicators of what students have been taught and have learned.* To accomplish this action, state assessment leaders and staff need to conduct or hire others to conduct periodic validity studies that consider the effect of OTL, testing accommodations, and test item accessibility on the test scores of students with varying levels of ability.

We believe all these actions and related innovations can be accomplished on a large scale. We have the knowledge and tools today to move forward; we now need to share this information and help educators create the infrastructure for using it to ensure more access for all students to the intended and assessed curriculum.

Conclusions

Drawing inferences about whether the students in a school are successfully learning the lessons indicated in content standards, based on the instruction that teachers are providing, is a complicated process. Any accountability system that is designed to assist in making such inferences must take into consideration the behaviors and characteristics of more than one group. The system must yield information about the instructional practices of teachers, most practically obtained using teacher-rating measures, as well as information about the knowledge and skills that students have learned. This latter piece of information has historically been obtained using achievement tests. These two pieces of information – OTL and achievement – must be aligned to the same grade level content standards and should be used together in an organized framework. One such framework would involve drawing different conclusions based on whether OTL at a school is higher or lower than one target threshold, along with whether student achievement is higher or lower than a second target threshold. Table 1.1 depicts such a framework.

When OTL is high but achievement is low (A), the school should thoroughly examine its tests and process for assigning testing accommodations, to insure that the assessment is accessible. It may also be important to examine the system for identifying and helping students with special needs, some of whom may be unidentified, and without proper intervention may be achieving at a very low level. A school is only providing optimal access to grade-level content standards when both OTL and achievement are high (B). When both OTL and achievement

Table 1.1 Proposed interpretive framework for student achievement by instructional quality

Instructional quality	Student achievement	
	Low percentage proficient	High percentage proficient
High mean OTL	A	B
Low mean OTL	C	D

are low (C), professional development may be needed in the delivery of high-quality instruction related to content standards. Examining tests and accommodations might also be a good step, but until high-quality instruction is established, it is difficult to know whether the scores are low due to lack of OTL or due to barriers to assessment. Lastly, when OTL is low but achievement is high (D), professional development in instruction of content standards is necessary. It may also be appropriate in this case to raise standards for achievement, in line with what could be obtained if OTL for students were higher.

Opportunity to learn and accessible achievement tests are both necessary to providing optimal access for students to grade-level content standards. Testing accommodations bridge these two constructs as individualized methods of improving access, which are best identified in part by transferring instructional accommodations to the test event. Although work remains in refining the measurement both of OTL and achievement, including the development of measures that are practical, formative, and accessible, the state of the field is such that large-scale assessment systems incorporating both constructs could become the norm in the near future.

References

- Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103–118.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Anderson, L. W. (1986). Opportunity to learn. In T. Husén & T. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies*. Oxford, UK: Pergamon.
- Assistive Technology Act, 29 U.S.C. §3001 *et seq.* (1998).
- Beddow, P. A. (2010). Beyond universal design: Accessibility theory to advance testing for all students. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques*. Charlotte, NC: Information Age Publishing.
- Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009). *TAMI accessibility rating matrix*. Nashville, TN: Vanderbilt University.
- Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test accessibility and modification inventory*. Nashville, TN: Vanderbilt University.
- Berliner, D. C. (1979). Tempus educare. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 120–135). Berkeley, CA: McCutchan.
- Bloom, B. S. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.
- Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*, 289–328.
- Boscardin, C. K., Aguirre-Muñoz, Z., Chinen, M., Leon, S., & Shin, H. S. (2004). *Consequences and validity of performance assessment for English learners: Assessing opportunity to learn (OTL) in grade 6 language arts* (CSE Report No. 635). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist, 41*(10), 1069–1077.
- Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.
- Brophy, J. E., & Evertson, C. M. (1978). Context variables in teaching. *Educational Psychologist, 12*(3), 310–316.
- Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*(8), 723–733.
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293–332.
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.
- Coker, H., Medley, D. M., & Soar, R. S. (1980). How valid are expert opinions about effective teaching? *Phi Delta Kappan, 62*(2), 131–149.
- Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries*. New York: Halsted Press.
- Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis, 2*(1), 7–25.
- Cushing, L. S., Clark, N. M., Carter, E. W., & Kennedy, C. H. (2005). Access to the general education curriculum for students with significant cognitive disabilities. *Teaching Exceptional Children, 38*(2), 6–13.
- Denham, C., & Lieberman, A. (Eds.). (1980). *Time to learn*. Washington, DC: National Institute of Education.
- Elliott, S. N. (2007). Selecting and using testing accommodations to facilitate meaningful participation of all students in state and district assessments. In L. Cook & C. Cahalan (Eds.), *Large scale assessment and*

- accommodations: What works?* (pp. 1–9). Princeton, NJ: Educational Testing Service.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., et al. (2010). Using modified items to test students with and without persistent academic difficulties: Effects on groups and individual students. *Exceptional Children*, 76(4), 475–495.
- Elliott, S. N., Kratochwill, T. R., & Gilbertson-Schulte, A. (1999). *Assessment accommodations checklist/guide*. Monterey, CA: CTB/McGraw-Hill [www.CTB.com].
- Elliott, S. N., Kratochwill, T. R., & McKeivitt, B. C. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology*, 39(1), 3–24.
- Elliott, S. N., Kurz, A., Beddow, P. A., & Frey, J. R. (2009, February). *Cognitive load theory: Instruction-based research with applications for designing tests*. Paper presented at the meeting of the national association of school psychologists, Boston.
- Elliott, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children*, 70(3), 349–367.
- Elliott, S. N., McKeivitt, B. C., & Kettler, R. J. (2002). Testing accommodations research and decision making: The case of “good” scores being highly valued but difficult to achieve for all students. *Measurement and Evaluation in Counseling and Development*, 35(3), 153–166.
- Feldman, E., Kim, J. S., & Elliott, S. N. (2009). The effects of accommodations on adolescents’ self-efficacy and test performance. *Journal of Special Education*. Advance online publication. doi:10.1177/0022466909353791
- Fisher, C. W., & Berliner, D. C. (Eds.). (1985). *Perspectives on instructional time*. New York: Longman.
- Fisher, C. W., Berliner, D. C., Filby, N., Marliave, R., Cahen, L., & Dishaw, M. (1980). Teaching behaviors, academic learning time, and student achievement: An overview. In C. Denham & A. Lieberman (Eds.), *Time to learn* (pp. 7–22). Washington, DC: National Institute of Education.
- Fuchs, L. S., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice*, 16(3), 174–181.
- Fuchs, L. S., Fuchs, D., & Capizzi, A. M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children*, 37(6), 1–8.
- Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., & Karns, K. M. (2000). Supplemental teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review*, 29(1), 65–85.
- Gagné, R. M. (1977). *The conditions of learning*. Chicago: Holt, Rinehart & Winston.
- Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, 19(4), 325–338.
- Gettinger, M., & Seibert, J. K. (2002). Best practices in increasing academic learning time. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (Vol. 1, pp. 773–787). Bethesda, MD: National Association of School Psychologists.
- Gibson, D., Haerberli, F. B., Glover, T. A., & Witter, E. A. (2005). The use of recommended and provided testing accommodations. *Assessment for Effective Intervention*, 31, 19–36.
- Haertel, E., & Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement*, 20(2), 119–132.
- Harnischfeger, A., & Wiley, D. E. (1976). The teaching–learning process in elementary schools: A synoptic view. *Curriculum Inquiry*, 6(1), 5–43.
- Herman, J. L., & Klein, D. C. D. (1997). *Assessing opportunity to learn: A California example* (CSE Technical Report No. 453). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Herman, J. L., Klein, D. C., & Abedi, J. (2000). Assessing students’ opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice*, 19(4), 16–24.
- Hollenbeck, K. (2002). Determining when test alterations are valid accommodations or modifications for large-scale assessment. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 395–425). Mahwah, NJ: Lawrence Erlbaum.
- Hollenbeck, K., Rozek-Tedesco, M. A., & Finzel, A. (2000, April). *Defining valid accommodations as a function of setting, task, and response*. Paper presented at the meeting of the council of exceptional children, Vancouver, BC, Canada.
- Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries*. New York: Wiley.
- Individuals with Disabilities Education Act Amendments, 20 U.S.C. §1400 *et seq.* (1997).
- Individuals with Disabilities Education Act, 20 U.S.C. §1400 *et seq.* (1975).
- Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment*, 4(2), 1–23.
- Ketterlin-Geller, L. R., Alonzo, J., Braun-Monegan, J., & Tindal, G. (2007). Recommendations for accommodations: Implications of (in) consistency. *Remedial and Special Education*, 28(4), 194–206.
- Kettler, R. J., & Elliott, S. N. (2010). Assessment accommodations for children with special needs. In B. McGaw, E. Baker & P. Peterson (Eds.), *International encyclopedia of education* (3rd ed.). Oxford: Elsevier.
- Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided

- and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, 84(4), 529–551. doi:10.1080/01619560903240996.
- Kettler, R. J., Niebling, B. C., Mroch, A. A., Feldman, E. S., Newell, M. L., Elliott, S. N., et al. (2005). Effects of testing accommodations on math and reading scores: An experimental analysis of the performance of fourth- and eighth-grade students with and without disabilities. *Assessment for Effective Intervention*, 31(1), 37–48.
- Kettler, R. J., Rodriguez, M. R., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (in press). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*.
- Kettler, R., Russell, M., Camacho, C., Thurlow, M., Geller, L. K., Godin, K., et al. (2009, April). *Improving reading measurement for alternate assessment: Suggestions for designing research on item and test alterations*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: Praeger.
- Kosciulek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Technical Report 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Kurz, A., & Elliott, S. N. (in press). Overcoming barriers to access for students with disabilities: Testing accommodations and beyond. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques*. Charlotte, NC: Information Age Publishing.
- Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *Journal of Special Education*, 44(3), 131–145. doi:10.1177/0022466909341196
- Lang, S. C., Elliott, S. N., Bolt, D. M., & Kratochwill, T. R. (2008). The effects of testing accommodations on students' performances and reactions to testing. *School Psychology Quarterly*, 23(1), 107–124.
- Lazarus, S.S., Hodgson, J., & Thurlow, M.L. (2010). *States' participation guidelines for alternate assessments based on modified academic achievement standards (AA-MAS) in 2009* (Synthesis Report 75). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Lazarus, S. S., Thurlow, M. L., Lail, K. E., Eisenbraun, K. D., & Kato, K. (2006). *2005 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 64). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Mace, R. L. (1991). *Definitions: Accessible, adaptable, and universal design (Fact Sheet)*. Raleigh, NC: Center for Universal Design, NCSU.
- McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis*, 17(3), 305–322.
- McKevitt, B. C., & Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review*, 32(4), 583–600.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
- No Child Left Behind Act, 20 U.S.C. §16301 *et seq.* (2001).
- O'Sullivan, P. J., Ysseldyke, J. E., Christenson, S. L., & Thurlow, M. L. (1990). Mildly handicapped elementary students' opportunity to learn during reading instruction in mainstream and special education settings. *Reading Research Quarterly*, 25(2), 131–146.
- Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, 7(2), 93–120.
- Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733–755). Westport, CT: Praeger.
- Pianta, R. C., Belsky, J., Houts, R., Morrison, F., & NICHD. (2007). Opportunities to learn in America's elementary classrooms. *Science*, 315, 1795–1796.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119.
- Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system*. Baltimore: Brookes.
- Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research*, 71(1), 53–104.
- Porter, A. C. (1991). Creating a system of school process indicators. *Educational Evaluation and Policy Analysis*, 13(1), 13–29.
- Porter, A. C. (1993). School delivery standards. *Educational Researcher*, 22(5), 24–30.
- Porter, A. C. (1995). The uses and misuses of opportunity-to-learn standards. *Educational Researcher*, 24(1), 21–27.
- Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–159). Mahwah, NJ: Lawrence Erlbaum.
- Porter, A. C., Kirst, M. W., Osthoff, E. J., Smithson, J. L., & Schneider, S. A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms* (Final Report). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.
- Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. (1978). *Impact on what? The importance of*

- content covered* (Research Series No. 2). East Lansing, MI: Michigan State University, Institute for Research on Teaching.
- Pullin, D. C., & Haertel, E. H. (2008). Assessment through the lens of "opportunity to learn". In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 17–41). Cambridge, MA: Cambridge University Press.
- Roach, A. T., Beddow, P. B., Kurz, A., Kettler, R. J., & Elliott, S. N. (2010). Using student responses and perceptions to inform item development for an alternate assessment based on modified achievement standards. *Exceptional Children, 77*(1), 61–84.
- Roach, A. T., Chilungu, E. N., LaSalle, T. P., Talapatra, D., Vignieri, M. J., & Kurz, A. (2009). Opportunities and options for facilitating and evaluating access to the general curriculum for students with disabilities. *Peabody Journal of Education, 84*(4), 511–528. doi:10.1080/01619560903240954
- Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools, 45*(2), 158–176.
- Rowan, B. (1998). The task characteristics of teaching: Implications for the organizational design of schools. In R. Bernhardt, C. N. Hedley, G. Cattaro & V. Svolopoulos (Eds.), *Curriculum leadership: Rethinking schools for the 21st century*. Cresskill, NJ: Hampton.
- Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *The Elementary School Journal, 105*(1), 75–101.
- Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher, 38*(2), 120–131.
- Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
- Schmidt, W. H., Houang, R. T., Cogan, L., Blömeke, S., Tatto, M. T., Hsieh, F. J., et al. (2008). Opportunity to learn in the preparation of mathematics teachers: Its structure and how it varies across six countries. *ZDM Mathematics Education, 40*(5), 735–747.
- Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457–490.
- Stevens, F. I. (1993). Applying an opportunity-to-learn conceptual framework to the investigation of the effects of teaching practices via secondary analyses of multiple-case-study summary data. *Journal of Negro Education, 62*(3), 232–248.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [2-1-11], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html>
- Thurlow, M. L., Laitusis, C. C., Dillion, D. R., Cook, L. L., Moen, R. E., Abedi, J., et al. (2009). *Accessibility principles for reading assessments*. Minneapolis, NN: National Accessible Reading Assessment Projects.
- Thurlow, M. L., Ysseldyke, J. E., Graden, J., & Algozzine, B. (1984). Opportunity to learn for LD students receiving different levels of special education services. *Learning Disability Quarterly, 7*(1), 55–67.
- U.S. Department of Education. (2007a, April). *Modified academic achievement standards: Non-regulatory guidance*. Washington, DC: Author.
- U.S. Department of Education. (2007b, revised July). *Standards and assessments peer review guidance*. Washington, DC: Author.
- Vannest, K. J., & Hagan-Burke, S. (2009). Teacher time use in special education. *Remedial and Special Education*. Advance online publication. doi:10.1177/0741932508327459.
- Vannest, K. J., & Parker, R. I. (2009). Measuring time: The stability of special education teacher time use. *Journal of Special Education*. Advance online publication. doi:10.1177/0022466908329826
- Walberg, H. J. (1986). Syntheses of research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 214–229). New York: Macmillan.
- Walberg, H. J. (1988). Synthesis of research on time and learning. *Educational Leadership, 45*(6), 76–85.
- Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*(3), 137–156.
- Ysseldyke, J. E., Thurlow, M. L., Mecklenburg, C., & Graden, J. (1984). Opportunity to learn for regular and special education students during reading instruction. *Remedial and Special Education, 5*(1), 29.

Part I

**Government Policies and Legal
Considerations**

U.S. Policies Supporting Inclusive Assessments for Students with Disabilities

2

Susan C. Weigert

U.S. Policies Supporting Inclusive Assessment of Students with Disabilities

This overview of U.S. policies supporting inclusive assessment practices traces the policies and supporting educational contexts within the historical framework of key legislation and regulations over the past 50 years.

Assessment Policies in the 1960s and 1970s: Inclusion and ‘Equal Terms’

The history of inclusive assessment policies in the United States has been a product of many political and practical influences, but at the root of their developments has been the central and fundamental tenet of equal protection. Title VI of the Civil Rights Act of 1964 prohibited discrimination on the basis of race, sex, color, or national origin, rather than on the basis of disability. Yet the spirit of the law swept students with disabilities (SWDs) into its strong political current and

promised to ensure and protect the equity of their educational opportunities. The passion for equal access to education as a civil right was best characterized by Chief Justice Warren’s opinion on *Brown v. Board of Education* in 1954:

In these days it is doubtful that any child may reasonably be expected to succeed in life if he is denied the opportunity of an education. Such an opportunity, where the State has undertaken to provide it, is a right that must be made available to all on equal terms. (*Brown v. Board of Education*, 347 U.S. 483 (1954), quoted in Russo & Osborne, 2008 p. 493)

While policies of inclusion constituted a popular solution to problems of inequity in public education, compulsory education statutes during the 1960s and early 70s left the authority to school districts to decide whether SWDs could ‘benefit’ from instruction (Russo & Osborne, 2008). The new inclusion principles were eventually codified in PL 93-112, Section 504 of the Rehabilitation Act of 1973, which prohibited discrimination against individuals with disabilities in federally funded programs, and required reasonable accommodations for students with physical or mental impairments that ‘substantially limited’ them in one or more major life activities, including learning (29 USC § 706 (7)(B)). In addition to physical and sensory handicaps, Section 504 of the Rehabilitation Act applied to persons with ‘mental’ disabilities such as mental retardation, traumatic or organic brain syndromes, emotional disturbance, specific learning

S.C. Weigert (✉)
U.S. Department of Education, Office of Special Education Programs, Alexandria, VA, USA
e-mail: susan.weigert@ed.gov

The views expressed in this chapter are solely those of the author in her private capacity. No official support or endorsement by the U.S. Department of Education is intended nor should it be inferred.

disabilities, and other cognitively disabling conditions (Phillips, 1994).

The Rehabilitation Act further defined the meaning of a free, appropriate, public education (FAPE), and specified that appropriate education included educational services designed to meet the individual education needs of students with disabilities ‘as adequately’ as the needs of non-disabled students were met. Yet the only assessment-related provisions of the Act were those requiring that assessments be provided in a child’s ‘normal mode of communication’ (including native language) unless it was clearly not feasible to do so.

The principle of inclusion in the Rehabilitation Act was later incorporated into the elementary and secondary education act (ESEA) amendments of 1974 (PL 93-380), which also mandated the ‘free appropriate public education (FAPE)’ in the ‘least restrictive environment (LRE).’ These provisions were codified a year later in the Education for All Handicapped Children Act (P.L. 94-142). Yet at the time, FAPE simply meant access to special education and related services—in conformity with individualized academic and behavioral goals stated in the student’s IEP, rather than connoting access to the general education curriculum. A new requirement for inclusive assessment of SWDs in PL 94-142 §612 (5)(C) mandated that testing and evaluative materials used for evaluation and placement not be the sole criterion for determining an appropriate educational program for a child with a disability.

The 1977 regulations amending the Rehabilitation Act of 1973 (§104.35) required that tests and other evaluation materials meet requirements for validity for the specific purpose for which the tests were used, and that they be administered by trained personnel in conformity with test-developer’s instructions. Further, the type of assessments to be used for educational evaluation of SWDs were to include those tailored to assess specific areas of educational needs, not merely those designed to measure a student’s I.Q. Finally, educational tests were to be selected and administered to ensure that the results of testing accurately reflected the student’s educational aptitude or achievement level (or

whatever educational factor was to be measured by the test), rather than merely reflecting the student’s impaired sensory, manual, or speaking skills (except when those skills were the factors measured by the test). The protections against ‘disparate impact’ of assessments for SWDs was restricted to ‘otherwise qualified’ individuals with disabilities, meaning that the student who might have impaired sensory, manual, or speaking skills, still had to be capable of meeting the standards required to pass the test. As U.S. Supreme Court justice Powell commented:

Section 504 imposes no requirement upon an educational institution to lower or to effect substantial modifications of standards to accommodate a handicapped person. (Southeastern Community College v. Davis 442 U.S. 397, 1979 Supreme Court of United States)

The 1980s and 1990s: IEP as Curriculum

As had been the case in the 1977 regulations, regulatory changes to the Rehabilitation Act in 1980 required that assessments used for college admissions constitute validated predictors of college aptitude or college achievement, rather than merely reflecting the applicant’s impaired sensory, manual, or speaking skills. Yet the 1980 regulations essentially permitted use of tests with established disproportionate adverse effects on SWDs, provided that an alternate test with a less disproportionate effect was unavailable. In 1980, when President Jimmy Carter established the Department of Education (ED) as a cabinet-level agency with a mission to ensure that educational opportunities were not denied on account of race, creed, color, national origin, or sex, disability was not included among the list of protected categories. Importantly, Section 103 (a) of the Department of Education Organization Act (PL 96-88) prohibited ED from exercising any control over the curriculum, or any program of instruction, or selection of instructional materials by any school system or educational institution.

Soon after the establishment of the new agency, Education Secretary Terrel Bell created a National Commission on Excellence in Education, which produced a report on the status of American education entitled ‘A Nation at Risk,’ which concluded that the country was threatened by a ‘rising tide of mediocrity,’ that over 10% of 17-year-olds were functionally illiterate, that SAT scores were declining across the country, and that many students required remediation courses even after entering college. The report concluded that comprehensive strategies to reform education across the country were needed (National Commission on Excellence in Education, 1983). It was believed that the needed reforms could be better comprehended after getting a fuller picture of student performance, by instituting the national assessment of all students (Ginsberg, Noell, & Plisko, 1988). During this decade and throughout the next, however, participation of SWDs in the national assessment (later, the National Assessment of Educational Progress or NAEP) was minimal (Shriner & Thurlow, 1993). In addition, most state assessment programs based their inclusion decisions for SWDs primarily upon those specified by the NAEP or on the basis of time spent in the regular classroom (Thurlow & Ysseldyke, 1993). Some factors that belied high exclusion rates on the NAEP cited by NCEO included unclear participation guidelines, sampling plans that systematically excluded students in separate schools or those not in graded classes, and an ‘altruistic’ motivation to reduce stress on students not expected to perform well (Zigmond & Kloo, 2009). Some states were simply unwilling to make accommodations to students to permit participation of SWDs in the NAEP (Ysseldyke & Thurlow, 1994a).

On the state assessment front, SWDs were included only slightly more often than on the NAEP. Shriner and Thurlow (1993) document that in the early 1990s, less than 10% of SWDs were being included in state assessments. As a consequence of widespread assessment exclusion policies for the first two decades after the establishment of the Office of Special Education Programs (OSEP), there was very little known about the academic outcomes of

SWDs (Ysseldyke, Thurlow, McGrew, & Shrine, 1994b).

The IDEA was reauthorized in 1990 as PL 101-476, and a focus remained on physical inclusion—greater inclusion in community schools, least restrictive placement of students, and transition services. Placements of SWDs in classes were to be age and grade appropriate with a minimum of placement in self-contained classrooms. Teaching methods for including SWDs in the general education classrooms began to involve cooperative learning and peer-instruction models. Yet an emphasis was placed on the physical inclusion of SWDs over the quality or effectiveness of their academic experience of SWDs (Danielson, personal communication, October 22, 2009). While teaching staff were expected to ‘adapt’ the curricular content, in doing so, they were encouraged to choose a grade-level curriculum that seemed developmentally most suited to meet each SWDs IEP objectives, rather than to ensure access to grade-level standards (Simon, Karasoff, & Smith, 1991).

IDEA 1990 also funded studies and investigations through which to collect information needed for program and system improvements by states and LEAs. The results of studies, such as the National Longitudinal Transition Study began to be available to OSEP prior to the 1997 authorization, and later shed light on the degree to which inclusion efforts were failing to ensure effective instruction of SWDs (Danielson, personal communication, October 22, 2009).

Prior to the 1993 ESEA reauthorization, Title I funds were to be distributed to schools on the basis of the poverty level and economic needs of students rather than on the basis of performance on State assessments. But the reauthorized 1993 ESEA shifted the focus to assessing outcomes for all children, including students with special needs, in key disciplines—mathematics, science, history, geography, civics, English, the arts, and other languages. The reauthorized ESEA attempted to ensure that ‘all students,’ including ‘special needs’ students, met high academic standards, that teaching and learning improved, that government offered flexibility coupled with responsibility for student

performance, that schools work cooperatively with parents and the community, and that Federal aid go to the poorest students (U.S. Department of Education, 1993).

ESEA 1993 endeavored to improve learning through reform approaches similar to those of other countries whose students were thought to be outperforming American students, particularly in the fields of science and mathematics. Thus, closely following the ESEA reauthorization of 1993 was the Goals 2000 Educate America Act (PL 103-227), which was signed into law on March 31, 1994. The essence of Goals 2000 was that by the year 2000 all students would leave grades 4, 8, and 12 with competency in English, mathematics, science, foreign languages, civics and government, economics, arts, history, and geography. Every student was to be prepared for responsible citizenship, postsecondary learning, and productive employment. The reforms of Goals 2000 were grounded in the expectation that States develop more challenging content and performance standards, design instruction and assessments aligned to those standards, and participate in accountability reporting on the extent to which schools and students were meeting the State standards (The White House, 1990; National Academy of Education, 1998). For some states, this was the first effort at trying to develop a broad framework for a general curriculum (National Academy of Education, 1998). Ultimately, under the ESEA Title I requirement, all states were expected to have valid, reliable, and aligned assessments based on their new content standards in the four core academic subjects by school year 2000–2001.

At the same time, among disability advocates, it was well understood that there was both an education gap as well as an ‘assessment gap’ for SWDs (Danielson, personal communication, October 22, 2009). The National Center on Educational Outcomes (NCEO) publicly posed the question of whether SWDs were seriously being considered in the standards-based reform movement, and pointed out that when identifying sources of data for monitoring progress toward the national goals, in 1991 the National

Education Goals Panel identified data collection programs that had excluded up to 50% of SWDs (McGrew et al., 1992). An NCEO Synthesis report ended with ‘Our nation in its quest to become first in the world has forgotten many of its students’ (Thurlow & Yesseldyke, 1993).

The Council for Exceptional Children testified to Congress in 1992 that the standards themselves should be constructed so as to accommodate all students, and it called for an investigation into alternative forms of assessments as well as ways to ensure that when educators worked on standards for assessments that at least one member be included who had expertise in working with individuals with disabilities (CEC testimony before House Subcommittee on Elem, Sec, and Vocational Education, 1992). By the time IDEA 1997 was reauthorized, most states had established content standards in the four core content areas, yet the question of which students with disabilities could access these standards, and participate in assessments based upon them, was a subject of debate. Moreover, the type of tests that was being constructed posed barriers over and above the content standards. States began moving away from flexible, or ‘authentic assessments,’ which held greater promise for inclusiveness of a range of student ability levels, in order to fulfill the pragmatic requirements of large-scale testing. Such ‘authentic’ assessments, popular during the era, were difficult to standardize across large numbers of diverse students. Moreover, while most special educators believed that performance-based assessments provided more accurate descriptions of student progress and were more helpful in informing classroom practice, their administration was expensive and overly time consuming for consideration in accountability testing.

In the midst of the standards movement, there was a pervasive concern that norm-referenced assessments were not appropriate to the goals of standards-based reforms, not just in the case of SWDs, but for all students. More importantly, norm-referenced tests were not well aligned to the curricula that students were to be taught

under the new standards movement. During the mid-1990s, there had been much concern about the fairness and the validity of norm-referenced scoring approaches for use with special populations. Many states also justified the exclusion of SWDs from standardized testing on the basis of fairness—that students had not received an opportunity to learn the material assessed on general assessments—and on the basis of utility—arguing that the results of assessment scores did not provide useful information about the academic performance or educational needs of SWDs. Advocates complained that the use of norm-referenced testing, in which SWDs were usually ranked lowest, led to the perpetuation of assumptions that SWDs were incapable of making any academic progress. Yet, excluding them from participation provided no information at all about their academic performance and, some argued, denied FAPE.

While many in the special education field were divided as to how SWDs should participate in the standards-based accountability movement of the 1980's, most later came to agree, as one state policymaker commented, that 'the removal of special education students from the "accountability track" resulted in their removal from the "curriculum track"' (Koehler, 1992).

Following Goals 2000, as most states embarked on a full-scale revision of their assessment systems and attempted to define 'what all students should know' in their new content standards, nearly all states shifted to the use of criterion-referenced assessments and began including a percentage of SWDs in these new assessments. In order to assist SWDs in accessing these new criterion-referenced assessments, States developed a list of 'standard accommodations.' The four classes of accommodations included the following: *presentation format*, which were changes in how tests were presented and involved accommodations like providing Braille versions of the tests or orally reading the directions to students; *response format*, which were changes in the manner in which students gave their responses and included accommodations such as having a student point

to a response or use a computer for responding; *setting of the test*, which could be alone, or in small groups; and finally, *timing of the test*, which could include extending the time allowed, or providing more breaks during testing.

In response to questions about the attainability of performance standards for all students with disabilities, ED advised states to implement alternative assessments for a 'small number' of students and to implement accommodations to ensure an 'equal playing-field' for those students, stating, 'Assessment accommodations help students show what they know without being placed at a disadvantage by their disability' (U.S. Department of Education, 1997). However, claims about the capacity of accommodations alone to overcome the disadvantages created by a student's disability were considered true for students with sensory, manual, or speaking skills, but not for SWDs with cognitive impairments.

The implementation of testing participation guidelines for SWDs was the subject of considerable controversy among disability advocates across states. Policy experts maintained, often based upon the Supreme Court ruling in *Southeastern Community College v. Davis* that standards could never not be lowered for SWDs taking accountability assessments, even if those assessments were also to be used to guide instruction (e.g., Philips, 2002). Yet most advocates maintained that, as students with disabilities were not included when the standards were developed, it seemed inappropriate to hold them to the standards.

Prior to the passage of the Americans with Disabilities Act of 1990 test developers were most familiar with the provision of accommodations for students with sensory impairments. However, following the passage of the ADA, advocates for the disabled argued that federal law should ensure the availability of testing accommodations and modifications for mental disabilities such as dyslexia and other learning disabilities. Yet policymakers responded again that the effect of accommodations for cognitive disabilities undermined the valid interpretation of a student's test score (e.g., Philips, 1993).

IDEA 97 and Options for Alternate Assessment

OSEP hoped, by providing funding opportunities, to spur the Research and Development Community to go beyond what had been considered technically feasible, to respond to the increasing demand for teaching tools and new approaches to the assessment of SWDs. The reauthorized IDEA of 1997 provided for the development of new assessments to both identify areas of academic need and to measure academic progress for children with disabilities. The new assessments were also to be used in educational program planning and for placement in special education, related services, and/or early intervention under § 641, (1)(G) of the law. Funds were to be made available for the development of alternative assessments for the inclusion of non-native English speakers and other minority students, to prevent misidentification of such students as SWDs. The mandates of IDEA 97 called for assessments to meet requirements that could not be entirely met by inclusion of SWDs in large-scale state assessments alone. Scientific measurement of student progress called for new types of classroom assessments to monitor academic progress (e.g., Fuchs & Fuchs, 2004), in response to intensive and evidence-based interventions. These new ‘curriculum-based’ assessments (CBM) were inherently inclusive, as they could be individualized for students working across a broad continuum of skill levels. The new CBM measures also represented a significant improvement from previous classroom assessments which relied on ‘mastery measurement,’ or summative, end-of-unit assessments, as the new measures permitted the monitoring of incremental progress over time by teachers (Stecker, 2005).

While progress-monitoring assessments permitted inclusiveness and informed decision-making in the classroom, some advocates maintained they had the disadvantage of isolating SWDs from the general education curriculum:

These assessments frequently were conducted in isolation from the larger general education curriculum. The assessments focused on immediate

and discrete skill deficits and IEPs often were a collection of isolated skill objectives that led to isolated instruction...Too often, the IEP *became* the curriculum for the student, instead of a tool for defining how to implement a general education curriculum. (Nolet & McLaughlin, 2000, p. 10)

Overall, the 1997 reauthorization was most significant for its endorsement of the participation of all children with disabilities in state assessments and for the requirement that alternate assessments be made available, by July 2000, for any of those students who could not meaningfully participate in the regular state assessment even with accommodations (U.S. Department of Education, 2000). The sanctioning of alternate assessments was to shift the views of federal and state policymakers on what constituted fair assessment practices by moving away from the principle that a single standard of performance on state standards, even for purposes of accountability, would by necessity apply to ‘all students.’

The states had concluded that, if all students were to be included in the new accountability systems, new assessments based on the standards would have to be developed for a small percentage of students with the most severe disabilities—generally students labeled with ‘severe-profound disabilities and trainable mentally handicapped’ (Quenemoen, 2009). According to Browder, & Wakeman, & Flowers (2009), prior to IDEA 97 there were three classes of SWDs: (a) those who pursued a general education curriculum with expectations for grade-level achievement, (b) those who required a remedial curriculum (e.g., a 7th grader working on 4th grade math), and (c) those who required functional life skills to prepare for independent living. While prior to 1997, teachers expected that only the first group of SWDs would participate in state assessments, after the 1997 reauthorization, the inclusion of SWDs in the new alternate assessments (Browder et al., 2009) constituted a major shift in assessment inclusion policies, permitting all subgroups of SWDs to participate with validity. The new alternate assessments were to be ‘aligned with the general curriculum standards set for all students and should not be assumed

appropriate only for those students with significant cognitive impairments' (34 CFR §200).

In spite of the inclusion accomplished by the 1997 IDEA, the disability community was torn as some advocates continued to maintain that exclusion from state assessments and substitution of measurements of progress toward IEP goals were the only appropriate responses to the new standards movement. Others contended that the substitution of IEP goals for state and National assessment participation would violate the spirit of inclusion—especially considering that IEP goals were often chosen from a list of useful skills of 'everyday life' (Browder et al., 2009), rather than being designed to provide equitable access to the full scope of the State content standards.

In response to the mandate for alternate assessments, states developed a variety of types of assessments which came in a variety of forms, including teacher checklists of functional skills, reports on progress toward IEP goals, portfolios of student work, and performance tasks moderately aligned to grade-level content standards (Thompson & Thurlow, 2000).

Significant policy input into the 1997 IDEA reauthorization came through David Hoppe, an aid to Senator Trent Lott, and a parent of a person with disabilities, who exercised a gentle touch in bringing diverse stakeholders to consensus. While the field grew to be unanimous in believing accountability and assessments based entirely on IEP rubrics did not make sense, at the same time, there continued to be debates about the validity of the scores of SWDs, especially those with cognitive impairments, who had not been exposed to the curriculum being assessed. Hoppe convinced policymakers to sidestep the partisanship in reauthorizing IDEA 97 and urged Congress to come up with a bill to please both sides of the debate (Danielson, personal communication, October 22, 2009).

The critical new elements in the 1997 IDEA amendments were accountability for inclusion of SWDs in general state and district-wide assessments, with appropriate accommodations and modifications, if necessary, and the establishment of performance goals for SWDs as a condition

of funding under IDEA Part B. In the infrequent cases when an IEP team or Section 504 team determined that standard assessments, even with reasonable accommodations, did not provide a student with an opportunity to demonstrate his or her knowledge and skill, the State or school district was to provide an alternate assessment. Yet whatever assessment approach was taken, the scores of students with disabilities were to be included in the assessment system for purposes of public reporting and school and district accountability.

The reauthorized IDEA 97 also required the consideration of assistive technology needs for participation, as well as the communication needs of children who were deaf, hard of hearing, or those with limited English language proficiency. A further requirement for assessments used for the evaluation of SWDs was the inclusion of information that was 'instructionally relevant' in the evaluation, in order to help a child become involved in and make progress in the general education curriculum. Assessment instruments used for disability evaluation were also required to be technically sound to assess the 'relative contributions' of both cognitive and behavioral factors, in addition to physical or developmental factors, on students' academic performance.

States were, once again, required to administer assessments to SWDs in a child's native language, or typical mode of communication. Importantly, any standardized tests given to a child were required to be validated for the specific purposes for which they were to be used. In addition, assessment tools and strategies that directly assisted teachers in determining the educational needs of the child were also to be provided under IDEA 97.

In response to the mandate for new alternate assessments, OSEP funded a variety of projects, such as supporting computer-adaptive assessments aligned to the state standards that would be capable of identifying, through 'dynamic' assessment techniques, learning issues of students with learning disabilities or other 'gap students' in order to uncover the instructional gaps they were manifesting in general education settings (e.g., see Tindal, 2008). It was known that

the population of students with specific learning disabilities (SLD) consisted of slow learners who could eventually address all content standards, though not necessarily in the time frame required to participate fairly in the summative end-of-year assessments. OSEP struggled with how to balance the learning needs of high-incidence SLD students with the mandate to include them in state assessments, as well as how to overcome the historical problem of low expectations. The answer OSEP arrived at to best address this was to mandate that instruction be provided by skilled teachers specifically trained to work with the SLD population. OSEP later funded the Access Center to help teachers adapt and individualize instruction aligned to standards that were appropriate for the student's grade and age-level, rather than 'out-of-level' standards, as had been a common teaching practice prior to 1997. Yet, to many advocates in the field, assessments based on standards that many SWDs could not master in the same time frame were also considered 'out-of-level' assessment, since such assessments required that a typical SLD student would need to make more than a year's worth of average progress in a year to learn enough grade-level material to be fairly assessed on the full scope of material being tested (Danielson, personal communication, October 22, 2009).

The effect of the 1997 IDEA was to shift reform efforts to the IEP, envisioning it not as a guide to what SWDs were to be learning, but rather rendering it into a tool to ensure inclusion and progress in the grade-level general education curriculum by defining each student's present level of performance, including how the student's disability affected his or her ability to be involved in and make progress in the general education curriculum. Additionally, the law required a statement in the IEP about the program modifications and supports to be used by school personnel to enable the child to be involved in and make progress in the general education curriculum and to participate with his or her non-disabled peers.

Subsequent to the IDEA Part B regulations in 1999, which mandated inclusion of all SWDs in standards-based reform programs, however, many SEAs did not succeed in ensuring that local

education agencies (LEAs) and schools taught SWDs the grade-level curriculum.

2001 No Child Left Behind Act

Under the 1994 ESEA, States were required to test only three times during a student's tenure in the K-12 educational system. For policymakers crafting the reauthorized ESEA, this left too many intervening years in which children's academic difficulties could go unaddressed, with the result that many children were being 'left behind,' academically. Under the 'No Child Left Behind Act' (NCLB) of 2001, States were obliged to enhance their existing assessment systems to include annual assessments in reading/language arts and mathematics for all public school students in grades 3 through 8 and at least once in grades 10 through 12 by the 2005–2006 school year. Additionally, by the 2007–2008 school year, all States were to annually assess their students in science at least once in grades 3 through 5, once in grades 6 through 9, and once in grades 10 through 12 (U.S. Department of Education, 2003).

The NCLB required annual testing in reading and mathematics, the demonstration of 'adequate yearly progress' against state-specified performance targets, and the inclusion of all students in annual assessments. Secretary of Education, Rod Paige, later succeeded by White House domestic policy advisor, Margaret Spellings, emphasized that the purpose of the NCLB provisions was to ensure that every child was learning 'on grade level.' The accountability for the SWD subgroup also required steps to recruit, hire, train, and retain highly qualified personnel, research-based teaching methods, and the creation of improvement programs to address local systems that fell short of performance goals.

During the same year, President Bush created the President's Commission on Excellence in Special Education, a program designed to improve the dropout rate among SWDs, who were leaving school at twice the rate of their peers, and whose enrollment in higher education

was 50% lower. Moreover, the SLD subgroup had grown over 300% since 1976, and 80% of those with SLD reportedly had never learned to read (President's Commission on Education, 2002). The claim was made that few children in special education were closing the achievement gap to a point where they could read and learn like their peers. A major thrust of the Commission was that although special education was based in civil rights and legal protections, most SWDs remained at risk of being left behind. Several findings of the Commission included criticisms that the reauthorized 1997 IDEA placed process above results and compliance above student achievement and outcomes. Further, Special Education did not appear to guarantee more effective instruction. The identification of students for special education services was criticized for being based upon a 'wait-to-fail' model. The criticism was launched that ED had become two separate systems instructionally, when it was critical that general education and special education share responsibilities for the education of SWDs. Among the recommendations of the report was a call for improved assessment policies to prevent exclusion from State and district-wide assessments, still a common practice in 2001.

2002–2003 Title I Regulations Permitting Alternate Achievement Standards in Accountability

The ESEA regulations of 2002 implementing the assessment provisions of NCLB authorized the use of alternate assessments in accountability assessments and required that States make available alternate assessments for any student unable to participate in the State's general assessments, even with accommodations. The subsequent ESEA regulations of 2003 permitted states to develop alternate achievement standards for students with the most significant cognitive disabilities. The 2003 regulations required that the alternate assessment be aligned with the State's academic content standards, promote access to the general curriculum, and reflect professional judgment of the highest achievement

standards possible (34 CFR§200.1). These regulations effectively forced most states to begin to revise the assessments they had originally created in response to the more liberal 1997 IDEA requirements. While the due date for the development of alternate assessments was 2000, there was little knowledge in the field of the appropriate academic content to base such tests on. While there had been some early work on how to teach general education curriculum content to students with severe disabilities (e.g., Downing & Demchak, 1996), the mandate for alternate academic achievement standards aligned to the state's academic content standards was to become a critical force in transforming the curriculum for SWDs with severe disabilities. Over the next decade, problems with developing a coherent academic curriculum appropriate to the diverse population of students with significant cognitive disabilities were prevalent. Among the most significant contributing factors in the delay in developing valid and aligned alternate assessments under the NCLB was the belief among special educators charged with developing the assessments that grade-level academic content standards were not relevant to these students, and that the appropriate content for these students consisted of 'life-skills' for independent living (Wallace, Ticha, & Gustafson, 2008). In response to this, ED set new standards for technical adequacy that alternate assessments were required to meet. Over the next decade, while the quality of many alternate assessments on alternate achievement standards (AA-AAs) improved, by 2010, many states still did not have technically adequate, peer-reviewed alternate assessments. At the same time, research carried out in the field indicated that students eligible for the AA-AAS could use symbolic communication systems and could learn to read and reason mathematically (Browder, Wakeman, Spooner, Ahlgrim-Delzell, & Algozzine, 2006; Kearns, Towles-Reeves, E., Kleinert, H. L., & Kleinter, 2009).

Students with significant cognitive disabilities continue to be inappropriately excluded from participation in alternate assessments, although one study of excluded students found that many could be using augmentative and assistive technologies

to speak or otherwise communicate, and that approximately 75% were learning sight words and using calculators to perform mathematical calculations (Kearns et al., 2009). Kearns recommends that to meaningfully include the population of students with significant cognitive impairment in state assessments, future assessments must include ‘authentic’ demonstrations of skills and knowledge aligned to the grade-level content standards, within the assessment context such students require. Scaffolding may be required for some within this population to enable them to show what they know and can do. The inclusiveness of alternate assessments for students with significant cognitive impairment depends largely upon whether or not the teacher has an understanding of how to teach a standards-based curriculum appropriate to the students within this diverse population (Kearns et al., 2009).

An additional problem with the implementation of alternate assessments for students with the most significant cognitive impairments has been the continued difficulty of establishing appropriately challenging achievement standards for the full range of ability levels manifested in this population. While states are permitted to set more than one achievement standard in order to make an alternate assessment more inclusive, teachers and schools have been cautious about assigning students to the more challenging achievement standard on a test used for making school accountability determinations—often assigning both higher- and lower-functioning students in the 1% population to the lowest achievement standard, to safeguard a favorable outcome in accountability under NCLB. As a result of these widespread practices, the proficiency rates on alternate assessments have been much higher compared to proficiency rates of SWDs taking the general assessment across the majority of states, suggesting that alternate assessments are simply not challenging enough for the majority of students taking them.

Currently available alternate assessments vary in the extent to which they inform parents and teachers about student academic progress. While advances in the development of classroom-based formative assessments, such as CBM, have been

widely used among students with high-incidence disabilities since IDEA 1997, comparable formative assessments have rarely if ever been available for teachers of students in the 1% population, though OSEP encouraged their development (but see Phillips et al., 2009). Assessment measures designed to measure a student’s level of mastery and progress on both prerequisite and grade-level content and skills at a challenging level for such students would greatly assist and support instruction of students with significant cognitive disabilities and help to demonstrate their capacity for progressing in an academic curriculum. Fundamentally, the central problem with alternate assessments has arisen from the need to understand and develop an appropriate academic curriculum for eligible students—one that is aligned to the same grade-level content standards intended for all students, yet reflects content and skills at an appropriate level of difficulty for each unique student, and is also capable of indicating progress toward an expected level of knowledge and skill.

IDEA 2004 and Assessments Measuring Responsiveness to Intervention

The reauthorization of the IDEA 2004 (PL 108-446) reiterated the NCLB mandate for inclusion of all SWDs in State and district-wide assessments, and clarified that IEP goals were not to be the only form of assessment. The most important assessment-related changes made in the IDEA 2004 reauthorization were those that pertained to requirements for eligibility determinations of SLD. These changes permitted states to move away from the IQ-performance discrepancy model of SLD identification that had been the subject of criticism in the Commission report. Under IDEA 2004, new assessments were to be developed for the purpose of assisting with the identification of students with SLD through the assessment of a student’s response to tiered, evidence-based instructional interventions (RTI). Such assessments were to be of several types, those to screen students in

basic skills (e.g., literacy or mathematics), and those to help define and monitor responsiveness to evidence-based interventions designed to help the child progress in significant areas of academic weakness and to inform a decision about the need for more intensive remedial instruction.

2007 Joint Title I IDEA Regulations Permitting Modified Academic Achievement Standards in Accountability

Nearing the end of the George W. Bush presidential term, Secretary of Education Margaret Spellings grew concerned at the slow pace with which states were advancing to the goal of ‘universal proficiency’ by 2014 intended by the NCLB. States and the special education community expressed concerns that a small group of SWDs who were enrolled in general education classes were, due to the nature of their disabilities, unable to demonstrate the same extent of academic progress on their state’s general assessment by the end of the school year. In 2006, the administration responded by announcing a new assessment policy option to provide states with the flexibility to include the scores of ‘persistently low-performing’ students with disabilities in alternate assessments based on modified academic achievement standards (AA-MAS). While states had struggled to improve teaching practices for a subgroup of low-performing students, both those with and without disabilities, it was widely accepted that some students, because of the effects of a disability, required additional time to learn content standards to mastery. State rationales for the new assessment included a need to develop academic achievement standards inclusive of this small group of SWDs, the majority of whom were enrolled in general education classes, yet had cognitive impairments such as mental retardation, autism, specific learning disability, or other health impairment for which accommodations alone could not ensure an ‘equal playing field.’

After the publication of the 2007 Joint IDEA Title I regulations, the field became split over

the potential consequences of permitting states to assess a portion of SWDs against a lower standard of performance. Concerns of some advocates in the disability community centered on the potential lowering of standards and educational tracking of students with SLD predominantly for the sake of making ‘Adequate Yearly Progress’ toward assessment performance targets specified under Title I accountability. States differed in the degree to which they included the SLD population in the new modified assessments—some argued that the SLD population in particular had been receiving below grade-level instruction and that the availability of such a test would unnecessarily lower academic expectations for these students, who, by definition, have normal or above average intellectual abilities. Others argued that problems manifested by this group of students in accessing general assessment items and their inability to master the expected content in the same time frame constituted the direct effects of their disabilities. They maintained that the new modified assessments, especially if used as a temporary measure, could help to illuminate the academic progress of this group of students, to help ensure that they received instruction aligned to grade-level content standards.

Subsequent to substantial investments by ED in the development of the AA-MAS, many states began investigating the population of SWDs who were persistently low achieving and conducted item accessibility and ‘cognitive lab’ studies to develop more accessible test items. Elliot et al., (2010) reported the results of a study in four states that indicated that certain modifications made to general test items improved the accessibility of the items and improved the accuracy of measurement for SWDs eligible for the AA-MAS. Modifications to regular test items did not change the content to be tested nor did they reduce the complexity of the test item (‘depth of knowledge’). The modifications made to test items were straightforward—including removing unnecessary words, simplification of language, addition of pictures and graphics, breaking long paragraphs into several, and bolding of key words. The study showed that an additional boost in student performance on the modified items

occurred when reading support (audio versions of text, except for the key vocabulary words being tested) was added to item modification. Results also showed a large boost in mathematics performance for the eligible group (Elliott et al., 2010).

‘Race To The Top’ Assessment Initiatives

In 2009, during the initial months of the Obama administration, a ‘new generation’ of standards and assessments was envisioned by which policymakers hoped all American students would become more competitive in the global marketplace. Many of the goals of the RTT assessment initiative reiterated those of the Goals 2000 era in this research. The RTT initiative endeavors to create assessments aligned to ‘fewer, higher, and clearer’ state standards held in common by most states, as well as becoming more inclusive and informative for students who typically perform at lower achievement standards, including SWDs.

In the words of Secretary Duncan,

The majority of students with disabilities take the regular state tests based on the state’s standards for all students, with appropriate accommodations to ensure that their results are valid. Students with the most significant cognitive disabilities can take alternate tests based on alternate standards and other students with disabilities may take an alternate test based on modified standards. (Duncan, 2010).

Our proposal would continue to hold schools accountable for teaching students with disabilities but will also reward them for increasing student learning. . . The Department plans to support consortia of states, who will design better assessments for the purposes of both measuring student growth and providing feedback to inform teaching and learning in the classroom. All students will benefit from these tests, but the tests are especially important for students with disabilities. (Duncan, 2010).

Policies supportive of inclusive assessments under the new administration promise to support the development of a new generation of assessments aligned to a range of achievement levels required for inclusion of all SWDs and to include measures of growth for students who make

progress at different rates. The history of inclusion in standards-based assessments shows that thoughtful inclusion decisions promote meaningful academic progress for SWDs. In the era of Goals 2000, 32% of SWDs graduated high school with a regular diploma, compared to nearly 60% of students in 2007. In 1987, only one in seven SWDs enrolled in postsecondary education, compared to a third of students in 2007 (Duncan, 2010). The challenge for growth-based accountability is to ensure that conceptions of growth are academically meaningful, in addition to being ‘statistically’ meaningful. To demonstrate consequential validity, the new assessments must provide results that specify clearly the knowledge and skills at each grade level that the student has attained and those requiring additional focus or individualized instructional intervention. Promising research and development since 2008 has resulted in empirically based assessment tools which can help to specify the accessibility of test items and to separate construct-relevant from irrelevant or extraneous cognitive processing demands of test items (Elliott, Kurz, Beddow, & Frey, 2009). To maximize inclusion of SWDs in the next generation of state assessments, a renewed commitment will be required by the disability community, to build upon the lessons learned over the history of inclusive assessment, and to methodically investigate and empirically establish expectations for their academic achievement and progress. Maximal inclusiveness in assessment will always be a function of successful inclusion in the general education curriculum, and of empirically-derived expectations for academic achievement and progress.

Acknowledgments The author gratefully acknowledges the contribution and support of Lou Danielson, who provided his unique historical perspective on assessment-related events during the first 40 years of OSEP.

References

- Browder, D., Wakeman, S., & Flowers, C. (2009). Which came first, the curriculum or the assessment? In W. D. Shafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: policy, practice, and potential*. Baltimore, MD: Paul H Brookes Publishing Co.

- Browder, D., Wakeman, S., Spooner, F., Ahlgrim-Delzell, L., & Algozzine, B. (2006). Research on reading instruction for individual students with significant cognitive disabilities. *Exceptional Children, 72*, 392–408.
- Brown v. Board of Education, 347 U.S. 483 (1954).
- Council for Exceptional Children. (1992). *Statement prepared for testimony to the house subcommittee on elementary, secondary, and vocational education*. Reston, VA: Author.
- Council for Exceptional Children. (1992). *Statement prepared for testimony to the technical advisory panel on uniform national rules for NAEP testing for students with disabilities*. Reston, VA: Author.
- Downing, J.E., & Demchak, M. (1996). First steps: Determining individual abilities and how best to support students. In J. E. Downing (Ed.), *Including students with severe and multiple disabilities in typical classrooms: Practical strategies for teachers* (pp. 35–61). Baltimore, MD: Paul H. Brookes Publishing Co.
- Duncan, A. (2010). *Keeping the promise to all America's children*. Remarks made to the council for exceptional children, April 21, Arlington, VA, www.ed.gov
- Elementary and Secondary Education Act of 1965, PL 89-10 (1965).
- Elliott, S. N., Kurz, A., Beddow, P., & Frey, J. (2009, February 24). *Cognitive load theory: Instruction-based research with applications for designing tests*. Paper presented at the annual convention of the national association of school psychologists, Boston.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., et al. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children, 76*(4), 475–495.
- Fuchs, L. S., & Fuchs, D. (2004). Determining adequate yearly progress from kindergarten through grade 6 with curriculum-based measurement. *Assessment for Effective Instruction*.
- Ginsberg, A. L., Noell, J., & Plisko, V. W. (1988). Lessons from the wall chart. *Educational Evaluation and Policy Analysis, 10*(1), 1–10.
- Hehir, T. (2005). *New directions in special education: Eliminating Ableism in policy and practice*. Cambridge, MA: Harvard Education Press.
- Kearns, J. F., Towles-Reeves, E., Kleinert, H. L., & Kleinter, J. (2009). Who are the children who take alternate achievement standards assessments? In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential*. Baltimore, MD: Paul H Brookes Publishing Co.
- Koehler, P. D. (1992). Inclusion and adaptation in assessment of special needs students in Arizona. In M. L. Thurlow & J. E. Yesseldyke (Eds.) (1993), *Can "all" ever really mean "all" in defining and assessing student outcomes?* (Synthesis Report No. 5). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- McGrew, K. S., Thurlow, M. L., Shriner, J. G., & Spiegel, A. N. (1992). *Inclusion of students with disabilities in national and state data-collection programs* (Technical Report 2). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- National Academy of Education. (1998). *Goals 2000: Reforming education to improve student achievement*. Washington, DC: National Academy of Education.
- National Commission on Excellence in Education. (1983). *A Nation at risk: The imperative for educational reform*. Washington, DC: Government Printing Office.
- Nolet, V., & McLaughlin, M. J. (2000). *Assessing the general curriculum: Including students with disabilities in standards-based reform*. Thousand Oaks: Corwin Press.
- Phillips, S. E. (1993). Testing accommodations for disabled students. *Education Law Reports, 80*(9).
- Phillips, S. E. (1994). High stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*(2), 93–120. Lawrence Erlbaum, Associates, Inc.
- Phillips, S. E. (2002). Legal issues affecting special populations in large-scale assessment programs. In G. Tindal & Thomas M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Phillips, G. W., Danielson, L., & Wright, L. (2009). *Psychometric advances in alternate assessments*. Washington, DC: American Institute for Research.
- President's Commission on Excellence in Special Education. (2002). *A new Era: Revitalizing special education for children and their families*. Washington, DC: Author
- Quenemoen, R. (2009). The long and winding road of alternate assessments: Where we started, where we are now, and the road ahead. In William D. Schafer & Robert W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential*. Baltimore, MD: Paul H Brookes Publishing Co.
- Russo, C., & Osborne, A. (Eds.). (2008). *Essential concepts and school-based cases in special education law*. Thousand Oaks: Corwin Press.
- Shriner, J., & Thurlow, M. L. (1993). *State special education outcomes: A report on state Activities at the end of the century*. Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.
- Simon, M., Karasoff, P., & Smith, A. (1991). *Effective practices for inclusion programs: A technical assistance planning guide*. Unpublished paper supported by U.S. Department of Education Cooperative Agreements #GOO87C3056-91 and #GOO87C3058-91.
- Skirtic, T. M. (1991). The special education paradox: Equity as the way to excellence. *Harvard Educational Review, 61*(2), 148–206.
- Southeastern Community College v. Davis, 442 U.S. 397, 413 (1979).
- Stecker, P. M. (2005). Monitoring student progress in individualized educational programs using curriculum-based measurement. In U.S. Department of Education, Office of Special Education: IDEAS that Work:

- Toolkit on teaching and assessing students with disabilities. National Center on Student Progress Monitoring.
- The White House. (1990). *National educational goals, office of the press secretary*. Washington, DC: Author.
- Thompson, S. J., & Thurlow, M. L. (2000). *State alternate assessments: Status as IDEA alternate assessment requirements take effect* (Synthesis Report 35). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M. L., & Ysseldyke, J. E. (1993). *Can "all" ever really mean "all" in defining and assessing student outcomes?* (Synthesis Report No.5). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Tindal, G. (2009). Reflections on the alternate assessment in Oregon. In W. D. Shafer, & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential*. Baltimore, MD: Paul H Brookes Publishing Co.
- U.S. Department of Education. (1993). *The reauthorization of the elementary and secondary education act, executive summary*. Washington, DC: U.S. Department of Education.
- U.S. Department of Education. (1997). *Elementary and secondary education. Guidance on standards, assessments and accountability*. Retrieved October 2, 2009 from http://www2.ed.gov/policy/elsec/guid/standardsassessment/guidance_pg4.html#disabilities3
- U.S. Department of Education. (1998). *Goals 2000: Reforming education to improve student achievement, April 30, 1998*. Washington, DC: U.S. Department of Education.
- U.S. Department of Education. (2000). *Summary guidance on the inclusion requirement for Title I final assessments*. Retrieved March 13, 2010 from <http://www2.ed.gov/policy/elsec/guid/inclusion.html>
- U.S. Department of Education. (2003). *Standards and assessments non-regulatory guidance March 10, 2003*. Washington, DC: Author.
- Wallace, T., Ticha, R., & Gustafson, K. (2008). *Study of general outcome measurement (GOMs) in reading for students with significant cognitive disabilities: Year 1* (RIPM Technical Report #27). Minneapolis, MN: Research Institute on Progress Monitoring, University of Minnesota.
- Ysseldyke, J., & Thurlow, M. L. (1994a). *Guidelines for inclusion of students with disabilities in large-Scale assessments* (Policy Directions # 1). Minneapolis: National Center on Educational Outcomes, University of Minnesota.
- Ysseldyke, J., Thurlow, M. L., McGrew, K. S., & Shriner, G. J. (1994b). *Recommendations for making decisions about the participation of students with disabilities in statewide assessment programs: A report on a working conference to develop guidelines for statewide assessments and students with disabilities* (Synthesis Report 15). Minneapolis, MN: National Center on Educational Outcomes.
- Zigmond, N., & Kloof, A. (2009). The 'two percent students': Considerations and consequences of eligibility decisions. *Peabody Journal of Education*, 84(4), 478–495.

U.S. Legal Issues in Educational Testing of Special Populations

3

S.E. Phillips

Introduction

Legal challenges to educational testing programs have focused primarily on the issues of adverse impact, parental rights, testing irregularities, non-standard test administrations, and testing English language learners (ELLs).¹ However, once a case goes to trial, issues of validity, reliability, passing standards, and adherence to other professional standards may be raised.² This chapter focuses on the legal, psychometric, and policy issues related to educational testing of special populations. Although prior federal cases and professional standards provide guidance, some tough policy decisions remain.

Nonstandard Test Administrations

State-mandated tests are typically administered under standard conditions that are the same for all students. The purpose of standard test administration conditions is to produce test scores that have comparable interpretations across students,

classrooms, schools, and districts. However, test administrators routinely receive requests for testing alterations for students with disabilities and ELLs. For example, a reader, calculator, word processor, extended time, or a separate room may be requested for students with reading, mathematics, writing, or other learning disabilities. Tests administered under such altered conditions are referred to as nonstandard test administrations.

The term *testing accommodation* refers to a nonstandard test administration that provides access for persons with disabilities while preserving the content and skills intended to be measured and producing comparable scores. Access involves the removal of irrelevant factors that may interfere with valid measurement of the intended content and skills. The validity of test scores as accurate measures of the intended content and skills is improved when the effects of irrelevant factors are removed but is reduced when relevant factors are altered or eliminated. When the intended inference from a test score is the attainment of a specified domain of content and skills, nonstandard test administrations that provide access to direct assistance with the tested content and skills may increase test scores, but those scores will lack validity for their intended interpretation.

The domain of content and skills intended to be measured by a test is referred to as the tested construct. For achievement tests, the tested construct (e.g., reading, mathematics) is defined by the content standards, objectives, and test

S.E. Phillips (✉)
Consultant, Mesa, AZ 85205, USA
e-mail: sephillips2@aol.com

¹ Portions of this chapter were adapted from Phillips (2010), Phillips (2006), and Phillips (2002).

² The APA/AERA/NCME 1999 *Standards for Educational and Psychological Testing*. In this chapter, the document is referred to as the *Test Standards*, and individual Standards are referenced by number.

specifications used to develop the test. Typically, the test specifications for an achievement test consist of a two-dimensional matrix of content by cognitive difficulty with the corresponding number of included test items given in each cell. A nonstandard test administration preserves the content and skills intended to be measured (construct preservation) when the resulting scores can be interpreted as indicators of attainment of the same domain of content and skills as specified for a standard test administration. Construct preservation means that the same content has been measured at the same level of cognitive difficulty.

In addition to construct preservation, score comparability is an essential characteristic of a testing accommodation. Comparable test scores are equivalent and interchangeable and have the same intrinsic meaning and interpretation. When test scores from standard and nonstandard test administrations are comparable, users are able to make the same inferences about the degree of attainment of the content and skills in the construct domain from identical scores. In essence, students who obtain comparable scores from standard and nonstandard test administrations know and can do the same things.

The evidence used to evaluate construct preservation and score comparability for nonstandard test administrations must be related to the purpose of the test for the intended audience. Such evidence may be logical (judgmental) and/or empirical evidence related to test score validity, reliability, and interpretation (*Test Standards*, p. 10; Standards 4.10 and 9.7). For achievement tests, construct preservation is typically evaluated via content validity evidence. Content validity evidence is obtained by asking content experts (usually teachers of the tested subject) to judge whether each test item is an appropriate measure of its designated content standard or objective for the intended grade level (*Test Standards*, pp. 10–11; Standards 1.2, 1.6 & 1.7). Positive content validity judgments support an inference that the test is measuring the intended construct at an appropriate level of cognitive difficulty.

For a nonstandard test administration, evidence of construct preservation can be obtained

by asking content experts to judge the fidelity of the altered items and/or testing conditions to the intended domain of content and cognitive difficulty applicable to the test administered under standard conditions. Information from relevant test statutes, administrative regulations, minutes of board meetings, item-writing guides, and test administration manuals for standard test administrations may also provide helpful guidance for content experts' judgments of preservation of the intended construct.

If content experts provide content validity judgments that differ for a standard versus a nonstandard administration, these judgments are logical evidence that the tested construct is different for the two test administrations and that numerically equivalent scores for the two test administrations should be interpreted as indicating attainment of qualitatively different content and skills and/or cognitive difficulty. Alternatively, if content experts provide content validity judgments indicating that a nonstandard test administration is measuring the same content but at a different level of cognitive difficulty, comparable scores may be obtained by statistically linking the test scores from the nonstandard administration to those from a standard test administration (Standard 4.10). Empirical evidence of the reliability of standard and nonstandard test administrations and comparisons between students with disabilities and low-achieving regular education students who have tested under standard and nonstandard conditions may also be useful for evaluating construct preservation and score comparability.

Unfortunately, the term *testing accommodation* has been used by some educators and policymakers to refer to any testing alteration provided to students with disabilities or ELLs during test administration. This is unfortunate because some alterations in testing conditions do not fit the legal or psychometric definitions of an accommodation. These distinctions are discussed in more detail below. Relevant federal legislation is reviewed followed by the analysis of a series of issues related to the decisions state and district testing staff must make to create a legally and psychometrically defensible nonstandard test

administrations policy. These issues highlight the trade-offs between competing policy goals advocated by different constituencies. In some cases, these competing policy goals cannot all be achieved simultaneously, so policymakers must prioritize goals and make difficult decisions that are consistent with the purpose(s) of the state or district tests, the state or district content standards, and the intended score interpretations.

Federal Legislation

There are three major federal statutes with specific provisions for persons with disabilities that are relevant to decisions about nonstandard test administrations. They include Section 504 of the Rehabilitation Act (1973), the Americans with Disabilities Act (ADA, 1990), and the Individuals with Disabilities Education Act (IDEA, 1991). Congress passed these statutes to correct serious abuses brought to its attention during hearings about the treatment of people with disabilities. For example, the IDEA was intended to provide educational services to students with disabilities who had been excluded, ignored, or inappropriately institutionalized by the educational system. Section 504 addressed discrimination by recipients of federal funding who, for example, refused to hire persons with disabilities even when the disability was unrelated to the skills required for the job. The ADA extended the protection against discrimination due to a disability to private entities. When Congress passed disability legislation, it was particularly concerned about mandating barrier-free access to facilities open to the public. However, all three federal disability laws also included provisions relevant to cognitive skills testing.

Section 504 of the Rehabilitation Act

Section 504 provides that no otherwise qualified disabled person shall be denied participation in or the benefits of any federally funded program solely due to the person's disability. In *Southeastern Community College v.*

Davis (1979), the U.S. Supreme Court defined *otherwise qualified* as a person who, despite the disability, can meet all educational or employment requirements. In that case, the Court held that the college was not required to modify its nursing program to exempt a profoundly hearing-impaired applicant from clinical training. The Court was persuaded that the applicant was not otherwise qualified because she would be unable to communicate effectively with all patients, might misunderstand a doctor's verbal commands in an emergency when time is of the essence, and would not be able to function in a surgical environment in which required facial masks would make lip reading impossible.

The *Davis* decision clearly indicated that an educational institution is not required to lower or substantially modify its standards to accommodate a disabled person, and it is not required to disregard the disability when evaluating a person's fitness for a particular educational program. The Court stated

Section 504 by its terms does not compel educational institutions to disregard the disabilities of [disabled] individuals or to make substantial modifications in their programs to allow disabled persons to participate. . . . Section 504 indicat[es] only that mere possession of a [disability] is not a permissible ground for assuming an inability to function in a particular context. (p. 413, 405)

Thus, a critical aspect in evaluating a requested nonstandard test administration turns on the interpretation of "substantial modification of standards." In a diploma testing case involving students with disabilities, *Brookhart v. Illinois State Bd. of Educ.* (1979), the court listed Braille, large print, and testing in a separate room as accommodations mandated by Section 504. However, paraphrasing the *Davis* decision, the *Brookhart* court provided the following additional interpretive guidance

Altering the content of the [test] to accommodate an individual's inability to learn the tested material because of his [disability] would be a "substantial modification" as well as a "perversion" of the diploma requirement. *A student who is unable to learn because of his [disability] is surely not an individual who is qualified in spite of his [disability].* (p. 184, [emphasis added])

This language in the *Brookhart* opinion indicated that the federal courts were willing to draw a line between format changes (*reasonable accommodations*) and substantive changes in test questions (*substantial modifications*).

The meaning of otherwise qualified was further explained in *Anderson v. Banks* (1982). In that case, students with severe cognitive disabilities in a Georgia school district, who had not been taught the skills tested on a mandatory graduation test, were denied diplomas. The court held that when the disability is extraneous to the skills tested, the person is otherwise qualified; but when the disability itself prevents the person from demonstrating the required skills, the person is not otherwise qualified. Using this definition of otherwise qualified, the *Anderson* court reasoned that the special education students who had been denied diplomas were unable to benefit from general education because of their disabilities. The court further reasoned that this should not prevent the district from establishing academic standards for receipt of a diploma. The fact that such standards had a disparate impact on students with disabilities did not render the graduation test unlawful in the court's view. The court stated

[I]f the [disability] is extraneous to the activity sought to be engaged in, the [person with a disability] is "otherwise qualified." ... [But] if the [disability] *itself* prevents the individual from participation in an activity program, the individual is not "otherwise qualified." ... To suggest that ... any standard or requirement which has a disparate effect on [persons with disabilities] is presumed unlawful is farfetched. The repeated use of the word "appropriate" in the regulations suggests that different standards for [persons with disabilities] are not envisioned by the regulations. (pp. 510–511, emphasis in original)

In *Bd. of Educ. of Northport v. Ambach* (1982), a New York Supreme Court justice concluded, "The statute merely requires even-handed treatment of the [disabled and nondisabled], rather than extraordinary action to favor the [disabled]" (p. 684).

Americans with Disabilities Act (ADA)

The ADA, which uses the phrase *qualified individual with a disability* in place of *otherwise*

qualified [disabled] individual, requires persons with disabilities to be given *reasonable accommodations*. Consistent with Section 504 cases, the legal requirement to provide reasonable accommodations for cognitive tests refers only to those variations in standard testing conditions that compensate for factors that are extraneous to the academic skills being assessed. Consistent with Section 504 case law, the ADA does not require nonstandard test administrations that substantially modify the tested skills.

Individuals with Disabilities Education Act (IDEA)

Although the IDEA and its predecessor, the Education for All Handicapped Children Act (EAHCA), clearly mandated specialized and individualized education for students with disabilities, the federal courts have held that federal law does not guarantee any particular educational outcome. Thus, under federal law, students with disabilities are guaranteed access to a free, appropriate, public education that meets their needs in the least restrictive environment, but not specific results (*Bd. of Educ. v. Rowley*, 1982) or a high school diploma (*Brookhart*, 1983). The *Brookhart* court also held that because students were required to earn a specified number of credits, complete certain courses mandated by the state, and pass the graduation test, the graduation test was "*not* the sole criterion for graduation" (p. 183).

A student with a disability who has received appropriate educational services according to an individualized education program (IEP),³ but who is unable to master the skills tested on a graduation test, may be denied a high school diploma without violating the IDEA. However, when appropriate, federal regulations do require good faith efforts by the educational agency

³ An individualized education program (IEP) is a written document constructed by a team of professionals to address the educational needs of a special education student. The IDEA mandates a separate IEP for each special education student and includes procedural requirements for its development and implementation.

to teach the tested skills to students with disabilities. Federal case precedents also indicate that an IDEA challenge to a graduation testing requirement for students with disabilities will be unlikely to succeed if professional testing standards have been satisfied.

Professional Standards

Three national professional organizations, the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have collaborated to produce consensus *Standards for Educational and Psychological Testing* (1999), referred to in this chapter as the *Test Standards*. Courts routinely recognize the *Test Standards* as an appropriate source of guidance and support for professional opinions in a testing case. Although the *Test Standards* are aspirational rather than prescriptive, judges tend to be skeptical of expert opinions that seriously conflict with reasonable interpretations of the *Test Standards*. The most appropriate edition of the *Test Standards* for evaluating a specific test is the edition in effect at the time the test was constructed and administered. Most of the cases discussed in this chapter involved tests constructed and administered when the 1985 or 1999 editions were current. Unless otherwise indicated, references to the *Test Standards* are to the most recent 1999 edition.

Introductory material in the 1999 *Test Standards* supports the exercise of professional judgment when evaluating tests:

Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. Specific circumstances affect the importance of individual standards, and individual standards should not be considered in isolation. Therefore, evaluating acceptability involves (a) professional judgment that is based on a knowledge of behavioral science, psychometrics, and the community standards in the professional field to which the tests apply; (b) the degree to which the intent of the standard has been satisfied by the test developer and user;

(c) the alternatives that are readily available; and (d) research and experiential evidence regarding feasibility of meeting the standard. (p. 4)

A construct is a skill, such as reading comprehension or mathematics computation, that is measured by a test. The *Test Standards* distinguish between factors intended to be measured by a test (construct-relevant factors) and factors extraneous to the construct intended to be measured (construct-irrelevant factors). When determining the appropriateness of nonstandard administrations, the *Test Standards* emphasize the importance of considering the construct validity of the inference from the test score the user wishes to make. The *Test Standards* indicate that the knowledge and skills intended to be measured (construct-relevant factors) should be preserved, but construct-irrelevant factors should be eliminated to the extent feasible. Standard 10.1 states

In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct [knowledge and skills] rather than any disabilities and their associated characteristics extraneous to the intent of the measurement. (p. 106)

The *Test Standards* distinguish between comparable and noncomparable scores in the context of determining when it is appropriate to place an identifying notation (flag) on test scores obtained from nonstandard administrations. Standard 10.11 states

When there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores.

Comment: . . . If a score from a modified administration is comparable to a score from a [standard] administration, there is no need for a flag. Similarly, if a modification is provided for which there is no reasonable basis for believing that the modification would affect score comparability, there is no need for a flag. . . . [I]f a nonstandard administration is to be reported because evidence does not exist to support score comparability, then this report should avoid referencing the existence or nature of the [student's] disability and should

instead report only the nature of the [modification] provided, such as extended time for testing, the use of a reader, or the use of a tape recorder. (p. 108)

Terminology

In the *Test Standards*, the terms *accommodation* and *modification* are used interchangeably to refer to nonstandard test administrations. But as indicated in Standard 10.11, the *Test Standards* do distinguish between nonstandard test administrations that produce comparable scores and those that do not (or for which evidence of comparability is lacking). For convenience in distinguishing between these two types of nonstandard test administrations, this author has urged testing programs to use the term *accommodation* for the former and *modification* for the latter.

Testing programs and test users routinely make decisions about whether scores obtained from nonstandard test administrations preserve the construct(s) intended to be tested and should be interpreted as comparable to scores obtained from standard test administrations. Many have found it helpful in communicating with students, parents, educators, professionals, policymakers, and the public to have different words to describe nonstandard test administrations that they judge do and do not result in comparable scores. This distinction has allowed them to explain more clearly to others why some scores count for satisfying requirements such as graduation testing while others do not.

The recommendation to use the term *accommodation* as a shorthand referent for nonstandard test administrations that produce *comparable* scores and the term *modification* as a shorthand referent for nonstandard test administrations that result in *noncomparable* scores is consistent with the plain English meaning of these terms. According to the *American Heritage Dictionary*, accommodate means to adapt or adjust while modify means to change in form or character. Making an *adjustment* for a paraplegic student who needs a taller table to make room for a wheelchair during the administration of a cognitive test is typically judged by psychometricians to produce a comparable score. On the other

hand, *changing* the construct of reading comprehension to listening comprehension by providing a reader for a reading test is generally viewed by psychometricians as producing noncomparable scores. Thus, the use of the term *accommodation* for the former nonstandard test administration and *modification* for the latter is consistent with prior case law and with the English meanings of those words as applied to the psychometric interpretations of score comparability described previously. This differential usage of the terms *accommodation* and *modification* is employed throughout this chapter.

Tension Between Accessibility and Construct Preservation/Score Comparability

As previously indicated, it is unfortunate that some advocates for students with disabilities have urged test administrators to classify all nonstandard test administrations as accommodations and to treat the resulting scores as comparable to scores from standard administrations. Such actions are not consistent with the legal requirement for the provision of reasonable accommodations. According to its legal definition, a *reasonable accommodation* must

- be *needed* by a disabled person to *access* the test
- while ensuring *construct preservation* and *score comparability*.

A testing condition variation needed to access the test means that the student with a disability is unable to meaningfully respond to the test questions without it. The phrase *needed for access* requires more than simply providing assistance that helps the student with a disability to obtain a higher score; the assistance must be necessary for participation in the testing program. The phrase *ensuring construct preservation and score comparability* means that the change in test administration conditions produces scores that are free from extraneous (content irrelevant) factors while preserving the knowledge and skills intended to be measured and producing scores that have the same interpretation

and intrinsic meaning as scores from standard test administrations. Language from the testing accommodations cases discussed earlier and the *Test Standards* support the interpretation of *reasonable accommodations* as providing accessibility for individuals with disabilities without compromising construct interpretation or score comparability (Davis, 1979; Brookhart, 1979; Ambach, 1982; Section 504 Regulations, 1997 and ADA Regulations, 1997; Rene, 2001 [discussed below]).

Labeling Nonstandard Test Administrations

The following questions should be considered when labeling a nonstandard test administration as an *accommodation* or a *modification*:

1. Will the test scores obtained under altered testing conditions have a different interpretation than scores obtained under standard test administration conditions? Are the test scores from standard and nonstandard administrations comparable?
2. Is the alteration in test format or administration conditions part of the skill or knowledge being tested? Is it construct relevant?
3. Would a nonstandard test administration also assist nondisabled, low-achieving students to demonstrate what they know and can do without changing the interpretation of their test scores?
4. Can valid and reliable decision procedures and appeals be established for determining which students will be allowed specific nonstandard test administrations?
5. Do students with disabilities included in regular education classrooms have any responsibility for adapting to standard testing conditions (and foregoing nonstandard test administrations) when feasible?⁴

Some observers have argued that any nonstandard test administration that helps a student with a disability to achieve a higher score is a

reasonable accommodation that should be treated the same as scores obtained by students with common accessories such as eyeglasses (Fraser & Fields, 1999). Unfortunately, this view fails to distinguish between variations for extraneous factors and variations that are closely related to the cognitive skill being measured. Eyeglasses are a reasonable accommodation for a mathematics estimation test because vision is not part of the skill the test is intended to measure. Alternatively, although a calculator on the same mathematics estimation test might assist a student with a learning disability to achieve a higher score, it would be a modification because its use changes the skill being measured from application of rounding and approximation techniques to pushing the correct buttons on the calculator.⁵

Some observers have also questioned policy decisions that exclude modified test administrations from high-stakes decisions. Yet the lack of availability of such scores for use in making high-stakes decisions is appropriate when those scores have different interpretations than scores obtained from standard or accommodated administrations. For example, it would be misleading and unfair to treat a modified test score from a reading comprehension test obtained with a reader (a measure of listening comprehension) as having the same meaning as scores obtained from standard administrations where students read the test material silently by themselves. Alternatively, when nonstandard test administrations are limited to only those testing variations that maintain test score comparability and preserve the intended construct (i.e., accommodations), the resulting scores should be interpreted the same as scores obtained from standard administrations.

⁵ Prior to becoming a psychometrician and attorney, the author was a mathematics educator. Consequently, many of the content examples of testing variations analyzed in this chapter using logical evidence are drawn from mathematics. Note also that the increased efficiency rationale for allowing students with learning disabilities to use calculators on a mathematics computation test is equally applicable to low-achieving regular education students and would be universally inappropriate if the content standards intended to be measured specified application of paper-and-pencil computational algorithms.

⁴ Adapted from Phillips (1993, p. 27, 1994, p. 104).

Skill Substitution

When a student is given a modification that changes the skill intended to be measured, the student has been permitted to substitute a different skill for the one tested. The following sections discuss examples of skill substitution related to extended time, readers, and calculators.

Extended Time. Extended time is a particularly difficult nonstandard test administration to classify. Deciding whether it is an accommodation or modification depends on the degree of intentional and unintentional speededness of the test. For example, if the purpose of the test is to measure how quickly a student can copy a pattern of numbers and letters, speed is part of the skill being measured. In this situation, an extended time administration would change the skill being measured from accuracy and speed to accuracy only. This change in the skill being measured would change the interpretation of the resulting test score, so in this case the extended time administration should be classified as a modification. Examples of speeded tests commonly administered in elementary schools are *math facts tests* for which the student is given a sheet of addition facts (e.g., $5 + 7 = \underline{\quad}$) or multiplication facts (e.g., $9 \times 8 = \underline{\quad}$) with a fixed amount of time to correctly answer as many items as possible.

Alternatively, an achievement test may intentionally be designed to be a power test with generous time limits. The purpose may be to measure pure academic knowledge irrespective of the time taken to demonstrate that knowledge. In this case, extended time may be judged to be an accommodation of a factor extraneous to the skill being measured. Even so, if a portion of the state test is norm referenced (e.g., Iowa Tests of Basic Skills (ITBS), American College Testing (ACT)), the corresponding normative score information would be valid only for scores obtained under the same timing conditions as the norm group.

Some educators argue that all achievement tests should be power tests with no speededness. But there are two reasons why allowing unlimited time may be counterproductive. First, if given unlimited time, some students will continue working on the test well beyond the point of

productivity. This behavior wastes instructional time and may unnecessarily tire and frustrate the student. Second, one of the goals of education is to help students automate skills so that they are readily available when needed. Thus, students who take 4 h to complete a task that most students can complete in 1 h may not have the same level of skill development. Assuming all other relevant factors equal, it is unlikely that an employer would be indifferent between these two potential applicants. Huesman and Frisbie (2000) demonstrated that both students with learning disabilities and students in regular education significantly improved their scores on a standardized, norm-referenced achievement test when given extended time.

An extended time alteration may be combined with other testing condition variations such as segmented administration with extended rest breaks. This combination of changes in standard testing conditions could result in a 1-h test, typically administered in a single sitting with no breaks, being administered over a period of 4 h spread across one school week. Such a nonstandard administration is often more expensive to schedule, and it may be difficult to ensure that the student has not obtained inappropriate coaching or assistance between testing sessions. In addition, in real-world college, professional licensure, and employment testing contexts, users often interpret scores from such a nonstandard test administration differently because they judge skill automaticity, efficiency, and speed of work to be construct relevant. Thus, it is not clear that students with disabilities are well-served by policies that permit major time extensions with segmentation for school tests when it is unlikely such options will be available or judged to produce comparable scores in postsecondary contexts.

To address this concern, the combined effects of taking four times as long in actual time and completing the task in much smaller increments, possibly with review or coaching in between, must be considered. Because most students find a final exam over all the material covered in a course much more difficult than a series of tests over individual units, one might conclude that the measurement of multiple skills at a single point in

time is different than measurement of a series of skills at different points in time. That is, extending testing time across multiple days in small segments appears more consistent with the characteristics of a modification. Thus, if additional time is permitted to compensate for an extraneous factor such as reading Braille, the time extension should have reasonable limits within a single day (and preferably within a single block of time) to be labeled a reasonable accommodation.

Readers. Another nonstandard test administration that is difficult to classify is a reader for a mathematics test. Again, classifying this testing variation as an accommodation or modification depends on the purpose of the test and may involve competing educational goals. On the one hand, curriculum specialists may argue for more authentic mathematics tasks (real-world problems) that require students to read text and graphs, apply mathematical reasoning, and explain their solutions in writing. On the other hand, advocates for students with learning disabilities may argue that assessing communication as part of a mathematics test penalizes students with reading/writing disabilities and does not allow them to fully demonstrate their mathematical knowledge.

Nonetheless, on real-world mathematics problems, poor performance by low-achieving students without disabilities may also be the result of poor reading or writing skills. But a reader, often available to students with learning disabilities, usually is not provided to a nondisabled student with poor reading skills. A policymaker might question why nondisabled students who read slowly and laboriously, have limited vocabularies, have difficulty interpreting symbols, respond slowly, suffer test anxiety, or have difficulty staying on task are less deserving of the opportunity to demonstrate maximum performance with a reader than is a student who has been labeled *learning disabled*. Perhaps no one has yet discovered the “disability” that causes the low-achieving students to experience the listed difficulties.

Some evidence suggests that students with disabilities profit more from a reader than do low-achieving students (Tindal, 1999). This may

be because testing variations are usually most effective when students have had practice using them. Instructional prescriptions for students with learning disabilities often require all text to be read to the student, whereas low-achieving students are typically expected to continue struggling to read text by themselves. Therefore, students with learning disabilities may have received extended practice listening to text being read aloud and little practice reading the text themselves, while the experience of low achievers was exactly the opposite. Nevertheless, Meloy, Deville, and Frisbie (2000) found that students with and without disabilities benefited from a reader on a standardized, norm-referenced achievement test.

The policy dilemma is whether to (a) confer a benefit on students with disabilities and potentially penalize low achievers by only allowing readers for students with disabilities on mathematics tests, (b) label a reader as a modification for all students or (c) allow a reader for any student who needs one and treat the resulting scores as comparable to scores from standard administrations. Choosing (a) creates the inconsistency of arguing that reading is part of the skill being measured for nondisabled students but not for students with disabilities. Choosing (b) preserves the intended construct of authentic problem solving but may be unpopular with advocates for students with disabilities because these students’ reading/writing disabilities would be considered part of the skill intended to be tested. Choosing (c) minimizes the communication component of mathematics and may be contrary to the test specifications and the corresponding state mathematics content standards. A more straightforward method for minimizing the impact of skills other than pure mathematics is to develop less complex items that use simplified text and/or more pictorial representations for all students. However, this option may also inappropriately alter the construct intended to be measured. Policymakers must make the final decision based on the purpose for testing, the tested content standards, the intended use(s) of the scores, and their evaluations of the relative merits of these competing policy goals.

Another important consideration in the formulation of a policy on readers for mathematics tests is their unintended effects on student and teacher behavior. If readers are allowed for all students or if mathematics items are simplified to reduce reading/writing complexity, teachers may be encouraged to engage in less real-world instruction in the classroom or may be encouraged to provide a read-aloud instructional strategy for most students who are having difficulty reading text themselves. Alternatively, disallowing readers for all students on a mathematics test might discourage teachers from providing read-aloud instructional strategies for students with disabilities and encourage them to give greater emphasis to improving students with disabilities' abilities to read text themselves. This latter unintended outcome might provide important long-term benefits to individual students with disabilities. Consider the following actual example from a statewide graduation testing program.

A first-grade student, Joey (not his real name), was struggling in school relative to the earlier performance of his gifted older brother. His parents requested a special education referral, and Joey was tested. His performance in mathematics was about one year below grade level, and his performance in reading was about half a year below grade level. There was a small discrepancy with ability in mathematics and none in reading. The IEP developed for Joey, with significant parental input, labeled him *learning disabled* in mathematics and required all written material to be read aloud to him. When Joey entered high school, an enterprising teacher chose to disregard the IEP and began intensive efforts to teach Joey to read. Within about 18 months, Joey went from reading at a second-grade level to reading at a sixth-grade level. Had this teacher not intervened, Joey would have graduated from high school with only rudimentary reading skills. Although his sixth-grade reading skills were not as high as expected of high school graduates, they provided some reading independence for Joey and allowed him to eventually pass the state's mathematics graduation test without a reader.⁶ (Phillips, 2010, chapter 5)

This example illustrates an important trade-off in providing debatable testing variations such as readers for mathematics tests. Joey's parents pressured his teachers to read all instructional materials aloud to him so he could avoid the frustration of struggling with a difficult skill. The downside was that Joey became totally dependent on the adults in his life to read everything to him. Yet, when he finally was expected to learn reading skills, he was able to develop some functional independence. Although he still may require assistance with harder materials, there are many simpler texts that he can read solo.

Calculators. Suppose a mathematics test measures paper-and-pencil computation algorithms and estimation skills. If calculators are not permitted for a standard administration of the test but are allowed for students with the learning disability of dyscalculia (difficulty handling numerical information), the resulting scores will represent different mathematical skills and will not be comparable. For example, consider this item:

About how many *feet* of rope would be required to enclose a 103' by 196' garden?

- A. 25 *B. 50 C. 75 D. 100

Students without calculators might round 103 and 196 to 100 and 200, calculate the approximate perimeter as $2(100) + 2(200) = 600$ inches, and convert to feet by dividing by 12. Students with calculators would likely use the following keystrokes: $103 + 103 + 196 + 196 = \div 12 = .$ The calculator would display 49.833... and the student would only need to choose the closest answer: 50. Clearly, use of the calculator would subvert the skills intended to be measured because the student would not have to do any estimation or computation to select the correct answer. In this case, the student's disability is not an irrelevant factor to be eliminated but rather an indication that the student does not have the cognitive skills measured by the mathematics test. Moreover, classifying calculator use as a modification for a mathematics test with such items is based on evaluation of the test specifications as they are *currently* written. If critics want students tested on different skills (e.g., calculator literacy

⁶ Incidentally, when the items on the state graduation test were divided into two categories, symbolic text and story problems, the percent of items Joey answered correctly was similar for both categories.

and problem solving rather than estimation and paper-and-pencil computation), they should first convince policymakers to change the test specifications and corresponding mathematics content standards.

Construct Fragmentation

In some cases, when a nonstandard test administration assists with or removes a single, minor factor from the construct measured on a test, a reasonable argument may be made that it is extraneous or not essential to the essence of the construct. However, if several parts of the construct are removed concurrently by providing multiple testing variations to the same students, the remaining parts of the construct that are tested may seriously distort the intended measurement. Allowing any student or group of students to choose which parts of the construct they will be tested on and which parts of the construct will be removed or receive assistance allows these students to be tested on different definitions of the construct and produces noncomparable scores.

For example, to assist students with learning disabilities on a mathematics problem-solving test, a test administrator might (a) permit calculator use, arguing computation is not part of the intended skill; (b) read the test aloud, arguing reading the problems is not part of mathematics knowledge; (c) provide a reference sheet with formulas and measurement unit conversions, arguing memorization of that information is unimportant; (d) remove extraneous information from each problem, arguing such information is an unnecessary distraction; and (e) eliminate one of the answer choices, arguing that with fewer choices, the student will be more focused on the task. If all these parts of the construct of solving mathematics problems, each judged individually to not be part of the intended measurement (i.e., construct irrelevant), are removed from the tested skill to provide greater “access” for students with learning disabilities, the only remaining tasks are to select the correct formula, plug the numbers into the calculator, and find the matching or closest answer choice. Similar

arguments can be made for administration of a reading comprehension test to ELLs with decoding, language complexity, more difficult vocabulary and nonessential text removed from the test, and each test question relocated to directly follow the paragraph containing the answer. In both examples, the intended construct has been fragmented into pieces, selected pieces have been altered or removed, and the reassembled pieces have produced a distorted construct. If the altered and lost fragments of the construct were part of the original test specifications and corresponding content standards, content validity judgments will characterize the tested construct as qualitatively different from that measured by a standard test administration.

Construct Shift

In addition to fragmenting the construct into removable parts with multiple testing variations, there is another problem with the logic sometimes used to identify “appropriate” testing variations. According to the *Test Standards*, the construct which defines the skills intended to be measured should be a property of the test and should be defined by the test’s content objectives and specifications. But, for example, when a reading test is administered with a reader for students with learning disabilities but not for regular education students who are poor readers, or when language assistance is provided to ELLs with English language skill deficiencies but not to non-ELLs with similar deficiencies, the construct has been redefined by group membership. For the reading test, the construct tested for students without learning disabilities is reading comprehension but for students with learning disabilities is listening comprehension. Similarly, for non-ELLs, the construct is content (reading or mathematics) in English, but for ELLs, the construct is content only with the effects of language removed to the extent possible. Thus, the definition of the construct has shifted from a property of the test to a property of the group to which the student is a member. When tests measuring different constructs

are administered to students in different subgroups, the resulting scores are not comparable and, according to the *Test Standards*, should not be interpreted as having the same meaning (e.g., proficient, passing) for all students.

Public Policy Exceptions

Normally, the award of diplomas, licenses, and credentials conditioned on the achievement of a specified test score should only be made when passing scores are obtained from standard administrations or nonstandard administrations that produce comparable scores. However, there may be extraordinary circumstances for which a special waiver of the testing requirement may be appropriate.

For example, suppose a student who has been taking accelerated college-preparatory courses and receiving “A” grades is involved in a tragic automobile accident just prior to the initial administration of the graduation test in eleventh grade. Suppose further that there was extensive evidence that the student’s academic achievement in reading and mathematics had already exceeded the standards tested by the graduation test and that the student had been expected to pass easily with a high score. However, due to the accident, the student is now blind and has no hands. The student can no longer read the material on the reading test visually (no sight) and is unable to learn to read it in Braille (no hands). But, administering the reading test aloud via a reader would alter the tested skills from reading comprehension to listening comprehension producing noncomparable scores.

Nonetheless, as a matter of public policy, such a case might be deserving of a waiver of the graduation test requirement in recognition of the extraordinary circumstances and compelling evidence of achievement of the tested skills. Based on the student’s medical records, transcripts, references, and a passing score on the graduation test obtained with the modification of a reader, an appeals board might determine that this student should receive a test waiver and be eligible to receive a high school diploma if all other

graduation requirements are satisfied. Rather than treating this modification as if it were an accommodation, it is preferable to grant a waiver of the testing requirement when a student is otherwise qualified. Waivers granted judiciously avoid pretense and preserve the integrity of the testing program. It should be a rare event for students with disabilities to be able to document achievement of high school level competencies but be unable to access the graduation test without a modification.

Leveling the Playing Field: Access Versus Success

One of the goals of federal legislation for students with disabilities is *access*. A goal of access reflects an understanding that there is value in having students with disabilities participate in testing programs under any circumstances. One reason policymakers value access may be that it results in greater accountability for school districts in providing instruction and demonstrating educational progress for special-needs students. If this is the case, any nonstandard test administration that moves a student from an exclusionary status to one of being included in the test may be desirable.

However, some advocates have interpreted the goal of federal legislation for students with disabilities to be increased *success* (higher test scores). Those who support this view argue that students with disabilities should not be penalized for biological conditions outside their control. They believe that the intent of federal legislation was to confer a benefit on students with disabilities that would change the skills being measured from tasks the student cannot do to tasks the student is able to do. The term *leveling the playing field* has often been used in this context to describe testing modifications that go beyond the goal of access to a goal of success and confuse removing extraneous factors with equalizing test scores.

Increasing access by removing the effects of extraneous factors supports the validity of test score interpretations by requiring students

with disabilities to demonstrate all relevant skills while not penalizing them for unrelated deficiencies. Alternatively, increasing success by providing a compensating advantage to offset relevant deficiencies so that students with disabilities have an equal opportunity to obtain a high or qualifying score relative to their nondisabled peers decreases the validity of the resulting test scores as indicators of achievement of the skills the test is intended to measure.

Leveling the playing field to equalize success is analogous to a golf handicap tournament. Based on prior performance, golfers are given a handicap that is subtracted from the final score. In a handicap tournament, a poor golfer who plays well on a given day can win over a good golfer with a lower score who has not played as well as usual that day. Handicapping prevents the good golfers from always winning.

Similarly, using a handicapping model for interpreting *leveling the playing field* would create a testing scenario in which the highest achieving students would not always receive the highest scores. In such a scenario, the test would be viewed as a competition for which all participants, regardless of their level of knowledge and skills, should have an equal chance of obtaining a proficient score. Under this interpretation, each student with a disability would be given whatever assistance was necessary to neutralize the effects of the student's disability. For example, students who were not proficient at estimation would be given a calculator for those items; students who had difficulty processing complex textual material would be given multiple-choice items with only two choices rather than the usual four. Low-achieving students whose poor performance could not be linked to a specific disability would not receive any assistance; they would be required to compete on the same basis as the rest of the regular education students. The settlement of a federal case in Oregon case illustrates some of the concerns with using the goal of *equal opportunity for success* rather than *equal access to level the playing field* for students with disabilities.

Oregon Case Settlement

Advocates for Special Kids v. Oregon Dep't of Educ. (2001) involved a ban on the use of computer spell-check on a tenth-grade writing test for which 40% of the score was based on spelling, grammar, and punctuation.⁷ Affected students and their parents claimed the ban discriminated against students with learning disabilities. A plaintiff student with dyslexia stated that “[w]hen they test me in spelling, they’re testing my disability, which isn’t fair. It’s like testing a blind man on colors.” Unconcerned that her spelling deficit would hinder her career, she also stated that “[a] lot of things I like are hands-on, [a]nd if I become a writer, they have editors for that” (Golden, 2000, p. A6).⁸

In a settlement agreement between the plaintiffs and the state testing agency, the parties agreed to changes in the state testing program based on recommendations of an expert panel convened to study the issue. In the settlement, the state agreed to permit all requested nonstandard administrations on its tests unless it had empirical research proving that a specific nonstandard administration produced noncomparable scores. Contrary to the *Test Standards*, which creates an assumption of noncomparability of scores from nonstandard administrations in the absence of credible evidence of score comparability, this agreement effectively created a default assumption that a requested nonstandard test administration produced comparable scores unless the state could produce empirical evidence that proved otherwise.

This settlement reflected the reality that negative publicity can motivate parties to settle a lawsuit contrary to accepted professional standards. Effectively, agreeing to these conditions meant

⁷ Technically, passing the tenth grade test was not required for receipt of a diploma. Nonetheless, retests were provided, and it functioned like a graduation test of minimal skills expected of high school graduates.

⁸ Note, however, that a color-blind individual cannot obtain a pilot's license.

that testing variations such as a reader for a reading test, a calculator for a computation or estimation test, or spell-check and/or grammar check for a writing test would automatically be provided to students with learning disabilities who regularly used them in instruction and requested them for the test, even though most psychometricians would agree that they altered the construct intended to be tested and produced noncomparable scores. In addition, because states often have few resources for research and often not enough students with a given disability requesting a particular testing variation to conduct a separate study, this settlement made it nearly impossible for the state to place limits on nonstandard test administrations for students with disabilities, even when logical evidence clearly indicated that the scores were not comparable. Moreover, if the state did impose any limits on nonstandard test administrations, the state was required to provide an alternate assessment for the test that probably would not have produced comparable scores to those obtained from standard test administrations.

To date, no court has ordered a state to interpret nonstandard administrations that altered the skills intended to be measured and produced noncomparable scores as equivalent to standard administrations because federal disability law does not require it. Had the Oregon case gone to trial, it is unlikely that the court would have done so. Unfortunately, despite the fact that this settlement applied only to the Oregon testing program, it was used by advocates to pressure other states to grant similar concessions.

No Child Left Behind (NCLB) Modified Tests

The issue of access versus success and its implications for the validity of the resulting test score interpretations is especially relevant to the modified tests permitted under the NCLB Act (2000) guidelines for students with disabilities who have had persistent academic difficulties. Federal guidelines require on-grade-level content but permit cognitive difficulty to be reduced. State may count up to 2% of students with disabilities as

proficient on such modified tests. Many states have reduced the cognitive difficulty of test items from standard administrations by reducing the number of answer choices, decreasing language complexity, removing nonessential information, adding graphics, shortening the test, and so on. However, as described previously, the intersection of content and cognitive difficulty defines the content coverage and grade-level constructs intended to be measured by most achievement tests. Thus, if a modified test retains the content designations but changes the cognitive difficulty of the items, the measured construct will change and the resulting test scores will not be comparable to scores from the on-grade-level test unless the two tests have been statistically linked. If so, the modified test score corresponding to *proficient* will typically be proportionally higher than for the on-grade-level test. Moreover, recent research has suggested that some of the item modifications intended to increase access (or success) may be ineffective (Gorin, 2010).

For example, consider the fifth-grade mathematics story problem and the simplified item for the same problem shown below.

Original Item: (The field test percent of students choosing each option is given in parentheses.)

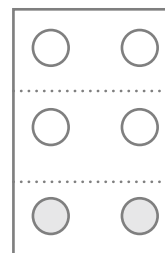
Tom divided his 12 cookies equally between himself and his 4-year-old sister Sara who then ate one-third of her share. How many cookies did Sara have left?

- A. 2 (22%)
- B. 3 (2%)
- * C. 4 (72%)
- D. 6 (4%)

Modified Item:

Sara has 6 cookies. She eats one-third of them. How many cookies are left?

- A. 3
- B. 4
- C. 6



In the original item, the student must analyze the problem; determine several relationships; use mathematical reasoning to understand what

happened; identify “4” as an irrelevant number; translate “divided,” “one-third of,” and “left” into the operations of division, multiplication of fractions, and subtraction; set up the problem; and solve for the answer. The modified item presents the same fractional computation but with less language complexity, fewer steps required to solve the problem, and fewer answer choices. Tom is removed completely, extraneous information has been eliminated, and the problem statement is broken into two simpler sentences using present tense verbs, no relational words, fewer modifying phrases, and fractions written as numbers rather than words. There are only three answer choices, the most attractive wrong answer has been eliminated, and the graphic demonstrates how to setup the problem.

The modified item still involves the same content (fractions), but the cognitive difficulty has been reduced because there are fewer steps, no extra information, and the student can obtain the correct answer from the graphic by counting. If, as part of a content validity judgmental review of these two items, elementary mathematics teachers stated that the modified item matched the initial introduction of fractions skills described in the fourth-grade content standards and the original item matched the fractions skills described in the fifth-grade content standards, these judgments would provide logical evidence that use of the modified item on the fifth-grade mathematics test would alter the construct intended to be measured and produce noncomparable scores. Thus, as judged by the mathematics teachers, students who could correctly answer the original item would have demonstrated more mathematical skill at a higher grade level than students who could correctly answer the modified item. Further, the modified item would be easier to guess correctly with fewer answer choices and the most attractive wrong answer eliminated.⁹

⁹ If the rationale for eliminating an answer choice is to reduce the reading load, reduce fatigue, and/or compensate for a short attention span, all of which might be considered construct-irrelevant factors for a mathematics test, this author’s opinion is that it would be more defensible to eliminate the least-often-chosen wrong answer. Eliminating the most attractive wrong answer appears to inappropriately address *success* rather than *access*.

The federal guidelines appear to intend these results, and there is no conflict with the *Test Standards* if scores from tests with modified items are interpreted as measuring a modified construct and as comparable only when statistically linked to the on-grade-level score scale. Some researchers who have investigated various item modifications in multiple states describe their modified items as providing greater access for students with disabilities with persistent academic difficulties (Elliott et al., 2010). However, their criteria for a useful item modification include a narrowing of the gap between average scores of students with and without disabilities. This criterion sounds more like one of *success* than *access* because it assumes that the only reason students with disabilities have performed poorly on the on-grade-level tests is due to lack of “access” to the test questions. However, some research studies have gathered evidence suggesting that the real problem for some of these students with disabilities is lack of “access” to instruction on the tested skills (Stoica, 2010). When this is the case, altering the content by modifying the items may not be addressing the real access problem faced by these students with disabilities. Thus, a greater narrowing of the achievement gap (success) for students with disabilities eligible for modified tests may be realized through a closer match between these students’ instruction and the on-grade-level tested skills (access through instruction rather than access through item modification).

Eligibility for Nonstandard Test Administrations

There is a difference between truly needing something and wanting something. Humans need food and water to survive, but they do not need to eat apple pie and drink coffee to survive. Some people may want the latter and be unhappy if they do not get it, but they can meet their survival needs with other less appealing foods and plain water. Similarly, students who are blind need an alternative method to access written text or they will be unable to provide any response to the test questions. However, a student with a

learning disability does not necessarily need to have a reader. The student with a learning disability may have sufficient reading skills to respond to the test questions but may want a reader to achieve a higher score and avoid the frustration of struggling with a difficult task.

The challenge for test administrators is to separate those students who truly cannot participate without a nonstandard test administration from those who simply want the additional assistance to make the task easier. Nonetheless, even a nonstandard test administration that is truly needed for a student to access the test may invalidate the intended score interpretation. In such a case, policymakers must determine whether the benefits of inclusion with a modification outweigh the costs of losing interpretive information and being excluded from aggregate results. In addition, if a modification confers a benefit on a specific subset of students, it is important to determine who qualifies for the benefit. In this case, relevant considerations may be these: recent written documentation of the disability by a trained professional, routine provision of the modification in the student's educational program, an appropriate relationship between the disability and the desired modification, and the availability of facilities for providing the modification.

In the past, most testing programs have relied on the student's IEP or 504 Plan to certify a disability. In many cases, achievement below expectations has been sufficient to label a student *disabled*. However, two Supreme Court decisions suggest a narrower interpretation of the term *disability*. In these cases, the Court held that persons are disabled only if substantial life activities are affected *after correction or mitigation of an impairment* (*Murphy v. United Parcel Service*, 1999; *Sutton v. United Airlines*, 1997). The Court held that an employee with a corrected impairment was not disabled even if the person theoretically would have difficulties in the absence of medication or corrective apparatus (e.g., medication for high blood pressure or eyeglasses). This holding suggests a duty to take feasible corrective action for a disability and then be evaluated for the level of impairment.

In an educational context, this holding may mean that a student would qualify as having a reading disability and be eligible for a reader only if it is impossible for the student to learn to read, not merely because the student might have difficulty learning to read or because the student's current skill level might be significantly below grade level. Nonetheless, a student who is blind and for whom reading printed text is impossible would still qualify for a Braille administration. Such an interpretation would have the advantages of reserving nonstandard test administrations for those whose conditions cannot be modified or corrected and reducing any incentive to label a student *disabled* whenever the student has difficulty with academic tasks.

Undue Burdens

Case law has established that an accommodation is not required if it imposes an undue burden. However, most interpretations have required extreme expense or disruption for a requested nonstandard test administration to qualify as an undue burden. Most testing programs are burdened with extra costs and scheduling difficulties for some nonstandard administrations. For example, more space and substantially increased numbers of test administrators are required when significant numbers of students are tested individually. An individual administration may be required due to frequent breaks, attention problems, readers, use of specialized equipment, medical monitoring, and so on, but such administrations may also pose a major resource allocation challenge for schools when testing windows are narrow. At issue is when such burdens rise to the level of an undue burden and need not be provided.

A situation in which a very expensive testing variation has been requested for a single student may qualify as an undue burden. For example, for one state graduation test, a student's parents requested a 99-point large-print version. This size print allowed only one to two words per page and required special large-paper versions for diagrams. The cost of producing such a

large-print version for a single test form was over \$5,000. Because it was a graduation test, there was also the possibility that additional versions would have to be created if the student was unsuccessful on the initial attempt. In this example, the student had a visual impairment that could not be further corrected. Thus, it was impossible for the student to take the test without the requested testing variation. However, the cost was significant for the state to pay for a single student. In such a case, the requested nonstandard test administration might be judged an undue burden and a state might not be required to provide it.¹⁰

Moreover, state policymakers must prioritize the use of available resources for nonstandard administrations. It may not be possible to provide the test in all requested formats, even when such requests are consistent with state policies. An alternative may be to allow local test administrators to decide which nonstandard administrations to provide. Such a policy would have the advantage of limiting the state cost, but would have the disadvantage of resulting in similarly situated students being treated differently depending on local policy in their geographic area. In addition, there would probably still be a duty for state policymakers to provide guidance to local policymakers regarding accommodation/modification decisions, appropriate use and interpretation of the resulting test scores, and informed consent of parents/students.

Graduation Testing Challenges

Some states with graduation tests award high school diplomas to students with disabilities who

have completed their IEPs, ELLs who have taken a translated test, or students who have passed the graduation test with modifications. However, the tendency of states to tie testing modifications to specific disabilities (or lack of language proficiency for ELLs) indicates that they implicitly recognize that test scores from modified administrations are not comparable to those obtained from standard test administrations. If a state allows scores from modified test administrations to be interpreted the same as scores from standard administrations, it has conferred a benefit on those students who received the modifications. When benefits are conferred, the state must establish eligibility criteria and evaluate the qualifications of persons who seek the benefit. Effectively, the conferring of such benefits means that if a student can document a disability or ELL status, the student is permitted to satisfy the state testing requirement by substituting an alternative skill and/or being exempted from portions of the tested skill.

Instead of awarding diplomas to students who have completed IEPs, passed translated tests, or received the benefit of modifications on a graduation test, some states substitute a certificate of completion for these students. This policy, challenged in recent litigation, was influenced by the early challenges to graduation testing of special education students referenced previously and challenges to graduation tests by minority students claiming racial/ethnic discrimination. The graduation test challenges claiming racial/ethnic discrimination, *Debra P. v. Turlington* (1984) and *GI Forum v. Texas Educ. Agency* (2000), are discussed next followed by additional graduation testing cases involving students with disabilities and ELLs.

The Debra P. and GI Forum Cases

In the *Debra P.* case, African-American students subject to a Florida graduation test requirement alleged that it was unconstitutional because they had received an inferior education in segregated schools and because they had too little time to prepare for it. Initially, the court announced

¹⁰ The stated reason for requesting the nonstandard administration was a parental desire for the student to attend a particular state university that required a high school diploma for admission. Due to the severity of the disability (the student could only work on the test for about ten minutes at a time and was tested in small segments over several weeks), some educators questioned the student's ability to handle college level coursework even if the student passed the modified test. Note also that a computerized test administration, unavailable at the time, might provide adequate magnification at less cost.

two new requirements for graduation tests, *notice* and *curricular validity*, and ordered Florida to award diplomas to students who had failed the graduation test but satisfied all other graduation requirements. The court defined curricular validity as evidence that the tested skills were included in the official curriculum and taught by the majority of teachers. Once Florida produced sufficient evidence of the test's curricular validity and students had been given at least 4 years' notice of the requirement, the court held that Florida could require students to pass the graduation test to earn a high school diploma.

In the *GI Forum* case, Hispanic and African-American students alleged that the Texas graduation test discriminated against them because they received inadequate educational preparation and the test failed to satisfy professional standards. The *GI Forum* court reaffirmed the holdings of the *Debra P.* court and ruled in favor of the state, finding

While the [graduation test] does adversely affect minority students in significant numbers, the [state] has demonstrated an educational necessity for the test, and the Plaintiffs have failed to identify equally effective alternatives. . . . The [state] has provided adequate notice of the consequences of the exam and has ensured that the exam is strongly correlated to material actually taught in the classroom. In addition, the test is valid and in keeping with current educational norms. Finally, the test does not perpetuate prior educational discrimination. . . . Instead, the test seeks to identify inequities and to address them. (p. 684)

Notice

Notice requires the state to disseminate information about graduation test requirements to all affected students well in advance of implementation. Notice periods of less than 2 years have been found unacceptable by the courts; notice periods of 4 years in the *Debra P.* case and 5 years in the *GI Forum* case were found acceptable. The courts have not mandated a specific length for the notice period; with extensive dissemination efforts, solid curricular validity, and systematic achievement testing of prerequisite skills in earlier grades, a 3-year notice period may be adequate.

There has been some debate about whether the notice period applies to the first administration of the graduation test or to students' scheduled graduations. The *Debra P.* and *GI Forum* cases referred to the latter, but satisfying curricular validity may require a longer notice period to allow students to complete all the coursework covering the tested skills. For example, if a graduation test administered in the spring of eleventh grade includes Algebra II content, and students must take courses covering Pre-algebra, Algebra I, and Plane Geometry prior to taking Algebra II, notice must be given by seventh grade so students can enroll in Pre-algebra no later than eighth grade. In this case, notice would occur 4 years before the first test administration and 5 years before the students' scheduled graduations.

Curricular Validity

The curricular validity requirement, also referred to as *opportunity to learn* (OTL), was included as Standard 8.7 in the 1985 version of the *Test Standards* and carried forward as Standard 13.5 in the 1999 revision. OTL means that students must be taught the skills tested on a graduation test. In practice, evidence of OTL is often gathered by examining the official curricular materials used in instruction and surveying the work of teachers to determine whether they are teaching the tested content. In the *GI Forum* case, the court held, on all the facts and circumstances, that the state had satisfied the curricular validity requirement with a mandated state curriculum, teacher committee item reviews that considered adequacy of preparation, remediation for unsuccessful students mandated by statute, and continuity of the graduation test with its predecessor which was based on the same curriculum and for which an OTL survey of teachers had been completed.

Retests and Remediation

Multiple retests combined with substantial remediation efforts were persuasive in the *Debra P.* and *GI Forum* cases. The *Debra P.* court stated

[The state's] remedial efforts are extensive Students have five chances to pass the [graduation test] between the 10th and 12th grades, and if they fail, they are offered remedial help All [of the state's experts] agreed that the [state's remediation] efforts were substantial and bolstered a finding of [adequate opportunity to learn]. (pp. 1410–1411)

In the *GI Forum* case, the Texas Education Code provided: “Each school district shall offer an intensive program of instruction for students who did not [pass the graduation test]” (§ 39.024 (b)), and the court held

[A]ll students in Texas have had a reasonable opportunity to learn the subject matters covered by the exam. The State's efforts at remediation and the fact that students are given eight opportunities to pass the [graduation test] before leaving school support this conclusion. (p. 29)

Notice and Curricular Validity for Special Education Students

Contemporaneous with the *Debra P.* and *GI Forum* cases, lawsuits in Illinois and New York addressed graduation testing of students with disabilities. These cases upheld the application of graduation test requirements to students with disabilities and addressed the notice and curricular validity requirements as applied to special education students.

The Brookhart Case

The *Brookhart v. Illinois State Bd. of Educ.* (1983) case addressed the due process requirements for graduation tests as applied to students with disabilities. It involved a minimum-competency test mandated by a local school district in Illinois. The testing requirement was imposed in the spring of 1978 and became effective for the spring 1980 graduating class. Prior to their scheduled graduations, regular and special education students had five opportunities to pass the three-part test of reading, language arts, and mathematics skills at a passing standard of 70%. Students who failed the graduation test received certificates of completion instead of a high school diploma.

Several students with disabilities who had successfully completed their IEPs but who had failed the graduation test and been denied diplomas filed a lawsuit to challenge the testing requirement and force their districts to award them diplomas. They alleged that their constitutional rights had been violated by insufficient advance notice of the graduation test requirement. A variety of disabilities were represented among the students challenging the testing requirement including physical disabilities, multiple impairments, mild mental retardation, and learning disabilities. The trial court ruled against the students, stating

It is certainly true that giving a blind person a test from a printed test sheet discloses only his [disability] and nothing of his knowledge. To discover a blind person's knowledge, a test must be given orally or in [B]raille This ... does not mean that one can discover the knowledge or degree of learning of a [cognitively] impaired student by modifying the test to avoid contact with the [cognitive] deficiency. To do so would simply be to pretend that the deficiency did not exist, and to fail completely to measure the learning This position does not involve any lack of compassion or feeling for people living with serious [disabilities]; it simply involves the avoidance of pretense, the integrity of a knowledge-testing program, and reserves some meaning for a high school diploma in relation to the attainable knowledge and skills for which the schools exist. (p. 729)

The trial court also held that awarding a diploma to a student with a disability who did not receive extended notice of the graduation testing requirement

is a misunderstanding and debasement of the concept of due process of law... [D]ue process of law does not require pretending that such standard has been achieved by a person whose [disability] clearly makes attainment of that standard impossible or inappropriate ... The means to avoid [the stigma of failing to receive a high school diploma] exists in attainment of the necessary knowledge. If capacity for such does not exist, the law does not require pretense to the contrary. (pp. 730–31)

The trial court seemed to be questioning whether extended notice would really change the learning outcomes for students with disabilities whose IEPs did not already contain provisions for academic instruction at a high school level.

The students appealed this decision, and the appeals court held that students with disabilities

may be subject to a graduation test requirement. But the court also held that (1) parents and educators must make an informed decision regarding whether the tested content should be included in the students' IEPs and (2) 1½ years is not sufficient notice for students with disabilities to prepare for a graduation test. The court stated

[I]n an educational system that assumes special education students learn at a slower rate than regular division students, a year and a half at most to prepare for the [test] is insufficient. . . . [P]arents and teachers may evaluate students and conclude that energies would be more profitably directed toward areas other than [test] preparation . . . Here however parents had only a year to a year and a half to evaluate properly their children's abilities and redirect their educational goals. . . . [T]his was insufficient time to make an informed decision about inclusion or exclusion of training on [the tested] objectives. (p. 187)

The court also found that up to 90% of the tested skills had not been included on the students' IEPs, an 18-month notice period was not sufficient to remedy this deficiency, and expert opinion suggested inappropriate school district predictions of failure without giving the students with disabilities an adequate chance to learn the tested skills. Because a number of the plaintiffs had been out of school for several years, the court determined that it would be unreasonable to expect these students to return to school for further remediation and that they should be awarded diplomas as the remedy for the notice violation.

The Ambach Case

The *Bd. of Educ. of Northport v. Ambach* (1981) case was filed by the Commissioner of Education in New York against a local school district. The school district had awarded high school diplomas to two students with disabilities who had failed to pass the mandatory state graduation tests in reading and mathematics. The two students with disabilities – one with a neurologic disability affecting computation ability and the other who was trainable mentally retarded – had successfully completed their respective IEPs. The Commissioner sought to invalidate the diplomas

awarded to these students with disabilities and any others who had not passed the required tests.

In challenging the Commissioner's order on behalf of these students with disabilities, the school district alleged a Section 504 violation. In holding that there was no denial of a benefit based solely on the students' disabilities, the trial court stated

Section 504 may require the construction of a ramp to afford [a person who is wheelchair bound] access to a building but it does not assure that once inside he will successfully accomplish his objective. . . . Likewise, § 504 requires that a [disabled] student be provided with an appropriate education but does not guarantee that he will successfully achieve the academic level necessary for the award of a diploma. (p. 569)

The trial court also found no constitutional violations because students with disabilities are not a protected class, and education is not a fundamental right. Nevertheless, after acknowledging that it normally does not interfere in academic decisions, the court found that "[t]he denial of a diploma could stigmatize the individual [students] and may have severe consequences on their future employability and ultimate success in life" (p. 574). Although the testing requirement became effective 3 years after its enactment, the court found that these students' constitutional rights had been violated by a lack of timely notice because written state guidelines covering the testing of students with disabilities were not issued until the year the testing requirement became effective.

The Ambach (1982) appeals court disagreed, holding

[I]t is undisputed that [the two disabled students] were performing at a level expected of elementary students and the record is clear that their mental deficits were *functional*, thereby depriving them of ever advancing academically to the point where they could be reasonably expected to pass the [graduation test]. Thus, the issue of whether the three-year notice [period was adequate] is irrelevant. No period of notice, regardless of length, would be sufficient to prepare [these two disabled students] for the [graduation test]. However, . . . [there may exist disabled] students with remedial mental conditions who, with proper notice, might [have] IEPs adopted to meet their needs so as to

prepare them to pass the [graduation test]. As to them, . . . we hold that the three-school-year notice given here . . . was not of such a brief duration so as to prevent school districts from programming [their IEPs] to enable them to pass the [graduation test]. (pp. 687–688)

Thus, the issues of adequate notice of the graduation test (which must be at least as long for students with disabilities as for nondisabled students) and the curricular validity of the test (evidence that students with disabilities have been given the option to be taught what is tested) are important legal requirements for students with disabilities in graduation testing cases.

Recent Challenges to Graduation Tests by Students with Disabilities

Historically, special education students could receive a high school diploma by remaining in school for 12 years and completing their IEPs at whatever level of achievement the writers of those plans deemed appropriate. So there was significant resistance when graduation testing requirements were applied to special education students in states where there was no exception for students with disabilities unable to demonstrate high school level skills or when a waiver process provided for those students was not automatic. The *Rene v. Reed* (2001) case in Indiana and the *Chapman (Kidd) v. State Bd. of Educ.* (2003) case in California are examples of such challenges.

The Indiana Case

The Indiana graduation test, adopted in 1997, was a subset of the English language arts (ELA) and mathematics items from the tenth-grade state accountability test. Students with disabilities could obtain a waiver of a failed graduation subtest if their teachers certified proficiency in the tested skills supported by documentation, the principal concurred, remediation and retesting were completed as specified in their IEPs, the students maintained a “C” average, and their attendance was at least 95%.

Rene had been in special education since first grade. Her IEP indicated that she was in the diploma program and that she was to be provided with a reader and a calculator for all tests. The graduation test was administered to Rene several times without a reader or calculator and she did not pass. Rene and other special education students challenged the graduation test requirement claiming that it violated their constitutional rights because they had previously been exempted from standardized tests, had not been given sufficient notice of the graduation test requirement to adjust their curricula to include the tested skills, and would not qualify for a waiver because their teachers were unable to certify that they had achieved proficiency on the tested skills. They also alleged a violation of the IDEA because the state had refused to allow certain nonstandard test administrations provided for in their IEPs. The state countered that the 5 years’ notice given to schools and the 3 years’ notice given to students and their parents were sufficient, and the state was not required to provide nonstandard test administrations which altered the tested skills and produced noncomparable scores.

By the time the case was decided, Rene was the only remaining class representative of the original four because one had dropped out of school, one had received a waiver and one had passed the graduation test. Citing the *Debra P.* case, the *Rene* trial court held there was no constitutional violation because (1) these students had received adequate notice and (2) if there was a lack of curricular validity because these students had not been taught the tested skills, the appropriate remedy was additional remedial instruction, not the award of a diploma.

The students appealed the trial court’s decision. The appeals court concurred with the trial court holding that 3–5 years’ notice provided adequate preparation time and that the remediation opportunities provided to students with disabilities were an adequate remedy for any prior failure of the schools to teach them the tested skills. The appeals court also held that the state was not required to provide students with disabilities modifications of the graduation test that the state had determined would fundamentally

alter the tested skills and produce noncomparable scores. The appeals court also ruled that the IDEA did not require the state to honor all modifications in students' IEPs (e.g., a reader for the reading test), even though some students with disabilities were unable to pass the graduation test without those modifications. The appeals court found

We note . . . that the IDEA does not require specific results, but instead it mandates only that disabled students have access to specialized and individualized educational services. Therefore, denial of a diploma to [students with disabilities] who cannot achieve the educational level necessary to pass a standardized graduation exam is not a denial of the "free appropriate public education" the IDEA requires. Further, the imposition of such a standardized exam [without honoring all modifications provided in instruction] does not violate the IDEA where, as in the case before us, the exam is not the sole criterion for graduation. (p. 745, citations omitted)

The appeals court further distinguished the appropriateness of modifications in educational services provided to students with disabilities through their IEPs and the appropriateness of modifications on a graduation test designed to certify academic skills. The court stated

We cannot say the trial court erred to the extent it determined the [s]tate need not honor certain [modifications] called for in the Students' IEPs where those [modifications] would affect the validity of the test results. The court had evidence before it that the State does permit a number of [testing variations] typically called for in IEPs. However, the State does not permit [modifications] for "cognitive disabilities" that can "significantly affect the meaning and interpretation of the test score."

For example, the State permits accommodations such as oral or sign language responses to test questions, questions in Braille, special lighting or furniture, enlarged answer sheets, and individual or small group testing. By contrast, it prohibits [modifications] in the form of reading to the student test questions that are meant to measure reading comprehension, allowing unlimited time to complete test sections, allowing the student to respond to questions in a language other than English, and using language in the directions or in certain test questions that is reduced in complexity. (p. 746, citations omitted)

The appeals court also cited Office for Civil Rights (OCR) decisions by hearing officers that

states were not required to provide readers for reading comprehension tests.

The appeals court summarized its distinction between educational services and graduation testing as follows:

The IEP represents "an educational plan developed specifically for the [student that] sets out the child's present educational performance, establishes annual and short-term objectives for improvements in that performance, and describes the specially designed instruction and services that will enable the child to meet those objectives." The [graduation test], by contrast, is an assessment of the outcome of that educational plan. We therefore decline to hold that [a modification] for cognitive disabilities provided for in a student's IEP must necessarily be observed during the [graduation test], or that the prohibition of such [a modification] during the [graduation test] is necessarily inconsistent with the IEP. (pp. 746–747, citations omitted)

The California Case

The *Chapman (Kidd)* case was a challenge to a graduation test requirement applied to special education students in California. It changed names from the *Chapman* case to the *Kidd* case during the course of the litigation when some of the named plaintiffs met the graduation test requirement and new named plaintiffs were added to take their places.

The California Graduation Test. The California High School Exit Examination (CAHSEE) was enacted in 2001 and was first administered to high school students early in the second semester of tenth grade. It consisted of two untimed tests, ELA and mathematics. The ELA test consisted of 72 multiple-choice items and one open-ended writing prompt that covered the ninth, tenth, and a few eighth-grade ELA content standards with reading and writing weighted approximately equally. The mathematics test consisted of 80 multiple-choice items covering the sixth- and seventh-grade content standards plus the Algebra I content standards. The passing standard was set by the State Board of Education (SBE) at 60% of the 90 ELA total score points (raw score = 54) and 55% of the

Table 3.1 California Graduation Test Passing Rates by Class and Subgroup

CAHSEE passing rates (both tests)	Class of 2004 (%)	Class of 2006 (%)
Special education	24	35
Diploma track special education	43	63
All students	68	78

80 mathematics items (raw score = 44) on the 2004 tenth-grade census administration and the equivalent scaled score on future test forms (Cal. Dep't of Educ., 2004–2005).

Cumulative passing rates through eleventh grade for both sections of the CAHSEE (ELA and mathematics) calculated by the independent evaluator (HumRRO, 2006) for the classes of 2004 and 2006 are shown in Table 3.1.

The cumulative passing rates for special education students on a diploma track were estimated using the California Department of Education (CDE) 2005 high school graduation rate of 56% (for special education students not required to pass the CAHSEE) as a proxy for the unknown percent of special education students on a diploma track (Phillips, 2010, Chapter 5). For example, in Table 3.1, the 63% at the middle of the column labeled *Class of 2006* was calculated by adjusting the 35% passing rate for all special education students (diploma track plus nondiploma track) to its corresponding value for the estimated 56% of special education students on a diploma track ($(35/56) \times 100 = 63\%$).

SBE Decisions. Final decisions about what the graduation test was intended to measure were made by the SBE, the entity with statutory authority to do so. The Board determined that reading, decoding, and paper-and-pencil mathematics computation were part of the skills intended to be measured (SBE, 2000). These skills were included in prerequisite standards from earlier grades. The high school standards specified that standards from all earlier grades were part of the knowledge and skills high school students were expected to have learned.

The Board also determined that students whose cognitive disabilities described the inability to learn decoding, paper-and-pencil computation, or other tested skills would not be permitted to substitute different skills for the ones intended to be measured by the graduation test. In particular, the Board determined that use of a reader for the reading test and a calculator for the mathematics test would change the construct being measured and would not produce comparable scores to those from standard test administrations (5 California Code of Regulations (CCR) § 1217 (c)).

Students with disabilities in the Class of 2006 who expected to earn a high school diploma had 7 years' notice of the requirement to learn the tested skills and at least five opportunities to pass the graduation test by the time of their scheduled graduation. Special education students could also remain in school beyond their scheduled graduation to continue remediation and work toward earning a high school diploma.

Alternate Assessments. The special education students who challenged the graduation test complained that some students with disabilities were unable to access the CAHSEE because they needed an alternate assessment. However, at that time, the NCLB and IDEA statutes requiring alternate assessments for school accountability did not require alternate assessments for graduation tests. In addition, it was not lack of access that resulted from a feature of the test that prevented these students with disabilities from responding to the test questions (e.g., a visually impaired student who cannot see the printed words or a quadriplegic who cannot blacken the ovals on the answer sheet). Rather, it was a situation where the student had not learned the tested high-school-level content. The fact that the disability may have prevented the student from learning academic skills at a high school level was not an access problem that required an alternate assessment but an indication that the student was not otherwise qualified for a skills-based diploma.

Waiver Policy. The plaintiffs also claimed that some otherwise qualified diploma-track students with disabilities had been denied modifications

needed to access the CAHSEE. This was apparently a misunderstanding because California had a waiver process. Students with disabilities were permitted to take the CAHSEE with any modifications specified in their IEPs. Students with disabilities who received the equivalent of a passing score on the CAHSEE with modifications that altered the skills intended to be tested could receive a waiver of the graduation test requirement through a local process specified by law (CEC § 60851 (c)). For diploma-track students with disabilities who had completed appropriate high school level courses such as Algebra I and could pass the CAHSEE with modifications, the waiver process should have been straightforward.

The plaintiffs proposed that students with disabilities in the Class of 2006 and beyond who were unable to pass the CAHSEE be awarded a high school diploma if they had met all other graduation requirements. However, the state's experts countered that exempting special education students in the Class of 2006 from the graduation test requirement while continuing to require passage of the graduation test for regular education students would lower standards for the special education students and exclude them from the benefits of remediation and the opportunity to earn a skills-based diploma provided to their nondisabled peers (Phillips, 2010). It would have in effect created a skills-based diploma for nondisabled students and a rite of passage (seat time) diploma for students with disabilities contrary to the intent of the CAHSEE statute.

In addition, the state's experts argued that creating different graduation standards for disabled and nondisabled students would provide a substantial incentive for low-performing nondisabled students to be reclassified into special education so they could receive a diploma without passing the CAHSEE. It would also have provided an incentive for limited remedial resources to be redirected away from educating students with disabilities who did not have to pass the CAHSEE to nondisabled students who did. Moreover, creating a double standard would also have devalued the diplomas of students with disabilities in the

Class of 2006 who had passed the CAHSEE. Finally, the state and its experts argued that if some students with disabilities who were on a *diploma track* had not been taught or had not learned all the skills covered by the graduation test, the appropriate educational remedy was additional remedial education, not the award of an unearned skills-based high school diploma (Phillips, 2010, Chapter 5).

Legislative Intervention and Settlement. Prior to trial, the legislature intervened and by law provided that students with disabilities in the Class of 2006 who met certain procedural requirements and all other graduation requirements were entitled to a high school diploma regardless of whether or not they passed the state graduation test (Senate Bill 517, Jan. 30, 2006, amending CEC § 60851 and adding CEC § 60852.3). These special conditions entitling students with disabilities to high school diplomas were extended by the legislature to the Class of 2007 with the proviso that state authorities study the issue and present a recommendation to the legislature for the Class of 2008 and beyond.

On May 30, 2008, the court approved a final settlement of the case. The settlement required the state to contract for an independent study of students with disabilities who had failed the graduation test with accommodations or modifications but satisfied all other high school graduation requirements. The recommendations from the study report were to be considered in formulating the state's response and recommendations to the legislature. In return, all claims against the state by the plaintiff class members from the Classes of 2001 through 2011 were released (see www.cde.ca.gov).

ELL Challenges to Graduation Tests

The 1999 *Test Standards* indicate that making a professional judgment about the validity and appropriateness of administering graduation or accountability tests to ELLs in English requires evaluation of the purpose for testing and the construct relevance of English language proficiency. Standard 9.3 from the chapter on *Testing*

Individuals of Diverse Linguistic Backgrounds states

The test generally should be administered in the [student's] most proficient language, unless proficiency in the less proficient language is part of the assessment. (p. 98)

ELL graduation test litigation has focused on evaluating the construct relevance of English language proficiency and the appropriate remedy for any OTL deficiencies.

The dilemma for policymakers in selecting appropriate testing condition variations for ELLs is similar to that faced in deciding how to test students with disabilities. There are competing policy goals and costs on opposite sides of each argument. Policymakers must determine which of the competing goals is most consistent with the purpose of the state test, is in the best interests of its students, and has affordable costs. For example, some policymakers might argue that by the time students reach high school, they should be receiving instruction in English. An exception might be newly arrived students who may not be literate in either English or their native languages if they have not received formal schooling in their home countries. Newly arrived students, who must master academic content and achieve proficiency in English, may need to spend additional years in school to earn a high school diploma that requires passing a graduation test administered in English. Alternatives, such as translated tests and testing waivers, may reduce student frustration and allow ELLs to graduate earlier, but they have the disadvantage of awarding a credential to students who have not yet achieved the knowledge and skills in English needed for postsecondary education and employment.

Cases in Texas (*GI Forum*, 2000) and California (*Valenzuela v. O'Connell*, 2006) have dealt with the application of graduation testing requirements to ELLs. California state law established a presumption that ELLs in elementary grades would be taught in English, but Texas law established a presumption of bilingual elementary education unless the student was proficient in English. Despite these differing views on ELL education, both states successfully defended

their high school graduation tests administered to ELLs in English.

The Texas ELL Case

The Texas Education Code required the provision of bilingual education for ELL elementary students and English as a second language (ESL) instruction for high school students (TEC § 29.053). At each elementary grade level, districts enrolling at least twenty ELL students with the same primary language were required to offer bilingual instruction (TEC §§ 29.055, 29.056). In schools required to offer it, bilingual instruction was expected to be a full-time, dual-language program with basic skills instruction in the primary language, "structured and sequenced English language instruction" and enrichment classes, such as art and music, with non-ELLs in regular classes. Electives could be taught in the primary language. Detailed criteria, including a home language survey, English language proficiency testing, and primary language proficiency testing, were provided for identifying eligible students. Multiple primary languages were represented in Texas districts, but the predominant one was Spanish. State accountability tests were administered in both English and Spanish at the elementary grades but in English only at the upper grades. Hispanic ELLs challenged the refusal of the state to provide a Spanish language translation for the graduation test.

State testing data indicated that ELLs had made progress on the Texas graduation test. For example, in 1994, 14% of ELLs passed all subtests of the graduation test at its initial administration in tenth grade, but by 1999, with more ELL students participating, 31% had done so. The state's ELL expert opined the following:

To suggest that students should graduate without demonstrating minimal knowledge and skills on a uniform measure is not acceptable for the current requirements of the technological and information age job market or for pursuing higher education . . . A policy of separating language minority students, many of whom are native born, from the rest of the student population when the [graduation test] is

administered is more likely to stigmatize and negatively impact the self-esteem of these students than is their inclusion in the tests. (Porter, 2000, p. 409)

The *GI Forum* court upheld the graduation test administered in English for all students, including ELLs, stating

[T]he Court finds, on the basis of the evidence presented at trial, that the disparities in test scores do not result from flaws in the test or in the way it is administered. Instead, as the plaintiffs themselves, have argued, some minority students have, for a myriad of reasons, failed to keep up (or catch up) with their majority counterparts. It may be, as the [State] argues, that the [graduation test] is one weapon in the fight to remedy this problem. At any rate, the State is within its power to choose this remedy. (pp. 682–683, citations omitted)

The California ELL Case

In the *Valenzuela* case, the California High School Exit Examination (CAHSEE), effective for the Class of 2006, was challenged in state court by a group of ELLs just prior to their scheduled graduation in the Spring of 2006. The ELLs sought a court order barring implementation of the graduation test requirement for students in the Class of 2006. They argued that they had not received an adequate opportunity to learn the tested material because they attended schools lacking fully aligned curricula and fully credentialed teachers. They asserted that this lack of OTL was a denial of their (state) fundamental right of equal access to public school education. In addition, they alleged that ELLs had been disproportionately affected by the scarcity of resources in poor districts.

The state argued that it was appropriate and valid to administer the graduation test in English to ELLs, even when English was not the ELLs' most proficient language, because state law required English proficiency for receipt of a high school diploma and the purpose of the CASHEE was to determine academic proficiency in English. The California Code of Regulations provided specific rules for the administration of the CAHSEE to ELLs with "accommodations" including supervised testing in a separate

room, additional supervised breaks and extra time within a testing day, translated directions, and translation glossaries if used regularly in instruction (5 CCR § 1217).

The *Valenzuela* lawsuit was filed in February 2006. On May 12, 2006, the trial court granted the requested injunction barring the state from imposing the testing requirement on *any student* in the Class of 2006. On May 24, 2006, the California Supreme Court stayed the injunction pending review and decision by the appeals court. The ELLs' request to the appeals court for an immediate hearing was denied, and the graduation test requirement remained in force for the Class of 2006.

In August 2007, the appeals court issued its decision vacating the injunction issued by the trial court. The appeals court noted that neither the graduation test requirement nor the validity of the CAHSEE was being challenged. Further, the court held that even if some ELLs had not received an adequate opportunity to learn the tested material, the appropriate remedy was provision of the missed instruction, not removal of the test requirement or the award of diplomas by court order. The court stated

Within the borders of California, until our schools can achieve [academic parity, the CAHSEE] provides students who attend economically disadvantaged schools, but who pass the [graduation test], with the ability to proclaim empirically that they possess the same academic proficiency as students from higher performing and economically more advantaged schools. Granting diplomas to students who have not proven this proficiency debases the value of the diplomas earned by the overwhelming majority of disadvantaged students who have passed the [test]... We believe the trial court's [order] erred by focusing its remedy on equal access to *diplomas* rather than on equal access to *education* (and the funding necessary to provide it)... The purpose of education is not to endow students with diplomas, but to equip them with the substantive knowledge and skills they need to succeed in life. (p. 18, 27)

The appeals court also found that the scope of the remedy (removal of the test requirement for *all students*) was overbroad because it provided a potential windfall to students who could not trace their test failure to inadequate school resources. The court stated

[T]he ostensibly interim relief of forcing the “social promotion” of [Class of 2006 students], by ordering that they be given diplomas, in fact does not maintain the status quo of the litigation, but ends it. Surely the trial court did not expect that if [the state] ultimately prevailed in the litigation, students would give back the diplomas they had received under the mandate of the court’s [order]. . . . [D]irecting [the state] to give [Class of 2006 students] diplomas . . . would inadvertently have perpetuated a bitter hoax: that the [court-ordered diplomas] somehow would have equipped them to compete successfully in life, even though they had not actually acquired the basic academic skills measured by the CAHSEE. . . . Plaintiff’s virtually concede[d] the overbreadth of the trial court’s injunction in their argument that some [class members] “actually know the material, but do not pass the [graduation test] due to test anxiety.” But plaintiffs have not argued, much less established, that there is any *constitutional* violation involved in depriving a student of a diploma when he or she has in fact received the educational resources required to pass the CAHSEE, but has not been able to do so because of “*test anxiety*.” (p. 19, 28, 30)

The court’s opinion concluded by urging the parties, with the active assistance of the trial court, to work together to provide all seniors in the Class of 2007 and beyond who had not yet passed the graduation test an equal access to effective remedial assistance. The case was settled in October 2007 with legislation providing up to 2 years of extra help beyond high school for students unable to pass the high school graduation test (Jacobson, 2007).

Accountability Testing Challenges

Challenges to state accountability tests used to evaluate schools have focused on the appropriateness of testing ELLs in English rather than their primary language. Early challenges were based on state accountability systems, and more recent litigation has focused on state tests mandated by the No Child Left Behind (NCLB) Act (2002). A discussion of cases from California and Pennsylvania follows after a brief review of the NCLB Act and its ELL provisions, ELL testing variations and related policy issues.

No Child Left Behind (NCLB) Act

In 2002, as a condition for receipt of Title I federal funding for remedial education for low-performing, disadvantaged students, the federal NCLB Act mandated that each state establish challenging academic content standards and annual assessments in reading and mathematics for all students in grades 3–8 and at least one grade in high school by 2006. The challenging academic content standards were required to describe what students were expected to know and do at each grade level and be the *same for all students*. Achievement standards applied to grade-level assessments which were aligned to the reading and mathematics state content standards were required to include at least three levels: advanced, proficient, and basic. NCLB assessments were also required to be valid, reliable, and consistent with professional standards and had to include alternate assessments for students with the most severe cognitive disabilities. Subsequently, U.S. Department of Education (U.S. DOE) regulations also permitted modified assessments (with content at grade level) for students with disabilities who did not qualify for alternate assessments but for whom the regular statewide assessments were not appropriate due to “persistent academic difficulties” (NCLB Regulations, 2006).

The NCLB Act also required each state to develop accountability measures of school progress toward the goal of *all* students achieving proficient test scores by 2014. The adequate yearly progress (AYP) determination for each school was required to be based on

- primarily test scores,
- plus graduation rates and at least one additional academic indicator,
- for all students,
- and for ethnic, ELL, students with disabilities (SD), and economically disadvantaged (ED) subgroups of statistically reliable and not personally identifiable size,
- using a baseline established by the higher of the 2002 subgroup with the lowest percent

proficient and above *or* the state's 20th-student percentile rank,

- and annual state proficiency targets meeting a timeline of all students proficient by 2014,
- with consequences for consistently failing schools, and
- participation in NAEP fourth and eight grade reading and mathematics assessments.

States were permitted to count up to 1% of their students as proficient on alternate assessments and up to 2% of their students as proficient on modified assessments.

NCLB ELL Provisions

Under the NCLB Act and its Regulations, all ELLs were required to be tested, but states were permitted to exempt from AYP calculations the scores of ELLs in the United States less than one year. ELLs were required to be tested

- on the same *grade-level* content standards as *all* other students,
- with measures most likely to yield *valid* and *reliable* results,
- with reasonable *accommodations*,
- to the extent practicable, in the language and form most likely to yield accurate data until they are English proficient. (20 U.S.C. § 6311) [emphasis added]

Similar to the decisions about content standards, proficiency standards, subgroup sizes for reporting results, and annual school targets that were left to the states, with respect to ELLs, the NCLB Act and its Regulations permitted each state to decide what was practicable, the criteria for English proficiency, *reasonable accommodations* for ELLs, and the language and form of testing that best aligned to the content standards required of all students in the state.

The NCLB permitted, but did not require, a state to use alternative tests for ELLs. For states choosing to administer alternative tests to ELLs, the NCLB Act and its Regulations specified that such tests must be valid, reliable, and aligned to content standards at grade level. Through its peer-review process, the U.S. DOE signaled its interpretation that states administering alternative

tests to ELLs for NCLB accountability purposes were required to provide evidence of alignment to grade-level content standards and comparability to the regular, on-grade-level tests administered to non-ELLs. The compliance status letters issued to states in July 2006 indicated that the U.S. DOE questioned 18 states' evidence of grade-level alignment and comparability of primary language or simplified English tests administered to ELLs.

ELL "Accommodations" and Modifications

Many states provide "accommodations" for ELL students. However, the term *accommodation* is probably inappropriate because lack of language proficiency is not a disability. Disabilities are generally thought to describe characteristics over which students have no control and generally are not reversible over time. However, ELL students can become proficient in English through instruction.

To the extent that a test intends to measure content skills in English, any nonstandard administration that provides assistance with English is providing help with a skill intended to be measured. Thus, the nonstandard administration is compensating for a construct-relevant factor, not an extraneous factor. This is clearly contrary to the definition of an accommodation. Therefore, if testing variations (i.e., bilingual dictionaries or responses in the native language) are provided to give ELLs greater access to a state test, they should be labeled and treated as modifications.

Majority and Minority ELLs

A few states have provided translated tests for some ELLs. However, existing resources typically support, at most, a handful of translated tests that meet professional standards. In many cases, there may be only enough resources to translate the state test for the majority ELL language group.

The equal protection clause of the U.S. Constitution requires similarly situated persons to be treated equally. Court cases based on this clause have invalidated educational programs that favored a majority ethnic group. In particular, any allocation of benefits based on race or ethnicity has been considered suspect, and the high standards required by the courts to justify such programs have rarely been met.

A testing program that provides a benefit of primary language testing to ELL students who speak one non-English language (Language 1), but denies that same benefit to ELL students who speak all other non-English languages, has treated similarly situated students unequally. In the context of the ELL classification, majority group ELLs (Language 1 speakers) would be treated differently than minority group ELLs (speakers of all other non-English languages). However, both majority and minority ELL language groups are similarly situated in that they lack proficiency in English. Using numerical dominance to justify providing translated tests in some languages and not others is unfair to the ELL students who do not receive this benefit and may constitute an equal protection violation.¹¹

Construct Shift

The argument is often made that students should not be assessed in a language in which they are not proficient (Fraser & Fields, 1999). The intended reference is to nonnative speakers of English. However, there are native speakers who perform poorly on tests given in English because they are also not proficient in English. Yet, for the native speaker, test administrators rarely worry about the effects of language proficiency on test

performance, and typically these students do not have the effects of poor language skills removed from their scores. Thus, the argument seems to be that the effects of lack of English proficiency should be removed from test scores for nonnative speakers but not native speakers, although both may need intensive additional instruction in English to achieve proficiency. This is another example of a construct shift – altering the tested construct based on group membership rather than following the guidelines in the *Test Standards* that refer to the tested construct as a property of the test. Cases from California and Pennsylvania considered the issue of construct relevance in deciding whether, for state accountability purposes, ELLs whose least-proficient language was English could be tested in English.

California State Law Case (2000)

A 1997 California law required that an achievement test designated by the State Board of Education be administered annually to all students in grades 2 through 11. The designated achievement test was the Stanford Achievement Test Ninth Edition (SAT-9) plus an additional set of items selected to measure state standards not measured by the SAT-9. Test scores were used for school accountability with scores for students enrolled for less than 12 months excluded from the computation of a school's accountability index. There were no state-imposed consequences for individual students.

The case began when the San Francisco Unified School District (SFUSD) refused to administer the SAT-9 in English to ELLs with less than 30 months (three school years) of public school instruction (ELLs<30) unless specially designated by their teachers. The state sought a court order to enforce the testing legislation. The Oakland, Hayward, and Berkeley Unified School Districts joined SFUSD in the lawsuit claiming that administration of the SAT-9 to ELLs<30 was unfair because the test measured English language proficiency in addition to content knowledge. They argued that ELL students who lacked English proficiency should

¹¹ Psychometric standards support testing students in the language in which they receive instruction. However, in many cases, bilingual instruction may only be available for a single language. Therefore, the question still remains whether it is fair and equitable to provide primary language instruction and testing to the majority ELL language group but not to other minority ELL language groups.

either be tested in their primary language or be exempt from testing. State law provided that ELLs with less than 12 months of public school instruction (ELLs<12) be administered a second achievement test in their primary language when available. ELLs<12 were also eligible for testing modifications when tested in English (*Cal. Dep't of Educ. v. San Francisco Unified Sch. Dist.*, 1998).

Plaintiffs' experts opined that ELLs<30 would suffer psychological harm from taking the SAT-9 in English because their low scores would be stigmatizing, would diminish their self esteem, and would cause them to be inappropriately placed in special education programs and portrayed as having inferior employment skills. In addition, plaintiffs' experts argued that ELLs<30 would score at the chance level resulting in unreliable test scores. In response, the State argued that a reasonable interpretation of state law indicated an intent to measure academic skills in English, a fair accountability system required the inclusion of all students, the districts and their ELL students benefited from the receipt of state funds targeted toward the improvement of academic skills for low-scoring students, the districts failed to show that any ELLs were harmed by the test administration, and available test data demonstrated that most ELLs scored above chance and their test scores were reliable. In addition, the state argued 30 months was an arbitrary exclusion criterion and that there was significant overlap in the performance of ELLs<30 and ELLs<30 in Plaintiff Districts. Further, over the 3-year period the SAT-9 had been administered statewide, ELLs had made substantial gains in some districts (Phillips, 2010, Chapter 6).

The *SFUSD* case subsequently settled out of court just prior to trial. In the settlement, the districts agreed to administer the state-designated achievement test to all ELL students as provided by state law. The state agreed to clarify the rules regarding educator communications with parents about exemptions, to consider English language proficiency test scores, among other factors, when evaluating school waiver requests and to make other minor modifications to program procedures.

NCLB Cases

The major testing cases under the NCLB Act to date have involved ELLs. Two state courts, Pennsylvania (*Reading Sch. Dist. v. Pa. Dep't of Educ.*, 2004) and California (*Coachella Valley v. California*, 2007), have ruled on the mandates of this federal law. Among other rulings, both state courts held that the NCLB Act did not require states to provide primary language testing for ELLs. To some extent, the California *Coachella Valley* case appeared to be a reprise of the same issues litigated in the *SFUSD* case, only with a different set of California school districts complaining about the state requirement to test ELLs in English.

The Reading School District Case

In a challenge by a school district with 69% economically disadvantaged and 16% ELL students, a Pennsylvania court determined that the state testing agency appropriately exercised its discretion under the NCLB Act when it made psychometric decisions related to ELL testing policy. Specifically, the court upheld the state's determination that primary language testing was not practicable with 125 languages represented in Pennsylvania schools and found no NCLB violation because primary language testing was not mandatory under the NCLB.

The Coachella Valley Case

Nine California school districts enrolling large numbers of Spanish-speaking ELLs asked a state court to order the state to provide NCLB tests for ELLs in Spanish or to provide these students with simplified English versions of the tests. The districts argued that the statutory language of the NCLB Act required the state to provide primary language testing for ELLs. On May 25, 2007, the court denied the request.

The court agreed with the state argument that the NCLB provided discretionary author-

ity to states to determine appropriate testing for ELLs and held that California's decision to test ELLs in English was not an abuse of its discretion. Therefore, the court held that it did not have the legal authority to issue an order to the state requiring a change in its ELL testing policy.

In reference to primary language testing of ELLs, the NCLB Act used the qualifying phrase "to the extent practicable." The American Heritage Dictionary defines *practicable* as "feasible and capable of being used for a specified purpose." The state argued that using primary language tests in Spanish as an alternative accountability test for some ELLs was not practicable in California because of the following:

- Existing Spanish language tests could not be used to assess ELLs with the *same* ELA and mathematics content and performance standards at *grade level* as non-ELLs as required by the NCLB accountability provisions and *in English* as provided by California law.
- It was not feasible to provide the same benefit to the significant numbers of California ELLs who spoke other primary languages due to insufficient resources to produce alternative tests in all relevant languages. Providing primary language tests for ELLs who spoke one language but not for ELLs who spoke other languages would have been contrary to the *Test Standards* fairness requirement that "The testing [process] should be carried out so that [students] receive comparable and equitable treatment . . ." (Standard 7.12, p. 84). Moreover, due to differences in language and culture likely to produce differential alignment to the content standards, inherent difficulties in establishing equivalent performance standards, and inconsistency with the mandates of California law requiring ELLs to be instructed primarily in English, satisfying NCLB peer review with even a single primary language test may have been unattainable in California.
- Providing primary language tests for ELLs who spoke one language but not for ELLs who spoke other languages may have been

an equal protection violation because ELL students who were similarly situated (lacked English language proficiency) would have been treated differently (Phillips, 2010, Chapter 6).

As in the *Valenzuela* case, the State in the *Coachella Valley* case argued that the appropriate remedy for ineffective instruction was additional, improved, remedial instruction, not less valid test scores that indicated achievement of different skills than intended. In refusing to issue an order compelling the state to change its ELL testing policy, the court stated

[G]iven that California has determined to teach students who lack English proficiency largely in English, it cannot be said that a decision to assess these same students in English for purposes of NCLB is arbitrary and capricious.

Further, given the extensive range of possible primary languages of students lacking English proficiency, it is certainly neither arbitrary nor capricious for California to determine that translation and evaluation of assessments in multiple languages is not practicable and that, accordingly, administration of assessments will be in English, the single language confirmed by the voters through [a ballot initiative] as the "official" language of our educational system. . . .

The task for this court . . . is not to choose among competing rational alternatives and then mandate the judicially chosen one. To the contrary, decisions such as how to assess student performance for purposes of NCLB are best left to other branches of the government that are better suited to such matters and, so long as they do not act in an arbitrary, capricious, unlawful or procedurally unfair manner, great deference must be afforded to their decisions. . . .

California's manner of conducting student assessment for the purposes of NCLB does not violate any ministerial duty created by statute, nor as a matter of law does it constitute an abuse of any discretionary authority. Therefore, . . . [the districts'] motion [to compel a change in policy] is denied. (p. 24, 27)

Recommendations

The following recommendations are based on current legal requirements and psychometric principles. Specific implementation details may

vary depending on the configuration of a state testing program and its purposes, implementing state legislation and administrative regulations. These recommendations can be used as a starting point for the development of state nonstandard test administration policies for students with disabilities and ELLs.

1. Require the IEP/504 committees or local ELL program directors to select one of the following exhaustive, mutually exclusive testing condition categories for each student with a disability or ELL student. Allow test administrators to provide the accommodations in Category II (below) for nondisabled students when appropriate.
 - I. Standard administration conditions.
 - II. Accommodated administration using one or more of the testing variations designated by subject matter test in the state test administration manual as accommodations that preserve the intended construct and result in comparable scores. Scores from test administrations in this category receive the same interpretive information as Category I scores and may be aggregated with Category I scores.
 - III. Modified administration using testing condition variations that are not on the state test accommodations list but are regularly provided in the student's instructional program. These may be modified tests as defined by U.S. DOE NCLB Regulations. Students tested with modifications should receive score reports with information corresponding to the particular test taken. If the on-grade-level state test has been modified, on-grade-level achievement levels, pass/fail designations, or associated normative information should not be reported and Category III scores should not be aggregated with Category I and II scores unless the modified tests have been statistically linked to the regular tests.
 - IV. Alternate assessment for students who cannot access the regular test with modifications or have been instructed with an alternative curriculum that consists of enabling or essence skills related to the academic skills measured by the regular state test. These may be alternate assessments as defined by U.S. DOE NCLB Regulations. Category IV scores should be reported and interpreted separately.
2. Develop a list that classifies specific nonstandard test administrations for the state test as accommodations or modifications. Consider written state content standards, test specifications, preservation of the intended skills (construct relevance), and score comparability. Aggregate and report scores accordingly. If a school/parent requests a testing condition variation not on the list, refer them to a contact person who has been designated to receive and act on written requests. The contact person may be aided by outside consultants in making a decision on each request. Add these decisions to the appropriate list for the next state test administration.
3. Permit the parent(s)/guardian(s) of any student (with or without a disability, ELL, or non-ELL) to request a category I, II, III, or IV state test administration by signing a written form that includes full disclosure of the options and their consequences. Require IEP/504 committees to include a form in the IEP, signed by school personnel and the parent(s)/guardian(s), documenting consideration of the available testing options and the final decision of the committee.
4. Collect information about the specific disability or ELL status and specific accommodations or modifications on the student answer sheet. Conduct separate research studies when sufficient numbers of students and resources are available.
5. Consider paying the costs of accommodated test administrations from state testing program funds and requiring local districts to pay at least some of the costs of modified test administrations (except state NCLB 1 and 2% tests). Provide detailed written guidelines to aid local districts in making accommodations/modifications decisions.

6. For ELLs, use the required NCLB English language proficiency test scores and primary language achievement test scores (where available for ELLs receiving bilingual instruction) to augment the interpretation of the state test scores. This information will assist users in judging the effects of language proficiency on the achievement of academic skills in English and assist them in determining appropriate future instructional strategies for these students.
7. If policymakers decide to allow scores from modified test administrations to be interpreted the same as scores from standard administrations, they should acknowledge that a benefit is being conferred on eligible students, and adopt policies that require appropriate written documentation and evaluation to determine which students qualify for the benefit. Eligibility criteria for which written documentation should be required include (a) certification of the disability by a trained professional, (b) confirmation of regular use of the modification by the student in the classroom, (c) explanation of the rationale for and relationship of the requested modification to the specific disability, and (d) certification of the impossibility of the student accessing the test without the requested modification. School and/or department officials may be aided by outside consultants when evaluating written documentation and should follow a written policy. Consider alternate credentials such as Certificates of Completion or transcript and/or diploma notations that identify scores obtained with *modifications*. Possible notations include general statements such as “tested with modifications” or descriptions of the modification (e.g., reader, calculator) without identifying the specific disability.
8. Establish an appeals procedure for persons desiring to challenge denial of a nonstandard test administration. Such a procedure might begin at the local school level and be reviewable by state officials upon request.
9. Create a written state ELL testing policy that answers the following questions for the state NCLB tests, graduation tests, and other state tests, if any, by content area (e.g., ELA, mathematics, science).
 - A. Is English language proficiency construct relevant for the tested content? If yes, administer the regular, on-grade-level test to ELLs. If no, consider alternatives.
 - B. If an alternative test is to be provided to ELLs, will it be a primary language test, translated test, or simplified English version of the regular test?
 - C. If an alternative test is provided, what are the criteria for ELL eligibility to take the alternative test in place of the regular test (e.g., length of time in U.S. schools, level of English language proficiency)?
 - D. How will the state establish comparability of its ELL alternative tests to its regular, on-grade-level tests (e.g., alignment studies, linking studies)?
 - E. What is the timeline for developing the state’s ELL alternative tests? Will the same test development procedures the state uses for its regular tests be followed? If not, how will consistency with psychometric standards be assured?
 - F. Has the state established expectations for annual progress of ELLs in English language proficiency? Are policies in place to identify schools where ELLs are remaining at beginning levels of English language proficiency for too long?
 - G. Does the state have policies and procedures for monitoring schools’ provision of appropriate English language and state content standards instruction to ELLs?
 - H. What assistance will the state provide to educators for professional development to administer the ELL alternative tests, to identify weaknesses based on test results, and to design/develop appropriate instructional programs to correct deficiencies?
 - I. If the state has a graduation test, will ELLs receive the same skills-based diploma as nonELLs, be awarded a certificate of completion if unsuccessful by their scheduled graduation, be permitted to remain in high school an extra year(s) or have

other options for earning a high school diploma after their senior year, and be allowed to substitute an alternative test or be exempted from the testing requirement?

- J. Have ELL state course credit requirements for a high school diploma been aligned with the skills tested on the state's graduation test, if any?
- K. Will the policies for recent immigrants be different than for other ELLs?
- L. Which ELL testing modifications (e.g., extra breaks, individual administration, word glossaries, reader) will be permitted on the regular or alternative tests and what criteria (e.g., regular use for classroom tests) must be satisfied for an ELL to qualify for a modification? Has the state (or have local schools) established an appeals procedure for persons desiring to challenge an unfavorable modifications decision?

Conclusion

Maintaining a defensible testing program is a challenging but achievable task. Tests are visible and accessible targets for those who disagree with state educational policy decisions. Challenges can come from a variety of groups on an assortment of issues. To be prepared, state testing programs must follow legal and psychometric standards, comprehensively document program decisions and activities, and work closely with legislators, administrators, and educators. Cooperation between staff members from the state agency, local school districts, and the testing contractor is essential for success. Carefully crafted policy documents covering nonstandard test administrations and ELL testing policies may assist states and school districts to provide accessible tests for special populations that satisfy legal and psychometric guidelines for construct preservation and comparable scores. When policymakers decide to treat test scores from modified test administrations the same as those from regular and accommodated test administrations, test score consumers should be alerted that the modified test scores are not comparable (or evidence of comparability is

lacking) and be assisted in interpreting and using these scores appropriately.

References

- Advocates for Special Kids v. Oregon Dep't of Educ., Settlement Agreement, No. CV99-263 KI (2001, February 1).
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing [Test Standards]*. Washington, DC: American Psychological Association.
- Americans with Disabilities Act (ADA), Pub. L. No. 101-336, 42 U.S.C. §12101 *et seq.* (1990).
- Americans with Disabilities Act (ADA) Regulations, 28 C.F.R. § 36.309 (1992).
- Anderson v. Banks, 520 F.Supp. 472 (S.D. Ga. 1981), *reh'g*, 540 F.Supp. 761 (S.D. Ga. 1982).
- Bd. of Educ. of Northport-E. Northport v. Ambach, 436 N.Y.S.2d 564 (S.C. N.Y. 1981), *rev'd* 458 N.Y.S.2d 680 (N.Y. App. 1982).
- Bd. of Educ. v. Rowley, 458 U.S. 176 (1982).
- Brookhart v. Illinois State Bd. of Educ., 534 F. Supp. 725 (C.D. Ill. 1982), *rev'd*, 697 F.2d 179 (7th Cir. 1983).
- Cal. Dep't of Educ., Interpreting CAHSEE Scores 2004–2005, available at www.cde.ca.gov
- Cal. Dep't of Educ. (CDE) v. San Francisco Unified Sch. Dist. (SFUSD), No. C-98-1417 MMC (N.D. Cal. 1998); San Francisco Unified Sch. Dist. (SFUSD) v. State Bd. of Educ. (SBE), Settlement Agreement, No. 99-4049 (Cal. Sup. Ct., Nov. 2000).
- Cal. State Bd. of Educ. (2000), Item 28 of December 6–7 Board Minutes, Item 31 of October 10–11 Board Minutes.
- Chapman (Kidd) v. Cal. Dep't of Educ., 229 F. Supp.2d 981 (N.D. Cal. 2002), No. C 01-01780 CRB (N.D. Cal. Mar. 12, 2003, Sept. 5, 2003).
- Coachella Valley v. California, No. CPF-05-505334 (Cal. Sup. Ct., May 25, 2007).
- Debra P. v. Turlington, 644 F.2d 397 (5th Cir. 1981), *aff'd*, 730 F.2d 1405 (11th Cir. 1984).
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., et al. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children*, 76(4), 475–495.
- Fraser, K., & Fields, R. (1999, February). *NAGB public hearings and written testimony on students with disabilities and the proposed voluntary national test October–November 1998, Synthesis Report*.
- G.I. Forum v. Texas Education Agency (TEA), 87 F.Supp.2d 667 (W.D. Tex. 2000).
- Golden, Daniel (2000, January 21). *Evening the score: Meet edith, 16; she plans to spell-check her state writing test*. The Wall St. J., at A1.

- Gorin, J. (2010, May). *Enhanced assessment item development: An item difficulty modeling approach*. Paper presented at the NCME annual meeting, Denver, CO.
- Huesman, R. Jr., & Frisbie, D. A. (2000). *The validity of ITBS reading comprehension test scores for learning disabled and non learning disabled students under extended-time conditions*. Paper presented at the AERA annual meeting, New Orleans, LA.
- HumRRO (2006, March 28). *January 2006 Update*, Press Release, Cal. Dep't of Educ.
- Individuals with Disabilities Education Act (IDEA), Pub. L. No. 102-119, 20 U.S.C. §1400 *et seq.* (1991).
- Jacobson, L. (2007). *California offers long-term help with exit exams*. Educ. Week, Oct. 24, 2007, at 23.
- Meloy, L. L., Deville, C., & Frisbie, D. (2000, April). *The effect of a reading accommodation on standardized test scores of learning disabled and non learning disabled students*. Paper presented at the NCME annual meeting, New Orleans, LA.
- Murphy v. United Parcel Service (UPS) Inc., 527 U.S. 516 (1999).
- No Child Left Behind (NCLB) Act, 20 U.S.C. §§ 6301–6578 (2002).
- No Child Left Behind (NCLB) Regulations, 34 C.F.R. § 200.1 *et seq.* (2006).
- Phillips, S. E. (1993, March 25). *Testing accommodations for disabled students*. 80 Ed. Law Rep. 9.
- Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled Rights. *Applied Measurement in Education*, 7(2), 93–120.
- Phillips, S. E. (2000). G.I. Forum v. TEA: Psychometric evidence. *Applied Measurement in Education*, 13, 343–385.
- Phillips, S. E. (2002). Legal issues affecting special populations in large-scale testing programs. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students* (Vol. 109, pp. 109–148), Mahwah, NJ: Lawrence Erlbaum.
- Phillips, S. E. (2010). *Assessment law in education*, Chapters 1, 2, 5, 6 and 9. Prisma Graphics, Phoenix, AZ, available at www.SEPhillips.dokshop.com
- Phillips, S. E. & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., Vol. 733, pp. 733–755) Westport, CT: American Council on Education/Praeger.
- Porter, R (2000). Accountability is overdue: Testing the academic achievement of limited-english proficient (LEP) students. *Applied Measurement in Education*, 13(4), 403.
- Reading Sch. Dist. v. Pa. Dep't of Educ., 855 A.2d 166 (Pa. Commw. Ct. 2004).
- Rene v. Reed, 751 N.E.2d 736 (Ind. App. 2001).
- Section 504 of the Rehabilitation Act, 29 U.S.C. §701 *et seq.* (1973).
- Section 504 Regulations, 45 C.F.R. § 84 *et seq.* (1997).
- Southeastern Community College v. Davis, 442 U.S. 397,406 (1979).
- Stoica, W. M. (2010, May). *Recommendations for AA-MAS item modifications: Utilizing information from focus groups and surveys with special educators*. Paper presented at the NCME annual meeting, Denver, CO.
- Sutton v. United Air Lines, Inc., 527 U.S. 471 (1997).
- Tindal, G. (1999, June). *Test accommodations: What are they and how do they affect student performance?* CCSSO large scale assessment conference, Snowbird, UT.
- Valenzuela v. O'Connell, No. JCCP 004468 (Cal. Super. Ct., May 12, 2006), rev'd, O'Connell v. Superior Court, 47 Cal. Rptr.3d 147 (Cal. App. 2006).

IEP Team Decision-Making for More Inclusive Assessments: Policies, Percentages, and Personal Decisions

Naomi Zigmond, Amanda Kloo,
and Christopher J. Lemons

IEP Team Decision Making for More Inclusive Assessments

Since the 2001–2002 school year, the accountability provisions of the No Child Left Behind Act (NCLB, 2001) have shaped much of the work of public school teachers and administrators in the United States. NCLB required each state to develop content and achievement standards in several subjects, to administer tests to measure students' attainment of those standards, to develop targets for student performance on those tests, and to impose a series of sanctions on schools and districts that did not meet the targets. Together, the standards, assessments, and consequences constitute a standards-based accountability system. State assessments are the mechanism for determining whether schools have been successful in teaching students the knowledge and skills defined by the content standards. The accountability provisions ensure that schools are held accountable for educational results. Many states had such a system in place before NCLB took effect, but since 2001–2002, every state in the United States has had to develop and implement a standards-based accountability system that meets the requirements of the law. This mandate has affected every public school

student, every public school, and every district in the nation.

The origins of federally required accountability for educational outcomes date to the 1994 reauthorization of the Elementary and Secondary Education Act (Improving America's Schools Act, Public Law 103-382, October 20, 1994). An often overlooked stipulation of the 1994 ESEA reauthorization was that each State ensures "participation in such assessments of *all* students" (Improving America's Schools Act, 1994, Title I, Subpart 1, section 1111 (b)(3)(F)(i)). Previously, students with disabilities had been exempt from participation in school level or district level standardized testing requirements. On-grade level performance was not expected of these students (Crawford, Almond, Tindal, & Hollenbeck, 2002; Thurlow, Seyfarth, Scott, & Ysseldyke, 1997; Ysseldyke & Thurlow, 1994). And, because the primary target of the ESEA regulation was the Title 1 program in a school or district, and accountability in federal law (PL 94-142 in 1975 and the Individuals with Disabilities Education Act [IDEA] of 1990) tended to focus on procedural compliance and not on achievement outcomes, little attention was actually paid to the ESEA 1994 requirement of full participation.

The situation changed in 1997. The reauthorization of IDEA (US Department of Education, 1997) reiterated the call for full participation of students with disabilities in statewide and district-wide assessment programs, through the use of reasonable adaptations and accommodations. The

N. Zigmond (✉)
University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: naomi@pitt.edu

IDEA amendments asserted that the historical underachievement of students with disabilities was linked to low expectations for learning and scant access to the general education curriculum (Koenig & Bachman, 2004). Mandating that students with disabilities participate in high-stakes accountability assessments would promote quality assurance in special education (Defur, 2002). The assumptions were that participation raises the stakes, in turn yielding higher expectations, leading to increased participation in general education, which promotes better teaching and results in improved academic outcome for students with disabilities (see Fig. 4.1).

This time, the special education community took notice and the general education community did as well when the assessment and accountability regulations in NCLB (2001) emerged as a logical extension of these IDEA 97 provisions. NCLB explicitly prohibited schools from excluding students with disabilities from the accountability system. NCLB restated the requirement for participation of *all* students in statewide accountability assessments and reporting of the results for students with disabilities with everyone else's and as a disaggregated group. Furthermore, students with disabilities were to be held responsible for the same academic content and performance standards as everyone else.

IDEA 2004, Section 612, Part B echoed the NCLB call for including all students with disabilities in state and district-wide assessment programs. It called for tests to adhere to universal design principles to the extent feasible and to contain bias-free test items; simple, clear instructions and procedures; maximum readability and comprehensibility; and optimal legibility from the

start. In that way, tests would be accessible for as many students as possible and the vast majority of students, including the vast majority of those with disabilities, would participate in the regular assessment or in the regular assessment with accommodations approved and widely disseminated by the State. The general idea was that the State had a responsibility to create a testing environment that ensured that as many students as possible could take the general assessment in a way that produced valid and meaningful results. This orientation was consistent with the growing commitment in law and public policy to full inclusion of students with disabilities in general education classrooms and access of *all* students, even those with severe disabilities, to the general education curriculum. The orientation also made sense. By excluding no one from the accountability requirement, it would be possible to answer truthfully the fundamental accountability question: How well is each school doing in bringing *all* of its students up to standard? The answer to this question lay in the annual reporting of the percent of students scoring proficient or advanced on the annual accountability test.

Despite the explicit commitment to all students being assessed on the regular test, from the start, the federal government introduced some flexibility. Beginning in December 2003 (Federal Register, 2003) States were given the option to develop alternate achievement standards to measure the progress of students with the most significant cognitive disabilities. A cap (1.0% of the number of students enrolled in tested grades) was set on the number of proficient and advanced scores based on alternate academic achievement standards that could be included in accountability reports. This cap not only protected the lowest

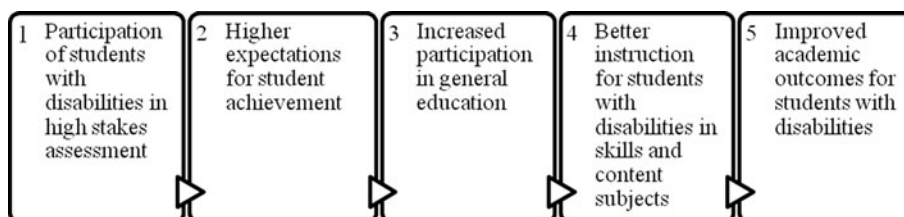


Fig. 4.1 Assumptions underlying more inclusive accountability assessments

performing students with disabilities but also provided a safeguard against inappropriately restricting the scope of educational opportunities for other lower-performing students. Permission to develop and use alternate achievement standards fundamentally changed the framework for assessing students with disabilities. It was no longer sufficient to ask whether a student would be assessed with the general assessment or the general assessment with a valid (or invalid) accommodation. Instead, each State needed to provide guidance to Individualized Education Program (IEP) teams to determine whether a student had a most significant cognitive disability and could be assessed using alternate achievement standards. The use of alternate achievement standards ensured that students with the most significant cognitive disabilities were appropriately included in State accountability systems and that schools and LEAs received credit for these students' achievement.

Then, in 2005, additional flexibility for inclusion of students with disabilities in statewide assessments was introduced. In recognition of "a small group of students whose disability precluded them from achieving grade-level proficiency and whose progress was such that they did not reach grade-level proficiency in the same time frame as other students" (US Department of Education, 2007, p. 8), States were given the option to develop another alternate assessment, with proficiency defined, this time, on the basis of *modified* achievement standards. Before this new flexibility, students with disabilities could take either a grade-level assessment or an alternate assessment based on alternate academic achievement standards. Neither of these options was believed to provide an accurate assessment of what at least some of these students know and can do. The grade-level assessment was too difficult and did not provide data about a student's abilities or information that would be helpful in guiding instruction. The alternate assessment based on alternate academic achievement standards was too easy and not intended to assess a student's progress toward grade-level achievement.

The new regulations permitted States to develop modified grade-level achievement

standards, to adopt such standards, and to develop an assessment aligned with those standards that was appropriately challenging for this particular group of students with disabilities. In the new assessment, expectations of grade-level content mastery would be modified, rather than the grade-level content standards themselves. The assessment had to cover the same grade-level content as the general assessment. The requirement that modified academic achievement standards be aligned with grade-level content standards was critical; for as many of these students as possible to have an opportunity to achieve at grade level, they must have access to and instruction in grade-level content. The regulations included a number of safeguards to ensure that students assessed based on modified academic achievement standards had access to grade-level content and were working toward grade-level achievement. These regulations also allowed teachers and schools to receive credit for the work that they did to help students with disabilities progress toward grade-level achievement.

Now there were several participation options from which to choose for students with disabilities who, even with accommodations, must be assessed with an alternate assessment: an alternate assessment based on grade-level academic achievement standards, an alternate assessment based on modified academic achievement standards, or an alternate assessment based on alternate academic achievement standards. Each of these achievement standards was an explicit definition of how students were expected to demonstrate attainment of the knowledge and skills reflected in the grade-level content standards. Grade-level achievement standards describe what *most* students at each achievement level know and can do. Modified achievement standards are expectations of performance that are challenging for the students for whom they are designed, but are less difficult than grade-level academic achievement standards. Alternate achievement standards reflect the "highest achievement standards possible" (US Department of Education, 2005, p. 18) for students with the most significant cognitive disabilities, no less challenging for the

students than grade-level standards are for students without disabilities. Because the choice of participation methods has to fit individual needs, both IDEA and NCLB charged the IEP team with the task of assigning a student with disabilities to the appropriate assessment method.

Historic Role of the IEP Team

Throughout the history of special education, attention to the unique needs of the individual has been paramount. PL 94-142 codified into federal law the rights of school-age individuals with disabilities (and in subsequent reauthorizations of IDEA of infants, toddlers, and preschoolers) to specially designed instruction, special education and related services, and special consideration in the curricula to be studied and assessments to be taken. To legitimize and make transparent the educational decisions for students with disabilities, legislators wisely introduced the concept of an IEP into the landmark special education law of 1975. The IEP is a written document detailing a student with disabilities' educational needs and the steps the school plans to take to meet those needs. Developed by the special education teacher in conjunction with other school personnel and the child's parents, it documents annually the student's level of performance, his/her goals and objectives, the degree to which those can be met in regular education, and any related services the student might need. It outlines the special curriculum and specially designed instruction that will be implemented to meet this child's individual educational needs. An IEP team, whose membership is individualized for every student with a disability, agrees to what is on the IEP. The team consists, at minimum, of a representative of the educational agency, often the principal; the student's special education teacher; the student's general education teacher; the student's parent or guardian; related service personnel; other persons at the discretion of parents or the educational agency; and for students nearing a transition, a representative of the agency that is likely to provide or pay for the transition services. The IEP that the team develops commits the education provider to focusing on the development of both

academic and practical life skills that are geared toward the goals and aspirations of that individual student. It ensures a meaningful educational curriculum aligned with the individual's needs and plans. "It is this aspect of individualization that distinguishes services authorized and provided through special education from those typically provided in general education" (Kohler & Field, 2003, p. 180).

Parents have always been designated as members of the IEP Team. Their role in planning their child's educational experience was reiterated, and perhaps strengthened, in the language of the most recently authorized version of the Individuals with Disabilities Education Act (IDEA 2004). Parents help to establish, and by signing off on the IEP, agree to the goals the team sets for a child during the school year, as well as any special supports needed to help achieve those goals.

IEP Team Decisions Regarding Participation in Statewide Assessment

The IEP Team is generally given wide berth in determining what a student with a disability will learn, where that learning will take place, and how the outcomes of that learning will be evaluated. As specified through IDEA, IEP teams have ultimate responsibility for making instructional, curricular, and assessment decisions for each student with a disability. However, since passage of the 1997 amendments to the *Individuals with Disabilities Education Act (IDEA, 1997)*, the breadth of IEP Team decision-making has been somewhat curtailed. For all but a small percentage of students with the most severe cognitive disabilities (and even for those students whenever possible), the agreed-upon IEP goals must provide access to the general education curriculum and address state-approved grade-level content standards. Furthermore, *all* students with disabilities must be included in state and district assessments. The IEP team cannot decide that a particular student will not participate in a statewide assessment, even if the child's parents insist. If it is the judgment of the team that the student is unable to participate in the regular assessment even with accommodations,

they must assign the student to participate in an alternate assessment. Some states may specify certain conditions under which parents may refuse to permit their children's participation; these exemptions apply to all students in a given state. For example, in Pennsylvania, students may be excused from the assessment if (and only if) their parents have reviewed the test's content and have declared it to be inappropriate based on religious grounds. Beyond that, it is inconsistent with federal law for an IEP team to exempt a student from state assessments.

In making the decision about how students with disabilities will take the statewide assessment, the IEP team operates in an environment in which academic content, academic achievement standards, and assessments are set by the State; the technical qualities of the State assessments are well established; there are policies in place on the use of accommodations that do not invalidate test results; and there are State guidelines regarding eligibility for alternate assessments.¹ Furthermore, each of these is subject to federal scrutiny and must meet the requirement specified in federal regulations.

Federal directives regarding participation in alternate assessments. Two guidance documents issued by the US Department of Education, *Alternate Assessment Standards for Students with the most Significant Cognitive Disabilities (2005)* and *Modified Academic Achievement Standards (2007)*, have sought to clarify both the nature of the students for whom the alternate assessment based on alternate academic achievement standards and the alternate assessment based on modified academic achievement standards were intended and the role of the IEP team in deciding who should take which state accountability assessment. The two guidance documents emphasize that all students with disabilities should have access to the general curriculum; that the participation decisions must be made by each student's IEP team and on an individual student

basis, and that decisions should not be based on disability category or other general qualities. Guidance documents assign to each State the responsibility to establish clear and appropriate guidelines for IEP teams to use when deciding if one or another alternate assessment is justified for an individual child. They suggest that each State's guidelines may contain the following:

- Criteria that each student must meet before participating in alternate assessments;
- Examples or case study descriptions of students who might be eligible to participate in such alternate assessments;
- Accommodations that are available for all assessments, and any special instructions that IEP teams need to know if such accommodations require special permission or materials;
- Flow charts for determining which assessment is appropriate and/or which accommodations are appropriate;
- Timelines for making the participation decisions;
- Consequences that affect a student as a result of taking an alternate assessment (e.g., eligibility for a regular high school diploma);
- Consequences that affect a test score as a result of using a particular accommodation;
- Approaches for ensuring that all students have access to the general curriculum;
- Commonly used definitions; and
- Information about how results are reported for individual student reports and in school or district report cards.

The 2005 Guidance specifies that "only students with the most significant cognitive disabilities may be assessed based on alternate achievement standards" (p. 23) and explains that these are the students with IEPs whose cognitive impairments may prevent them from attaining grade-level achievement standards, even with the very best instruction. The 2007 Guidance stipulates two primary requirements for participation in an alternate assessment based on modified academic achievement standards. The student must have a disability that precludes attainment of grade-level achievement standards. And, the student's progress to date in response to appropriate instruction, special education, and related services must be

¹ States are not required to use every assessment method. Some states may have alternate assessments but choose not to use alternate or modified achievement standards. State guidelines should clarify which methods are available.

such that, even if significant growth occurs, there is reasonable certainty that the student will not achieve grade-level proficiency within the year covered by the student's IEP. Both documents emphasize the importance of communicating the assessment participation decision to students' parents and to inform them of any educational consequences of the decision.

State Guidelines for IEP Team Decision Making, 2007–2009

In 2007, 2008, and 2010, NCEO published a synthesis of decision-making guidelines posted by states on their State Department of Education websites (Lazarus, Rogers, Cormier, & Thurlow, 2008; Lazarus, Thurlow, Christensen, & Cormier, 2007; Lazarus, Hodgson, & Thurlow, 2010). The 2007 report was a quick snapshot of the decision-making guidelines regarding participation in the alternate assessment based on modified academic achievement standards of six states (Kansas, Louisiana, North Carolina, North Dakota, Oklahoma, and Maryland) just a few months after the 2007 regulations on the AA-MAS were finalized. One year later, the report compared and contrasted participation guidelines for nine states (the original six, plus California, Connecticut, and Texas), although none had yet successfully completed the peer review process that determines whether an accountability assessment fulfills the necessary requirements for the state to receive federal funds. By 2009, one state (Texas) had passed peer review though the number of states with publically available guidelines for student participation in alternate assessments was up to 14 (Arizona, Indiana, Michigan, Ohio, and Tennessee added to the nine of the previous year). In that third report, Lazarus, Hodgson, and Thurlow (2010) describe common elements within the 14 sets of guidelines and provide samples of the checklists (usually a series of yes/no questions) or flow charts and decision trees (conceptual representations of the decision-making process) recommended for use by IEP teams. Not surprisingly, the published guidelines

each contain some or all of the elements specified in the federal guidance documents (see Table 4.1). Across the 14 states, elements of the guidelines to IEP teams fall into four categories: (a) indicators that qualify the student for the alternate assessment based on alternate achievement standards, instead; (b) factors that should not influence the participation decision; (c) indicators of opportunity to learn grade-level content of the general education curriculum; and (d) indicators of limited academic growth or progress.

Recommendations to IEP Teams

The decision about how a student with a disability will participate in the annual statewide accountability assessments is not made in isolation. It is part of the total plan that defines an appropriate education for the student. As such, it should be made at the annual IEP meeting, with all the important actors present, and in conjunction with the development of IEP goals. If the student is making adequate progress in the general education curriculum, the recommendation regarding participation is indisputable; the default decision is the general assessment, with or without accommodations, or for students who cannot manage a pencil/paper test, an alternate assessment based on grade-level achievement standards (see Fig. 4.2). The discussion at the IEP team meeting would focus on the nature of the accommodations routinely used to enhance this student's typical educational experience. These accommodations would be recommended to make the assessment more accessible and to have the assessment results better reflect what the student knows and is able to do. This seems an obvious point. However, in a 2001 study of the relationship between the types of accommodations typically provided to students with IEPs during instruction and those offered on the accountability tests, Ysseldyke and colleagues (2001) reported a significantly high tendency for students with IEPs to receive testing accommodations that were not provided during instruction.

The second decision is also fairly straightforward. If the student has a significant cognitive

Table 4.1 Common elements in 14 state assessment participation guidelines

Number of State guidelines <i>n</i> = 14	Element to be considered in making participation decision
14 states	Student has a current IEP
12 states	Permit “combination participation”: separate decisions made by subject area [3 states allow selection from all options; 9 states require selection from regular assessment and AA-MAS, only]
9 states	Decision requires parental notifications
8 states	Decision requires consideration of consequences for meeting graduation requirements
<i>Indicators that qualify the student for the alternate assessment based on alternate academic achievement standards</i>	
4 states	Decision based on presence of significant cognitive disability
6 states	Decision based on whether receiving instruction on extended or alternate standards
7 states	Decision based on whether student receives specialized or individualized instruction
<i>Factors that should not influence the participation decision</i>	
8 states	Decision not based on category label
7 states	Decision not based on excessive absences, social, cultural, language, economic, or environmental factors
6 states	Decision not based on placement setting
<i>Indicators of opportunity to learn grade-level content of the general education curriculum</i>	
11 states	Decision based on whether student is learning grade-level content
6 states	Decision based on whether student receives accommodations during classroom instruction
9 states	Decision based on whether student has IEP goals based on grade-level content standards
9 states	Decision based on whether student has IEP goals based on grade-level content standards
<i>Indicators of limited academic growth or progress</i>	
12 states	Decision based on previous performance on multiple measures
11 states	Decision based on evidence that student not progressing at rate expected to achieve grade-level proficiency
9 states	Decision based on whether student has IEP goals based on grade-level content standards
3 states	Decision based on evidence that student’s performance is multiple years behind grade-level expectations
4 states	Decision based on previous performance on state assessment

disability, has IEP goals and is receiving instruction based on extended or alternate academic content standards, and requires significant scaffolding to participate meaningfully in the general education curriculum, the team should recommend that the student participate in annual statewide accountability assessments through the

alternate assessment based on alternate academic achievement standards (AA-AAS).

Of course, if the student is not making adequate progress in the general education curriculum, but is not eligible to be assigned the AA-AAS, the IEP team may be obligated to consider recommending a different alternate assessment

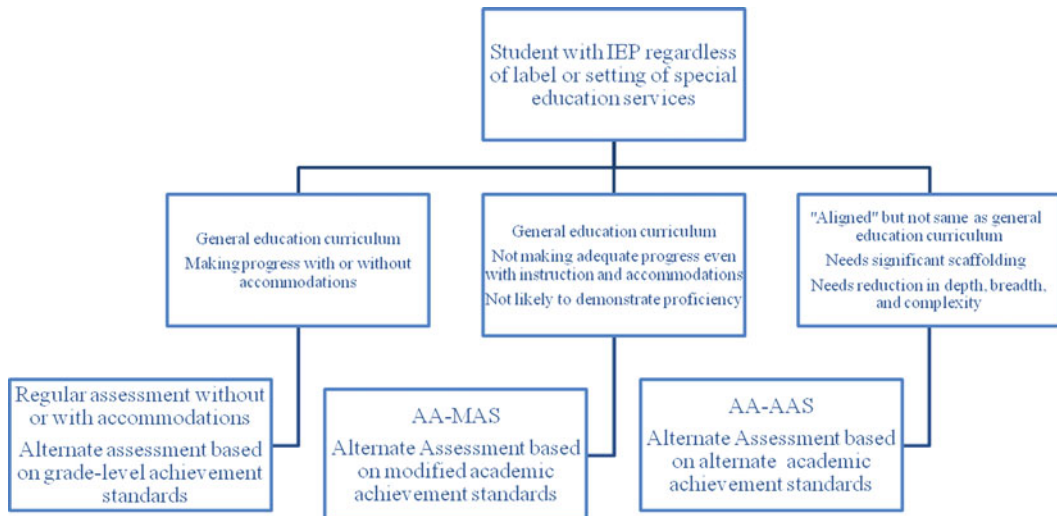


Fig. 4.2 Primary decisions regarding participation in annual accountability assessment

by reviewing and following guidelines provided by their State Department of Education, using the flow charts, decision trees, or checklists designed uniquely for use in their state. This decision is not nearly as clear-cut. Within the parameters of the state guidelines, IEP teams would be wise to consider the following:

Make certain all members of the IEP team are clear about what the participation decision for these 'middle' students is supposed to accomplish. State guidelines are (have to be) consistent with federal guidelines and regulations, but an IEP team is making a singular decision in a very specific context of district, school, and individual student. Is the student actually competent in the general education content and skills but needs an even more accessible test than is possible with standard accommodations? Is this student so far from competent that he/she will not be proficient regardless of the definition of proficiency on the alternate assessment, so the student might as well participate in an 'easier' assessment? Is there pressure (spoken or unspoken) within the school or district to increase the number of students with disabilities who score proficient on the annual assessment by finding ways to assign them en masse to an alternate assessment? Is there a strong commitment within the school or district to maintain uncompromisingly high standards for all the students, teachers, and curriculum, by

assigning as few students as possible to alternate assessments? Answers to these questions clarify the climate in which the participation decision needs to be made.

Do not prejudice the outcome of the decision based on the student's diagnostic label or the setting in which special education services are provided. The type of assessment recommended for a student should not be based on the student's disability classification. Most students with disabilities do not have intellectual impairments. They have learning disabilities, speech/language impairments, other health impairments, emotional/ behavioral disabilities, and/or physical, visual, and hearing impairments. When given appropriate accommodations, services, supports, and specialized instruction, these students may be capable of learning the grade-level content in the general education curriculum and thus achieve proficiency on the grade-level content standards. In addition, research suggests even a small percent of students with disabilities who have intellectual impairments (i.e., generally includes students in categories of mental retardation, developmental delay, and some with multiple disabilities and/or autism) might also achieve proficiency when they receive high-quality instruction in the grade-level content, appropriate services and supports, and appropriate accommodations (Thurlow, 2007). There is no basis in research

for defining how the specific categorical labels differentiate how students learn and demonstrate their learning, or how, as required in the regulation, this disability prevents a particular student from attainment of grade-level achievement within the current year (National Association of School Psychologists, 2002). Instead of focusing on label, the IEP team's decision about assessment should be based on the types and intensity of support the student needs to show academic learning during ongoing instruction.

The decision about the most appropriate type of assessment for students with disabilities should be based on neither current placement nor the setting in which the student receives instruction. As noted previously, every student has the right to access the general curriculum. Students also have the right to receive instruction from highly qualified teachers who are trained in the content area and the right to an appropriate education in the least restrictive environment (LRE; IDEA, 2004). LRE does not define the way a student participates in a statewide assessment. Every type of educational setting includes students with disabilities who can be recommended for any of the assessment methods.

Do not let the team decision be unduly influenced by the quality and alignment with grade-level standards of previous years' IEPs or the limited educational accomplishment of the student that resulted from those previous educational experiences. The regulations (CFR §200.1(e)(2)(ii)(A)) stipulate that a student should not be assigned to an alternate assessment based on modified academic achievement standards if the child's IEP is not of "high quality" and designed to move the child "closer to grade-level achievement" (Federal Register, 2007, p. 17749). Does that mean a struggling student with a poorly written IEP will be precluded from taking the new alternate assessment for at least a year during which time the IEP team writes a better IEP? Further, Section 200.1(e)(2)(ii)(A) stipulates that students should not be assigned to an alternate assessment based on modified academic achievement standards if they have not had the opportunity to learn grade-level content. Does that mean that a struggling

student being taught reading or mathematics at his/her instructional level rather than at "grade-level" would be precluded from taking the new alternate assessment for at least a year during which time he is provided less appropriate and 'special' grade-level instruction instead? Section 200.1(e)(2)(ii)(B) requires IEP teams to examine a student's progress in response to high-quality instruction over time and using multiple measures before assigning that student to an alternate assessment based on modified academic achievement standards. Does that mean that a struggling student whose teacher delivers "less than high-quality instruction" would be precluded from assignment to the assessment based on modified achievement standards until she or he gets a better teacher? And how should measuring student progress be standardized? Given the pressures of accountability and the time and effort devoted to preparing for and administering state tests, many schools have stopped administering additional standardized achievement tests. And, although Response to Intervention (RTI) and Reading First have promoted increased use of curriculum-based measurement and progress monitoring in reading and math, these models are often restricted to the elementary level. In fact, the field is only beginning to scratch the surface in developing and implementing comprehensive progress monitoring/intervention frameworks for middle and secondary students – the primary cohort for accountability testing (Mastropieri & Scruggs, 2005; Vaughn et al., 2008). How will IEP teams determine how much progress is sufficient? And for how long should an IEP team wait for signs of progress before recommending the student for the new alternate? There are currently no definitive answers to these questions.

The recommendation regarding which assessment will be used for a particular student's participation in statewide accountability should not be made in a separate meeting; this decision-making is an integral part of the entire IEP team planning and decision-making for the student for the upcoming year. The same team that is making the participation decision is writing the IEP goals and objectives for the next academic

year, making decisions about where and how the student will be taught. The fact that many students with disabilities currently do not achieve at proficiency should raise questions of whether they were receiving the required high-quality instruction in the grade-level content, appropriate services and supports, and appropriate accommodations, not confirm preformed opinions about the capacity of these students to learn or perform. Anne Donnellan (1984, p. 142) concluded using her “criterion of least dangerous assumption” that barring proof to the contrary, educators need to assume that poor performance is due to instructional deficits instead of student deficits. The IEP team should correct deficiencies in past IEPs while making assessment decisions for the future.

Craft the IEP so that students have opportunity to learn what will be tested (i.e., this section of the IEP has ramifications for other sections of IEP). An important principle of assessment is that students have the opportunity to learn the material on which they will be tested. English and Steffy (2001) call this the “doctrine of no surprises” (p. 88). Because the student being discussed is to be assessed based on grade-level content standards, instruction for the students with disabilities should be aligned with grade-level content in reading and math (albeit reduced in breadth, depth, and/or complexity for some students). The IEP team must ensure that the student’s IEP includes goals that address the content standards for the grade in which the student is enrolled. Since the decisions about how a student will participate in State and district-wide assessments are to be made at the student’s annual IEP meeting, the IEP team will have the time to also develop IEP goals that are based on grade-level content standards. This will help to ensure that the student has had an opportunity to learn grade-level content prior to taking an alternate assessment based on modified academic achievement standards.

Remember that an IEP team cannot “stop the clock” (i.e., suspend participation in the annual accountability assessments until the student has appropriately written IEP goals and access to appropriate instruction and accommodations). Participation in statewide assessment is an annual event from third grade through eighth

grade (and continues in at least one grade of high school). It is true that the reality of many students with disabilities currently not achieving at proficiency raises questions of whether they have been receiving the required high-quality instruction in the grade-level content, appropriate services and supports, and appropriate accommodations. Nevertheless, participation in statewide accountability assessment cannot be postponed until past deficiencies or inappropriate educational decisions/opportunities have been corrected.

Do not be intimidated by the ‘cap’ on how many students’ scores from alternate assessments can be counted as proficient. Selecting a method for participation in statewide assessments, like all educational decisions made by the IEP team, should focus on creating an appropriate education for each individual student with a disability. Although limits are in place on the number of students who can be reported as proficient using modified or alternate achievement standards, there is no limit on the number of students who can be assigned to these alternate assessments. In general, the Department of Education estimates that about 9% of students with disabilities (approximately 1% of all students) have significant cognitive disabilities that qualify them to participate in an assessment based on alternate achievement standards. An additional 18% of students with disabilities are eligible to take an alternate assessment based on modified academic achievement standards. They even suggest “an LEA that assesses significantly more than 3% of all students with an alternate assessment based on modified academic achievement standards should prompt a review by the State of the implementation of its guidelines to ensure that the LEA was not inappropriately assigning students to take that assessment” (US Department of Education, 2007, p. 37).

These percentage estimates are misleading for two reasons. First, since they actually represent the number of students that can be *counted* toward proficient, they presume that every student assigned to an alternate assessment will score in the proficient range. If a state, in designing the alternate state assessments and defining alternate

or modified achievement standards, is sincerely setting challenging standards, that would hardly be the case. Second, every IEP team should have the flexibility to select the assessment method that is best for a particular student. The IEP team is individualized, the IEP goals are individualized, and the needs of the individual student under consideration are unique. In the spirit and letter of IDEA, an IEP team must make the assessment recommendation free of external considerations like a percentage participation cap.

IEP team members must understand that there will be intended and unintended consequences to the assessment participation decision. There is considerable research evidence that, when the stakes are high, assessment drives instruction (rather than the more desired scenario of instruction driving assessment). Without intending to teach to the test, in classrooms today, “what you test is what you get” (Mislevy, 2008, p. 5). Tests determine not only the content but also the format of instruction. Shepard (1989) reported that in response to externally mandated tests, “Teachers taught the precise content of the tests rather than underlying concepts; and skills were taught in the same format as the test rather than as they would be used in the real world. For example, teachers reported giving up essay tests because they are inefficient in preparing students for multiple-choice tests” (p. 5). Others have documented that in schools or classrooms using multiple-choice tests, instruction tends to emphasize drill and practice on decontextualized skills, reflecting the emphasis of many multiple-choice tests (see Romberg, Zarinna, & Williams, 1989).

Consequences will, and should, follow the assessment participation decision. Both the alternate assessment based on alternate academic achievement standards and the alternate assessment based on modified academic achievement standards assume that the student with disabilities has been learning the general education curriculum, but each defines very different expectations for mastery of that curriculum in terms of breadth, depth, and complexity. The reason given for developing these assessment options was the recognition that some student with disabilities should not be expected to master the full scope

of the grade-level general education curriculum in the 1-year time frame of annual assessments. The type of student work that defines proficient performance on an alternate assessment based on alternate academic achievement standards is substantially different from the type of work that defines proficient performance on an alternate assessment based on modified academic achievement standards. That, in turn, is substantially different from the performance criteria associated with proficiency on the regular assessment. If definitions of proficient performance are different, it follows that the scope of instruction will be different, especially if accountability assessments are designed to assess mastery of what has been taught.

Narrowing the curriculum may be a positive consequence. If over the course of the academic year, fewer concepts need to be mastered, more time can be spent on working to mastery. On the other hand, a narrowed curriculum at an early grade level may limit educational opportunities down the road. Assignment to a particular alternate assessment may set the trajectory of academic accomplishment too soon, before the student has had the opportunity to be exposed to, and possibly achieve proficiency in, a broader and more complex curriculum.

In the same way, changing expectations for proficiency (i.e., using alternate or modified achievement standards to define proficient performance) may be useful for some students and their families. It makes “proficiency” more attainable to students with chronic difficulties in school performance, giving the student a sense of pride in accomplishment, even if the “proficiency” designation actually carries a different meaning. On the other hand, telling a family or a student that his/her academic performance in reading, or mathematics or science is “proficient” could be harmful when the performance is only judged as proficient because the definition of proficiency has been altered. It gives the family and the student a misleading accounting of relative accomplishment, an unwarranted sense of what may be achievable in the future. With parents present at the IEP meeting, the implications of the assessment decision can be discussed openly and

considered in the final assessment participation recommendation.

Concluding Comments

NCLB (2001) set the ambitious but unachievable goal of having all students scoring proficient on accountability assessments by the year 2014. The reason for introducing the flexibility of alternate assessments into the accountability system was to help schools, districts, and states “count more students as proficient” (Quenemoen, 2010, p. 22). The regulatory language leaves it to the states to define the target populations for these alternate assessments.

In the current standards-based accountability-driven reform model, any option that encourages or rewards less-challenging standards for any student who could achieve at grade level (assuming they have access to the curriculum and are instructed effectively) undermines the entire system of school reform. For most students, with and without disabilities, we cannot yet predict with any accuracy whether they will achieve at grade level when instructed effectively. Thus, the least dangerous assumption requires that all students receive that effective instruction. The implementation of alternate assessments should expressly *raise* expectations and result in realistic but higher achievement for students who participate in the options. Ultimately, state-defined modified or alternate achievement standards should be policy statements of what are appropriately high expectations for some small, defined groups of students, and they should lead to improvement of achievement outcomes for these students, in order to be consistent with the letter and the spirit of NCLB and IDEA.

Careful monitoring of consequences of the assessment decisions is essential to ensuring that the intended positive consequences are occurring and unintended negative consequences are not. But IEP teams in annual IEP meetings are not making grand assessments of the efficacy of the standards-based

accountability system. They are making recommendations and decisions to ensure an appropriate education for a particular student in a particular district, in a particular state. Each member of the team, the LEA representative, the educators, and the student’s parents, should be fully informed about the nature of the decisions that need to be made and the implications of those decisions in shaping the student’s educational experiences. There are federal, state, and perhaps even district guidelines that should influence the decision, but in the end, the decision needs to be a very personalized one, reflecting only the team consensus of what is appropriate for this particular student at this particular time in his/her educational career, until the next year when the decision gets made all over again.

References

- Crawford, L., Almond, P., Tindal, G., & Hollenbeck, K. (2002). Teacher perspectives on inclusion of students with disabilities in high stakes assessments. *Special Services in the Schools, 18*(1/2), 95–118.
- Defur, S. H. (2002). Education reform, high-stakes assessment, and students with disabilities: One states approach. *Remedial and Special Education, 23*(4), 203–211.
- Donnellan, A. (1984). The criterion of the least dangerous assumption. *Behavior Disorders, 9*, 141–150.
- English, F. W., & Steffy, B. E. (2001). *Deep curriculum alignment: Creating a level playing field for all children on high-stakes tests of educational accountability*. Lanham, MD: Scarecrow Press.
- Federal Register (2003, December 9). Department of Education, 34 CFR Part 200, Title I – Improving the academic achievement of the disadvantaged; final rule, pp. 68968–68708
- Federal Register (2007, April 9). 34 CFR Parts 200 and 300, Title I – Improving the academic achievement of the disadvantaged; rules and regulations, pp. 17747–17781
- Improving America’s Schools Act, Pub. L. No. 103–382. (1994). Retrieved September 10, 2008, from <http://www.ed.gov/legislation/ESEA/toc.html>
- Individuals with Disabilities Education Act, Pub. L. No 101-476. (1990). Retrieved September 10, 2008, from <http://www.ed.gov/policy/speced/leg/edpicks.jhtm>
- Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 105-7. (1997). Retrieved

- September 10, 2008, from http://www.cec.sped.org/law_res/doc/law/index.php
- Individuals with Disabilities Education Improvement Act, Pub. L. No. 108-446. (2004). Retrieved September 10, 2008, from <http://www.ed.gov/policy/speced/guid/idea/idea2004.htm>
- Koenig, J. A., & Bachman, L. F. (Eds.). (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments*. Washington, DC: National Academies Press.
- Kohler, P. D., & Field, S. (2003) Transition-focused education: Foundation for the future. *Journal of Special Education, 37*, 174–183.
- Lazarus, S. S., Hodgson, J., & Thurlow, M. L. (2010) *States' participant guidelines for alternate assessments based on modified academic achievement standards (AA-MAS) in 2009* (Synthesis Report 75). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Lazarus, S. S., Rogers, C., Cormier, D., & Thurlow, M. L. (2008). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2008* (Synthesis Report 71). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Lazarus, S. S., Thurlow, M. L., Christensen, L. L., & Cormier, D. (2007). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2007* (Synthesis Report 67). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Mastropieri, M. A., & Scruggs, T. E. (2005) Feasibility and consequences of response to intervention: Examination of the issues and scientific evidence as a model for the identification of individuals with learning disabilities. *Journal of Learning Disabilities, 38*(6), 525–531.
- Mislevy, R. L. (2007). What you test is what you get: Q&A with Robert J. Mislevy, Ph.D. *Endeavors, 9*(16), 4–5.
- National Association of School Psychologists (NASP). (2002). *Position statement: Rights without labels*. Retrieved April 19 2010, from http://www.nasponline.org/about_nasp/pospaper_rwl.aspx
- No Child Left Behind Act, Pub. L. No. 107-110 (2001). Retrieved September 11, 2008, from <http://www.ed.gov/policy/elsec/leg/eseas02/index/html>
- Public Law 94-142 (1975) *Education of All Handicapped Children Act of 1975*. Retrieved 1/09, from <http://users.rcn.com/peregrin.enteract/add/94-142.txt>
- Quenemoen, R. (2009, July). Identifying students and considering why and whether to assess them with an alternate assessment based on modified achievement standards. In M. Perie (Ed.), *Considerations for the alternate assessment based on modified achievement standards (AA-MAS)* (Chapter 2, pp. 17–50). Retrieved from the New York Comprehensive Center: http://nycomprehensivecenter.org/docs/AA_MAS_part2.pdf
- Romberg, T., Zarinnia, A., & Williams, S. (1989). The influence of mandated testing on mathematics instruction: Grade eight teachers' perceptions. In T. Romberg & L. Wilson (1992, September). Alignment of tests with the standards. *Arithmetic Teacher, 40*(1), 18–22.
- Shepard, L. A. (1989, April). Why we need better assessments. *Educational Leadership, 46*(7), 4–9.
- Thurlow, M. (2007). The challenge of special populations to accountability for all. In D. Clark (Ed.), *No child left behind: A five year review* (Congressional Program, Vol. 22, No. 1, pp. 39–44). Washington, DC: The Aspen Institute.
- Thurlow, M. L., Seyfarth, A., Scott, D. L., & Ysseldyke, J. E. (1997). *State Assessment policies on participation and accommodations for students with disabilities: 1997 update* (Synthesis Report No. 29). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from <http://education.umn.edu/NCEO/OnlinePubs/Synthesis29.html>
- US Department of Education. (1997). Individuals with disabilities education act, from <http://www.ed.gov/offices/OSEP/policy/IDEA/index.html>
- US Department of Education. (2005). *Alternate assessment standards for students with the most significant cognitive disabilities: Non-regulatory guidance*. Washington, DC: Author. Retrieved September 10, 2007, from <http://www.ed.gov/policy/speced/guid/nclb/twopercent.doc>
- US Department of Education. (2007). *Modified academic achievement standards: Non-regulatory guidance*. Washington, DC: Author. Retrieved September 10, 2007, from <http://www.ed.gov/policy/speced/guid/nclb/twopercent.doc>
- Vaughn, S., Fletcher, J. M., Francis, D. J., Denton, C. A., Wanzek, J., Wexler, J., et al. (2008). Response to intervention with older students with reading difficulties. *Learning & Individual Differences, 18*(3), 338–345.
- Ysseldyke, J., & Thurlow, M. (1994). *Guidelines for inclusion of students with disabilities in large-scale assessments* (Policy Directions No. 1). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Ysseldyke, J., Thurlow, M., Bielinski, J., House, A., Moody, M., & Haigh, J. (2001). The relationship between instructional and assessment accommodations in an inclusive state accountability system. *Journal of Learning Disabilities, 34*(3), 212.

Australian Policies to Support Inclusive Assessments

5

Michael Davies and Ian Dempsey

Introduction

Australia has recently adopted a National Assessment Program for Literacy and Numeracy (NAPLAN) that involves assessment of students in Years 3, 5, 7, and 9. While Australian legislation and policies are designed to support inclusive assessments for all, it is evident that in practice this is not the case. This chapter provides information about Australian legislation and ensuing policies, how they interact with national achievement testing in Australia, and proposes recommendations for the development of a more effective and inclusive assessment regime for Australian students.

The Australian Education Landscape

Australia is a federal parliamentary democracy comprising six states and two territories and had the eleventh-largest gross domestic product per capita in the world in 2008 (World Bank, 2009a). In that year, the population of Australia was 21.4 million people, the vast majority of whom lived in the country's capital cities (World Bank, 2009b).

Australia follows a two-tier system of school education which includes primary education

(generally till 12 years of age), and secondary education (generally till 18 years of age). Schooling is compulsory between the ages of 6 and 15 or 16, depending on the jurisdiction. Each of Australia's states and territories has a department of education that provides free public education. Public school students comprise about two-thirds of the total student population. The remaining students attend fee-paying religious and secular private schools which also receive substantial government funding. The Programme for International Student Assessment for 2006 ranked Australia 6th on a worldwide scale for Reading, 8th for Science, and 13th for Mathematics (Australian Council for Educational Research, 2009).

While there is no national curriculum in Australia, there is agreement across the states and territories on the broad content of curriculum (Ministerial Council for Education, Early Childhood Development and Youth Affairs, 2009a). Although the Federal government makes a relatively small direct contribution to the states' and territories' education budgets, conditions are often tied to this funding. This mechanism of influence has, among other things, facilitated the introduction in recent years of a national testing program and the use of a common student academic grading system.

The majority of states and territories provide three enrolment options for students with a disability: regular classes, support classes (separate classes in a regular school), and special schools. For reasons explored in the next section,

M. Davies (✉)
School of Education and Professional Studies, Faculty of Education, Griffith University, Brisbane, QLD 4122, Australia
e-mail: m.davies@griffith.edu.au

Australia has a very poor record of reporting both the number of students with a disability in Australian schools and the nature of enrolment of those students. The limited evidence available suggests that over 3.5% of Australian school students have a disability (there is some variation in the definition of disability across the states and territories), there has been a large increase in the number of students with a disability identified in regular schools, the majority of students with a disability are educated in mainstream schools, and as a percentage of the total school population there has been a recent increase in the proportion of students placed in segregated settings (Australian Government Productivity Commission, 2004; Dempsey, 2004, 2007). It is likely that the increased number of identified students with a disability in regular schools have always been in the regular school system. However, many have more recently been identified with a disability so as to receive funding support and also because of improved awareness of disability and special needs by education professionals.

Australian Legislation Relevant to Students with a Disability

Disability Discrimination Act

While many Australian states and territories had legislation that directly related to discrimination of a variety of groups in the community, by the late 1980s there was increasing concern that without a law focussing on disability rights little progress could be made in addressing discrimination. Consequently, in 1992 federal legislation was passed in the *Disability Discrimination Act 1992* (DDA) (Australasian Legal Information Institute, 2009) that directly addressed a range of areas, including education. In relation to education, the DDA states the following:

- (1) It is unlawful for an educational authority to discriminate against a person on the ground of the person's disability:
 - (a) by refusing or failing to accept the person's application for admission as a student; or
 - (b) in the terms or conditions on which it is prepared to admit the person as a student.
- (2) It is unlawful for an educational authority to discriminate against a student on the ground of the student's disability:
 - (a) by denying the student access, or limiting the student access, to any benefit provided by the educational authority; or
 - (b) by expelling the student; or
 - (c) by subjecting the student to any other detriment.
- (3) It is unlawful for an education provider to discriminate against a person on the ground of the person's disability:
 - (a) by developing curricula or training courses having a content that will either exclude the person from participation, or subject the person to any other detriment; or
 - (b) by accrediting curricula or training courses having such a content (Australasian Legal Information Institute, 2009).

Although this legislation effectively mandates that educational services are provided to students with a disability, there are some aspects of the original legislation that are less clear about the nature of the educational setting and the support that should be provided to Australian students with a disability. For example, the DDA includes a component that identifies unjustifiable hardship as a legal basis for discrimination. In determining whether an education provider may experience unjustifiable hardship in enrolling a student with a disability, the DDA notes that the following need to be taken into account:

- (a) The nature of the benefit or detriment likely to accrue to, or to be suffered by, any person concerned;
- (b) the effect of the disability of any person concerned;
- (c) the financial circumstances, and the estimated amount of expenditure required to be made, by the first person (i.e., the education provider);
- (d) the availability of financial and other assistance to the first person (Australasian Legal Information Institute, 2009).

By 2000, there had been one high-profile Federal Court case related to the enrolment of a student with a disability in a regular school. In *Hills Grammar School v. HREOC* (Australian Human Rights Commission, 2009a), a private school was deemed to have discriminated against a young student with spina bifida by refusing her enrolment in a regular class. The school did not provide a support class placement option. Enrolment was refused on unjustifiable hardship

grounds with the school arguing that the student required a range of additional supports that were beyond the capacity of the institution. The Federal Court noted that the institution was a large national education provider with substantial resources. The school was fined and ordered to provide a placement for the student, if she wished. The precedent set by this case means that the unjustifiable hardship argument is unlikely to be successfully used by large education providers.

Other aspects of the Australian DDA also warrant examination. The DDA defines disability as including a range of more traditional impairments (e.g., physical, intellectual, psychiatric, and sensory), as well as some impairments that are traditionally not recognised as disabilities in educational settings in Australia (e.g., learning disabilities, behaviour problems, attention deficit hyperactivity disorder, and physical disfigurement). In addition, the DDA recognises disabilities that individuals may presently experience, may have had in the past, or may have in the future (e.g., a disease that is yet to physically manifest itself).

Because the original DDA legislation addressed educational enrolment, and not educational services in schools, there was little incentive for school systems to broaden their definitions of disability beyond traditional, medically oriented categories. Indeed, for many state and territory education departments there was a powerful disincentive to broaden their definition; that it may place excessive financial demands on educational authorities to meet the needs of a perceived additional group of students with special needs. Such concerns were misplaced because Australian public education systems already provided a range of extensive supports to students with a disability, as defined by the DDA (e.g., services to students with reading difficulties and behaviour problems). Regardless, the unpreparedness of the states and territories to adopt a common definition of disability means that in Australia, at a national level, it is not possible to report on the total number of school students with a disability, or to determine the

number of students with a disability in different educational settings.

In its present form, the DDA provides several avenues of redress for individuals, groups, and organisations. In the first instance, complainants are encouraged to lodge a grievance with the Australian Human Rights Commission, the body with responsibility for oversight of the DDA. Such complaints are reviewed by the Commission, which seeks to conciliate a settlement between the parties, usually without admission of liability (Australian Human Rights Commission, 2009b). If conciliation cannot be achieved, the matter may then be heard by the Australian Federal Court. By 2010, there had been 14 court decisions and 81 conciliated outcomes.

In 2002, a Productivity Commission review of the DDA concluded that the legislation had been “reasonably effective in reducing discrimination (p. xxvi)” although its effectiveness was highly variable across sectors of the economy and within disability groups (Australian Government Productivity Commission, 2004). Education was the third most common subject of complaints made under the DDA after employment and the provision of goods and services. The Commission also noted that the introduction of education standards to supplement the DDA was a pressing need and that the net impact of such standards would be positive.

Education Standards

Another important feature of the DDA worthy of discussion is that it provided a mechanism to develop standards to assist organisations to understand their responsibilities in avoiding discrimination. Education standards were eventually legislated in 2005 (Australian Government Attorney-General’s Department, 2005), and they had a ‘chequered’ history. From the time of the release of a discussion document in 1996, it took 9 years and many attempts at negotiation between the Commonwealth and State and Territory governments before the standards were enacted.

However, the standards address enrolment, participation, curriculum, student support services, and elimination of harassment and victimisation (Australian Human Rights Commission, 2009c).

An extract from the education standards, as they relate to participation, appears in [Box 5.1](#).

Box 5.1. Standards for participation

- (1) The education provider must take reasonable steps to ensure that the student is able to participate in the learning experiences (including assessment) of the courses or programs provided by the educational institution, and use the facilities and services provided by it, on the same basis as a student without a disability, (and include measures that ensure that . . . the assessment and certification requirements of the course or program are appropriate to the needs of the student and accessible to him or her; and (f) the assessment procedures and methodologies for the course or program are adapted to enable the student to demonstrate the knowledge, skills, or competencies being assessed)
- (2) The provider must:
 - (a) consult the student, or an associate of the student, about whether the disability affects the student's ability to participate in the courses or programs for which the student is enrolled and use the facilities or services provided by the provider; and
 - (b) in the light of the consultation, decide whether an adjustment is necessary to ensure that the student is able to participate in the courses or programs provided by the educational institution, and use the facilities and services provided by it, on the same basis as a student without a disability; and
 - (c) if:
 - (i) an adjustment is necessary to achieve the aim mentioned in paragraph (b); and
 - (ii) a reasonable adjustment can be identified in relation to that aim; make a reasonable adjustment for the student in accordance with Part 3.
- (3) The provider must repeat the process set out in subsection (2) as necessary to allow for the changing needs of the student over time.

An important feature of the participation standards is the notion that students with a disability are entitled to participate on the same basis as students without a disability, and without

discrimination. The ramifications of this requirement are that school courses and activities must be flexible enough to meet the student's needs, that the school must consult with the student and/or the student's advocate in determining those needs, and that the participation of the student with a disability should be comparable to the participation of a student without a disability. Students with a disability are to be afforded the same opportunities to participate in school or a course as other students. The opportunity to participate in assessments, including national assessment tests should also be provided. "This may mean making adjustments to . . . how students will be assessed" (DDA Education Standards, 2009, p. 5). The Standards also provide information regarding exceptions for education providers.

Another important feature is the requirement that education providers must make 'reasonable adjustments' to facilitate the engagement of students with a disability in school activities. The standards note that an adjustment is reasonable in relation to a student with a disability when it balances the interests of all parties affected. For example, education authorities are not obliged to adjust courses or programs and assessment requirements to the extent that the academic integrity of the program is threatened. However, the student's disability, the student's wishes and the wishes of their advocate, the effect of the adjustment on the student with a disability and other students, and the cost of the adjustment should be considered in coming to a conclusion about the 'reasonableness' of an adjustment. That the texture of this aspect of the DDA is so open demonstrates that the standards themselves are a significant compromise, without which, the Australian community would still be waiting for their enactment.

At the time of writing, some educational jurisdictions are still coming to terms with the ramifications of the education standards. However, as the next section demonstrates, the standards are already having an impact on the way in which schools support students with a disability. In addition to impelling schools to embrace a wider definition of disability, schools are now

legally obliged to provide a minimum level of educational support to students with a disability (Nelson, 2003). Just how schools interpret this level of service will, no doubt, be open to interpretation and perhaps subject to legal challenge.

By 2009, there had been no Federal or High Court decisions that related to the participation of students with a disability in educational testing (there were several Federal Court decisions relating to other aspects of the education of students with a disability). Usually such matters are addressed through policy and administrative procedures and rarely reach courts or tribunals (Cumming, 2009a). However, several conciliated settlements had been made between the parties in this area (Australian Human Rights Commission, 2009d). Nine separate students with different disabilities at various schools complained that they had not been given sufficient adjustments by their respective schools to allow them to participate in exams. The schools were instructed to provide appropriate adjustments that included additional time, a reader, a writer, access to digital recordings of exams, a separate exam room, use of coloured examination paper, taking exams in a familiar setting, and the use of a word processor. A student who was deaf complained that he was not provided with a sufficiently qualified teacher to help interpret exams. The education authority subsequently made available a teacher with the necessary skills.

That the Australian DDA is not as prescriptive as US legislation about the rights of students with a disability can be seen as both a weakness and a strength in meeting the educational needs of these students. Supporters of stronger legislation say that laws guarantee a minimum standard and that they may create the circumstances by which attitudes to students with a disability in the general and education communities may change for the better. Those critical of stronger legislation point out that there is a difference between ‘following the letter of the law’ and improved attitudes and practice. The US experience following the mandating of individualised education programs for students with a disability is an example of how behaviour change does

not always follow from compliance with the law (Sopko, 2003). Whatever the view taken on legislation, the reliance in Australia on open-textured laws and education policy to provide educational services for students with a disability is a reflection of cultural standards and historical precedent in this country.

Australian Policy Relevant to Students with a Disability

National Goals for Schooling in the Twenty-First Century

The 1999 Adelaide Declaration on national goals for schooling followed a meeting of State, Territory, and Commonwealth Ministers of Education. The Declaration provided broad directions to guide schools and education authorities and to maintain a commitment to the reporting of comparable educational outcomes for all students, including students with a disability (MCEECDYA, 2010). Of interest to this chapter is the common and agreed commitment to

- continuing to develop curriculum and related systems of assessment, accreditation, and credentialling that promote quality and are nationally recognised and valued; and to
- increasing public confidence in school education through explicit and defensible standards that guide improvement in students’ levels of educational achievement and through which the effectiveness, efficiency, and equity of schooling can be measured and evaluated.

One of the goals specifically relates to students with a disability and calls for schooling to be socially just, so that students’ outcomes from schooling are free from the effects of negative forms of discrimination based on sex, language, culture and ethnicity, religion, or disability. This Declaration was the precursor to the plan to introduce national reporting of educational outcomes in numeracy and literacy and other curriculum areas.

The Melbourne Declaration on Educational Goals for Young Australians (Ministerial Council

on Education, Early Childhood Development, and Youth Affairs, 2008) was released in December, 2008 and superseded the Adelaide Declaration to set the direction for Australian schooling for the next 10 years.

The MCEETYA 4 Year Plan (2009–2012) supports the Melbourne Declaration and outlines the key strategies and initiatives Australian governments will undertake in eight inter-related areas in order to support the achievement of the educational goals for young Australians. The two areas of interest to this chapter are

- promoting world-class curriculum and assessment and
- strengthening accountability and transparency.

In the area of *promoting world-class curriculum and assessment*, the goals of a national curriculum are identified, before stating that “assessment of student progress will be rigorous and comprehensive” (p. 14, MCEETYA 4 Year Plan, 2009). The plan also identifies the need for government to work with all school sectors to develop and enhance national and school-level assessment that focuses on

- assessment for learning—enabling teachers to use information about student progress to inform their teaching,
- assessment as learning—enabling students to reflect on and monitor their own progress to inform their future learning goals and
- assessment of learning—assisting teachers to use evidence of student learning to assess student achievement against goals and standards.

Some of the agreed strategies and actions to support this plan are as follows:

- To establish the Australian Curriculum, Assessment and Reporting Authority (ACARA) to deliver key national reforms including development of plans to improve the capacity of schools to assess student performance and to link assessment to the national curriculum and to manage the National Assessment Program.
- Development of high-quality diagnostic and formative assessment tools and strategies to support teachers’ skills and understanding in the use (of) assessment as a tool for student learning and classroom planning and in

adapting instructional practice in classrooms to focus on specific student needs.

In terms of *strengthening accountability and transparency*, the plan identifies the need for good-quality information on schooling. Schools and students need reliable and rich data on student performance so as to improve student outcomes. Parents and families, however, need information about the performance of their son or daughter at school, of the school he or she attends, and of the system, to help parents and families make informed choices.

Through the Council of Australian Governments (COAG), all jurisdictions agreed to a new performance-reporting framework and agreed that ACARA will be supplied with the information necessary to enable it to publish relevant, nationally comparable information on all schools to support accountability, school evaluation, collaborative policy development, and resource allocation.

While the Melbourne Declaration and associated plans and policies identify students from low socio-economic backgrounds, including indigenous youth and other disadvantaged young Australians, students with disabilities receive minimal attention. This cohort, however, does receive some recognition in the conduct of the National Assessment Program.

The National Assessment Program

In 2008 the National Assessment Program—Literacy and Numeracy (NAPLAN) commenced in all Australian schools. The program continued in 2009 with all Australian students in Years 3, 5, 7, and 9 being assessed using common national tests in reading, writing, language conventions (spelling, grammar, and punctuation), and numeracy. Samples of students must also participate in other national and international tests, such as in Information Technology, Maths, and Science, but NAPLAN is the most comprehensive national assessment process that aims to include all students in the nominated grades each year. Commonwealth legislation requires

testing of *all* students in identified grades for federal school funding to be provided to the states.

The tests are developed collaboratively by the states and territories, the non-government education sectors, and the Australian Government. The NAPLAN tests broadly reflect aspects of literacy and numeracy within the curriculum in all states and territories, and the types of test questions and test formats are chosen for familiarity to both teachers and students across Australia. Content of the tests is also informed by the National Statement of Learning in English and Mathematics. Questions are multiple-choice type or require a short written response. National tests of literacy assessed language conventions (spelling, grammar, and punctuation), writing (knowledge and control of written language), and reading (comprehension). In numeracy, the content areas assessed were number, measurement, chance and data, space, algebra, function, and pattern. Students are expected to develop a range of important skills and strategies before sitting the tests. Resources to help students develop these skills are made available, and teachers are also able to use the preparation materials from previous NAPLAN and state tests to familiarise students with the types of questions and response formats on the tests.

Each state and territory has a test administration authority that is responsible for printing the NAPLAN tests each year, for test administration, data capture, marking, and the delivery of reports to the central Commonwealth body, the MCEETYA. National protocols for test administration ensure that the administration of the tests by the eight authorities is consistently applied. Principals and teachers are offered information on protocols by the relevant authority through a range of mediums including information sessions, written information, and web-based materials. Students are tested in their own schools, administered by their own school teachers, in mid-May of the school year. In terms of marking, the tests for reading, language conventions, and numeracy are marked using optical mark recognition software to score multiple-choice items. Writing tasks are professionally

marked by expert, independent markers using well-established procedures to maintain marker consistency.

National data are collected by the relevant test administration authority, and the de-identified student data are then submitted to an independent national data contractor for analysis. Comparative data showing the performance of each state, to determine national achievement scales, national means, and achievement of the middle 60%, for each domain of reading, writing, spelling, grammar and punctuation, and numeracy, are then provided to each testing authority. The national achievement scales each span Years 3, 5, 7, and 9. The skills and understandings assessed in each domain from Year three to Year nine are mapped onto achievement scales with scores that range from 0 to 1000 (MCEETYA, 2008).

The scores for individual students across the national achievement scale are provided to enable comparisons over time. Students are able to be located on a single national scale for each domain, and achievement could be assessed against national and state means and national minimum standards. Specific scale scores determined cutoffs for achievement bands from 1 to 10 for each domain. Reporting scales were constructed so that any given scaled score represents the same level of achievement over time, enabling monitoring of student achievement as students advance through year levels.

Later in the year schools receive statements of performance of their individual students and year levels as a whole in relation to the national minimum standards and the number and percentage of children not reaching national benchmarks. While decisions regarding intervention varied across states, regions, and schools, the decision regarding intervention is clear if the student achieves a level that is identified as below the national minimum standard. Schools then provide individual students (and their parents/carers) with statements of performance in relation to the national minimum standards. The means and standard deviations for each state and territory compared to national means and standard deviations for each domain are published in the full

report that is publicly available on the web (see MCEETYA, 2008; MCEECDYA, 2009b).

Student Participation

The policies on student participation are provided in more detail in materials developed by the educational authorities responsible for each state and territory. While based on the same information, these materials are presented differently. One of the states, Queensland, provided a Test Preparation Handbook that states that the tests are “structured to be inclusive of all students, within budgetary and administrative limitations...and all eligible students must sit for the tests, unless they are exempt or withdrawn by parents/carers” (QSA Test Preparation Handbook, 2009, section 4.0). Further, eligible students include those students of equivalent chronological age to a “typical” Year 3, 5, 7, or 9 student and involved in a special education facility or program.

Exempt Students

The Queensland Handbook indicates that students may qualify for exemption from one or more of the tests because of their “lack of proficiency in the English language, or because of significant intellectual and/or functional disability”. However, students with disabilities should “be given the opportunity to participate in testing if their parent/carer prefers that they do so” (QSA Test Preparation Handbook, 2009, section 4.2). Principals must consult with parents/carers on all matters of exemption, and then use ‘professional judgment’ when making decisions about a student’s participation in the tests. Principals are required to obtain signed forms

from parents/carers to allow students who meet the criteria to be exempted. In Australia, exempt students are judged to have achieved in each exempted test at the level “below national minimum standard” and are reported within this subgroup of the population of students who had participated in the tests.

Students may also be withdrawn from the testing program by their parents/carers in consultation with the school. Withdrawals are intended to address concerns such as religious beliefs and philosophical objections to testing. However, information guides from some educational authorities link withdrawal with students with disabilities. For example, the 2007 information guide for New South Wales, another Australian state, documented that “Students with confirmed disabilities or difficulties in learning are expected to participate in the testing. However, parents do have the right to withdraw their children from testing. This is classified as a parent withdrawal and not as an exemption” (p. 2).

Students who are withdrawn are not counted as part of the population. In terms of reporting, they are grouped with those students who were absent or suspended, and despite the principals’ facilitation were unable to complete the test(s) in the days immediately following the standard National testing day.

The numbers and percentages of students who are exempted or absent/withdrawn for each of the five domains are reported within participation statistics in the NAPLAN annual report. To provide a summary, the average percentages of students in each grade who were exempted, absent, or withdrawn across the five domains for the 2009 NAPLAN across Australia are provided in the Table 5.1.

In the 2009 report, some comparisons were made regarding participation levels of the

Table 5.1 Percentages of Australian student exemptions, absences/withdrawals, and assessed across year levels for 2009 NAPLAN

	Year 3 (<i>n</i> = 260,000)	Year 5 (<i>n</i> = 265,000)	Year 7 (<i>n</i> = 255,000)	Year 9 (<i>n</i> = 250,000)
Exempt	1.9	1.7	1.2	1.3
Absent/withdrawn	3.6	3.2	3.6	6.1
Assessed	94.5	95.1	95.2	92.6

2009 cohort compared with 2008. In general, these participation rates were very similar, although some year-to-year variations among indigenous populations in some states were noted.

No details were publicly available as to the reasons for exemptions or absences and withdrawals. Testing authorities mentioned that the written parent applications for exemptions or withdrawals subsequently approved by principals were kept at the school level and were not centrally recorded at the state or national level. It is unknown as to how many exempted students had identified disabilities, or were students with learning disabilities, or language difficulties. Similarly, the reasons for students being withdrawn or absent (or suspended) and parental philosophical objections for withdrawal are unknown. How many of those absent or withdrawn who might have had learning disabilities is also unknown.

Special Provisions/Considerations

Special provisions were to be made where these were identified as necessary to comply with the legislative requirements of the Disability Standards for Education (DDA). A range of support and differentiated resources was to be made available to enable all students to complete the NAPLAN testing. The information guide for NSW (2007) stated that special provisions were provided to students with disabilities or special needs in line with arrangements for existing state-based tests. Special provisions could be accessed by a student for all or part of a test, using more than one provision in any one test. Special provisions should also reflect the type of support the student regularly accesses in the classroom.

Queensland Special Provisions and considerations were included in the Test Preparation Handbook and also made freely available on a website: <http://www.learningplace.com.au/deliver/content.asp?pid=43511> (accessed Jan 18, 2010). The stated goal of special provisions/considerations is the maximisation of

student access to the NAPLAN tests. Schools were encouraged to consider the principles of equity and inclusivity in meeting the needs of all students by recognising that many students without disabilities might also require special provisions/consideration. While this extension of potential support is noteworthy, the notion of “reasonable adjustment” provides schools with the opportunity to avoid offering accommodations. After a process of consideration as to whether an adjustment was necessary, a reasonable adjustment, as previously outlined, was then to be identified and put into place. Schools were able to provide the following accommodations or reasonable adjustments, as were considered necessary (QSA Test Preparation Handbook, 2009, section 5.10):

- Reading support
- Use of a scribe.
- Extra time (up to 50%) including rest breaks
- Braille and large-print test materials
- Separate supervision or special test environment
- PCs/laptops (no spell check or speech-to-text software)
- Assistive listening devices
- Specialised equipment or alternative communication devices
- ‘Signed’ instructions.

However, a number of accommodations were *not permitted*. In terms of reading, it was not permitted to

- read numbers or symbols in numeracy tests
- interpret diagrams or rephrase questions
- read questions, multiple-choice distractors, or stimulus material in the reading or language conventions tests
- paraphrase, interpret, or give hints about questions or texts;
- literacy questions could not be read or signed to students with moderate/severe-to-profound hearing impairment.

Many of these disallowed accommodations would be judged by many to be quite “reasonable” in providing students with disabilities with the chance of an even playing field in understanding what is required and to then be able to complete assessment tasks.

The numbers and percentages of students who are afforded special consideration accommodated, and the types of accommodations are not provided in the annual report. Some testing authorities have reported the numbers who were given special consideration on at least one

Table 5.2 Queensland students afforded special consideration and exemptions, across year levels for 2009 NAPLAN

Year level	Total participants	Numbers given special consideration	Percentage of total given special consideration	Numbers exempted	Percentage of total exempted
3	56,368	7388	13.1	1123	1.9
5	57,467	6730	11.7	1068	1.8
7	58,182	6121	10.5	976	1.7
9	59,997	2834	4.7	1015	1.7

test booklet. For instance, the Queensland Studies Authority (QSA, 2009) provided the details for the Table 5.2 to be constructed.

No details were publicly available as to the types of special consideration afforded to students, and the types of disability of these students. Written parent applications for special consideration subsequently approved by principals were kept at the school level and were not centrally recorded at the state or national level.

Issues Arising from Exemptions/Absences/Withdrawals

Some concerns have been expressed by politicians and others in the print media regarding an increase in NAPLAN absenteeism in some schools in some states (*Sydney Morning Herald*, October 20, 2009). In this article published in the *Sydney Morning Herald*, a politician suggested that because of the publication of school performance data that there is “a fear that the school’s average and reputation will be damaged by individual poor results”. The Council of Australian Governments (COAG) agreement to a new performance reporting framework to publish relevant, nationally comparable information on all schools to support accountability and school evaluation is potentially threatening to poor performers. Some teachers, schools, school principals, education authorities, and governments will potentially experience ramifications if performances at their level of responsibility in the system fall below expectation and/or the national

standards. To avoid such negative outcomes, strategies at each level of the system are being put into place. For example, since the first NAPLAN testing in 2008, there is anecdotal evidence that teachers are spending more time on training their children on NAPLAN-type tasks from very early in the school year. School principals in some states are being offered financial and other incentives to increase the performance of their school. In the future, school funding and teacher performance pay will potentially be linked to performance on national tests (Cumming, 2009a). There is concern that at the student level, many poor-performing students and their parents will be discouraged by teachers/schools to participate. While these issues raise the potential for negative outcomes for students, many schools and teachers are adopting a more positive approach and identifying strategies to help students learn the curriculum and ultimately perform well on assessment tasks.

The figures provided in Tables 5.1, and 5.2 point to the fact that substantial numbers of students are not participating in national achievement testing. Non-participation is an indictment on the whole system and flies in the face of the legislation (DDA, Education Standards) and policies (National Schooling for the twenty-first century, Adelaide and Melbourne Declarations, MCEETYA 4 Year Plan) that promote assessment for all.

When students are exempted from national testing or are withdrawn or are absent because the tests are deemed not to be appropriate for those students, the national assessment program fails those students. Their lack of inclusion

places them and their parents in a position of obscurity. These students and their performance levels become less important to schools, and their parents do not receive meaningful performance information in terms of national standards. In terms of the school performance data, many of these students could be written off as not having the potential to achieve the current minimum national standards, and so when resources are allocated to improve performance, the learning needs of these students may be considered less important, compared to other students who might have potential to achieve these standards.

These are similar concerns to those raised in the United States more than 10 years ago in the report from the Committee on Goals 2000 and the Inclusion of Students with Disabilities for the National Research Council (McDonnell, McLaughlin, & Morison, 1997). In the Executive Summary of this report, concerns were identified about the exclusion from participation of many students with disabilities. When data on achievement levels of students with disabilities are absent, then “judgments about the effectiveness of educational policies and programs at local, state and national levels (p. 6)” are neither valid nor fair.

Issues Arising from Special Provisions/Considerations

There is evidence of special provisions and accommodations being applied to support students with additional needs. However, these sanctioned accommodations are regarded as minimal by some judges. Moreover, validity issues have been raised in relation to the limited application of accommodations and adaptations to NAPLAN among other assessment approaches (Cumming, 2009a). From a cultural perspective, while students with disabilities have diverse ways of knowing, we still frame accommodations from particular constructions of ways of knowing, “expected patterns of ‘normal’ development, and how the demonstration of knowledge should occur” (Cumming, 2009a, p. 5).

While legislation and policy calls for appropriate accommodations for all students with disabilities, alternative assessment forms to allow students to demonstrate their knowledge in other ways are not provided in NAPLAN. Cumming also raises concerns about the form of some NAPLAN assessment items dominating the task and precluding the demonstration of knowledge. Such assessment formats are not uncommon in large-scale testing, but they do not allow students with disabilities to demonstrate their skills and knowledge.

There is a need to develop alternative assessments for students with additional needs. In the United States, the No Child Left Behind Act (NCLB) requires all students to participate in statewide assessments. The challenge is to design an assessment system that signals high expectation of performance but still provides useful data about the progress of students at the lower end (McDonnell et al., 1997). Participation for some students with disabilities will require some form of testing accommodation that entail non-standard forms of test administration and response. Other students will require alternative assessments to accurately measure performance at the low end of the scale. The US Department of Education has developed guidelines for states that permit alternate and flexible assessments for special education students (Cortiella, 2007). Some students will require alternate assessment based on modified academic achievement standards (AA-MAS) that are determined, justified, and documented by the teaching or Individualised Education Plan (IEP) team. Assessment is aligned to grade-level content standards for the grade in which the student is enrolled, but may be less challenging than the grade-level achievement standard. This form of modified assessment must have at least three achievement levels. States can modify standards, or design a totally different assessment, or adapt the existing regular assessments. Some states adapted the regular assessment by reducing the number of test questions, others simplified the language of test items, reducing the number of multiple choice options, using pictures to aid understanding, and providing more white space

on the test booklet. The focus of the special education service is to accelerate learning to overcome achievement gaps.

A few students will require alternate assessment based on alternate achievement standards (AA-AAS) whereby the expectation of performance is lower than the grade-level achievement standard, and “usually based on a very limited sample of content that is linked to but does not fully represent grade-level content” (Cortiella, 2007, p. 7).

In essence, assessment needs to give credit for what knowledge can be demonstrated when not bound by the constraints of comparisons with other children, or a curriculum that does not reflect different ways of knowing (Cumming, 2009a). Educators and assessors will need to demonstrate clearer understandings of children and their ways of knowing before assessments can effectively meet the needs of all children.

Conclusion

The Ministerial Council for Education, Early Childhood Development and Youth Affairs (MCEECDYA) under its previous name (MCEETYA) had a facilitative role in the development of the Education Standards. Subsequent to the passing of the Education Standards into legislation, MCEECDYA has developed a national testing regime. It is concerning that a number of significant inconsistencies between the Standards (and other legislation and policies) and the application of NAPLAN have been identified.

While the legislation and policies require equity of opportunity in all school activities, many students with a disability are exempted from national testing. While the overall number of exemptions is known, the reasons for these exemptions are not known. Some students are offered adjustments and accommodations, but these are not centrally recorded or monitored for quality assurance in the states that have been reviewed. No alternative assessments for NAPLAN have been developed, and there appears to be no agenda to do so.

So while there would seem to be a national agenda of assessment of achievement for all, the practical application of these policies falls well short of the mark. It may well be that testing authorities in each state could justify their lack of alternative assessment items or tests behind the DDA Education Standards’ exception of “unjustifiable hardship” in not carrying out the assessment obligation because it “would be very expensive” (DDA Education Standards, 2009, p. 6).

National assessment needs to become more inclusive to meet the requirements of the DDA Education Standards. Public reporting of assessment results for students with disabilities along with those who participate in different or modified assessments are “key to ensuring fair and equitable comparisons among schools, districts, and states; in addition, all students should be accounted for in the public reporting of results” (McDonnell et al., 1997, p. 7). With the establishment of ACARA and the goal of alternative assessments, there is the promise that inclusive national assessment for all can take place in the near future. If alternate modes of assessment are not developed, there is an expectation of court challenges about the adequacy of current provisions to assess student achievement (Cumming, 2009b).

So the agenda is clear. Exemptions need to be minimised, and existing assessment protocols need to be adjusted, in particular, the level of expected performance needs to be reduced for each grade level. Conditions of testing also need to be adjusted, and test items need to be reviewed and modified to accommodate students with additional needs. Alternative assessment tools need to be designed to accommodate such students. While national testing has been newly introduced and is evolving within complex political and educational contexts, this testing regime needs to achieve full validity and provide equity for all. For this to be realised, it is essential that national assessments are truly inclusive.

References

- Australasian Legal Information Institute. (2009). *Disability Discrimination Act 1992*. Retrieved on December 3, 2009 from the World Wide Web: <http://www.austlii.edu.au/>
- Australian Council for Educational Research. (2009). *Programme for international student assessment*. Retrieved on December 3, 2009 from the World Wide Web: <http://www.acer.edu.au/index.html>
- Australian Government Attorney-General's Department. (2005). *Disability standards for education*. Retrieved on December 4, 2009 from the World Wide Web: http://www.ag.gov.au/www/agd/agd.nsf/Page/Humanrightsandanti-discrimination_DisabilityStandardsforEducation
- Australian Government Productivity Commission. (2004). *Review of the disability discrimination Act 1992*. Retrieved on December 4, 2009 from the World Wide Web: <http://www.pc.gov.au/>
- Australian Human Rights Commission. (2009a). *Court decisions*. Retrieved on August 19, 2009 from the World Wide Web: http://www.hreoc.gov.au/disability_rights/decisions/court/court.html#cted
- Australian Human Rights Commission. (2009b). *Disability complaint outcomes*. Retrieved on December 4, 2009 from the World Wide Web: http://www.hreoc.gov.au/disability_rights/index.html
- Australian Human Rights Commission. (2009c). *Disability standards and guidelines*. Retrieved on December 3, 2009 from the World Wide Web: <http://www.hreoc.gov.au/>
- Australian Human Rights Commission. (2009d). *Disability rights*. Accessed on August 19, 2009 from the World Wide Web: http://www.hreoc.gov.au/disability_rights/index.html
- Cortiella, C. (2007). *Learning opportunities for your child through alternate assessments: Alternate assessments based on modified academic achievement standards*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Cumming, J. J. (2009a). *Adjustments and accommodations in assessments that counts: Needed creativity in consideration of equity*. Paper presented at the 35th annual conference for the international association for educational assessment, Brisbane, 13–18 September 2009. Retrieved January 19, 2010 from the World Wide Web: <http://www.iaea2009.com/abstract/187.asp>
- Cumming, J. J. (2009b). Assessment challenges, the law and the future. In C. M. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st Century: Connecting theory and practice* (pp. 157–179). Dordrecht, The Netherlands: Springer International.
- DDA Education Standards. (2009). *Your right to an education: A guide for students with a disability, their associates, and education providers*. Retrieved January 19, 2010 from the World Wide Web: www.ddaeducationstandards.info
- Dempsey, I. (2004). Recent changes in the proportion of students identified with a disability in Australian schools. *Journal of Intellectual and Developmental Disability*, 29, 172–176.
- Dempsey, I. (2007). Trends in the placement of students in segregated settings in NSW government schools. *Australasian Journal of Special Education*, 31, 73–78.
- McDonnell, L. M., McLaughlin, M. J., & Morrison, P. (1997). *Educating one and all: Students with disabilities and standards-based reform*. Washington, DC: National Academy Press.
- Ministerial Council on Education, Early Childhood Development, and Youth Affairs. (2008). *National declaration on the educational goals for young Australians*. Retrieved January 19, 2010 from the World Wide Web: http://www.mceecdy.edu.au/verve/_resources/National_Declaration_on_the_Educational_Goals_for_Young_Australians.pdf
- Ministerial Council for Education, Early Childhood Development and Youth Affairs. (2009a). *Statements of learning*. Retrieved on December 3, 2009 from the World Wide Web: <http://www.mceecdy.edu.au/mceecdy/>
- Ministerial Council for Education, Early Childhood Development and Youth Affairs. (2009b). *NAPLAN 2009 report*. Retrieved on January 19, 2010 from the World Wide Web: http://www.mceecdy.edu.au/mceecdy/naplan_2009_report.29487.html
- Ministerial Council for Education, Early Childhood Development and Youth Affairs. (2010). *The Adelaide declaration on national goals for schooling in the twenty-first century*. Retrieved on January 27, 2010 from the World Wide Web: http://www.mceecdy.edu.au/mceecdy/adelaide_declaration_1999_text.28298.html
- Ministerial Council on Education, Employment, Training and Youth Affairs. (2008). *National assessment program: Literacy and numeracy. Achievement in reading, writing, language conventions and numeracy*. Retrieved on 18 March 2009 from the World Wide Web: <http://www.mceetya.edu.au>
- Ministerial Council on Education, Employment, Training and Youth Affairs. (2009). *MCEETYA four year plan 2009–2012*. Retrieved on January 14, 2010 from the World Wide Web: http://www.mceecdy.edu.au/mceecdy/action_plan.25966.html
- Nelson, B. (2003). *Most state and territory education ministers vote against disability standards*. Media release, July 11. Retrieved on January 28, 2010 from the World Wide Web: http://www.dest.gov.au/ministers/nelson/jul_03/minco703.htm
- NSW Dept of Education and Training. (2007). *National assessment program: Primary newsletter*. Retrieved on January 18, 2010 from the World Wide Web: http://www.curriculumsupport.education.nsw.gov.au/policies/nap/assets/nap_pri_newshi.pdf

- Queensland Studies Authority (QSA). (2009). *Test preparation handbook*. Brisbane, Australia: Retrieved on July 10, 2009 from the World Wide Web: QSA. <<http://qsa.qld.edu.au/assessment/8021.html>, and on Jan 18, 2010 on the learning place website: <http://www.learningplace.com.au/deliver/content.asp?pid=43511>
- Sopko, K. M. (2003). *The IEP: A synthesis of current literature since 1997*. ERIC Document Reproduction Service, No. ED476559.
- Sydney Morning Herald. (2009). *Schools play truant to avoid bad marks*. Anna patty. Media release. Retrieved on October 20, 2009 from the World Wide Web: <http://smh.com.au/national/schools-play-truant-to-avoid-bad-marks>
- World Bank. (2009a). *Gross domestic product*. Retrieved on December 3, 2009 from the World Wide Web: <http://www.worldbank.org/>
- World Bank. (2009b). *Population*. Retrieved on December 3, 2009 from the World Wide Web: <http://www.worldbank.org/>

Part II
Classroom Connections

Access to What Should Be Taught and Will Be Tested: Students' Opportunity to Learn the Intended Curriculum

Alexander Kurz

The annual large-scale assessment of student achievement for accountability purposes is expected to provide reliable test scores that permit valid inferences about the extent to which students have achieved the intended academic content standards. The fact that schools are open to sanctions and rewards on the basis of student achievement also indicates that these test score interpretations include inferences about what students know and can do as a result of the learning opportunities provided by schools (Burstein & Winters, 1994). That is, teachers and school administrators are considered to contribute to student achievement and, if properly incentivized and supported, are assumed to promote student achievement more effectively (Linn, 2008). To support the *validity* of such inferences, test users must either account for differences in learning opportunities or provide evidence that all test-takers had access to a comparable opportunity to learn the intended and assessed curriculum (Kurz & Elliott, in press; Porter, 1995).

The issue of opportunity to learn (OTL), especially in the context of test-based accountability, has also prompted researchers and other stakeholders to raise *equity* concerns for subgroups such as students identified with disabilities and English language learners (Heubert, 2004;

Herman & Abedi, 2004; Kurz, Elliott, Wehby, & Smithson, 2010; Porter, 1993; Pullin & Haertel, 2008; Stevens & Grymes, 1993). The differential achievement of students with and without disabilities on state and national achievement tests (Abedi, Leon, & Kao, 2008; Malmgren, McLaughlin, & Nolet, 2005; Ysseldyke et al., 1998) underscores this concern and has kept the question of access and OTL on the forefront of debates in special education: "Are students [with disabilities] receiving equitable treatment in terms of access to curriculum . . . and other inputs critical to the attainment of academic and educational outcomes important to their postschool success?" (McLaughlin, 2010, p. 274).

For nearly five decades, research related to OTL has focused on unpacking the "black box" of schooling processes that translate inputs into student outcomes (McDonnell, 1995). The conceptual and methodological challenges of OTL notwithstanding (Elliott, Kurz, & Neergaard, in press; Kurz & Elliott, 2011; Roach et al., 2009), measurement related to this concept has evolved over the years allowing stakeholders to ascertain relevant information about OTL at the system, teacher, and student level. Broadly defined, OTL refers to "the opportunities which schools provide students to learn what is expected of them" (Herman, Klein, & Abedi, 2000, p. 16). Researchers from various disciplines, however, have focused on different aspects of OTL that facilitate student learning of what is expected, which has resulted in a wide array of definitions

A. Kurz (✉)

Department of Special Education, Peabody College
of Vanderbilt University, Nashville, TN 37067, USA
e-mail: alexander.kurz@vanderbilt.edu

and ways to measure OTL. Moreover, the issues of *alignment* and *access* (to the general curriculum) intersect with OTL making it all the more difficult for stakeholders to understand and apply OTL and its measurement tools in the service of promoting student achievement for different subgroups. Hence the main objectives of this chapter: (a) situate the concept of OTL in the context of a comprehensive curriculum framework; (b) clarify the conceptual and substantive relevance of OTL; (c) synthesize aspects of OTL emphasized in the literature into a coherent and empirically based concept; (d) highlight OTL measurement options; and (e) conclude with a prospective view on OTL. Although McLaughlin (2010) recently noted that we can only “speculate” about the factors that may have contributed to the current disparate educational outcomes for students with disabilities and that we “cannot know with certainty whether greater access to rigorous courses and higher expectations would make a difference to both in school and postschool outcomes” (p. 274), I contend that the concept of OTL and its measurement, as discussed throughout this chapter, can move us *beyond speculation* toward actual measurement of access and the role it plays in promoting student achievement.

Student Access to the Intended Curriculum

The educational lives of students with disabilities are embedded in a legal framework based on legislation, such as the American with Disabilities Act (ADA, 1990), the Individuals with Disabilities Education Improvement Act (IDEA, 2004), and the No Child Left Behind Act (NCLB, 2001), which mandates physical and intellectual access at all levels of the educational environment including curriculum, instruction, and assessment (Kurz & Elliott, 2011). That is, students with disabilities are to be provided with a free and appropriate public education that meets their individual needs and offers them access to, and progress in, the general curriculum. Moreover, students with disabilities are to be fully included in test-based accountability and

offered content that is maximally aligned with the grade-level content standards of their general education peers. To some, this legal framework signals “a clear presumption that all students with disabilities should have *access to the general curriculum* [emphasis added] and to the *same opportunity to learn* [emphasis added] challenging and important content that is offered to all students” (McLaughlin, 1999, p. 9). The extent to which standards-based accountability along a general curriculum for all students can be reconciled with the notion of an individualized education for students with disabilities continues to be a matter of debate (e.g., Fuchs, Fuchs, & Stecker, 2010; McDonnell, McLaughlin, & Morison, 1997; McLaughlin, 2010). To shed light on this debate and to develop the conceptual and substantive relevance of OTL, it is necessary to set the stage via a comprehensive curriculum framework.

The Intended Curriculum Model

Several researchers have introduced curriculum frameworks (e.g., Anderson, 2002; Porter, 2002; Webb, 1997) that emphasize the content of three major elements of the educational environment: the *intended curriculum* (i.e., the content designated by state and district standards), the *enacted curriculum* (i.e., the content of teacher instruction), and the *assessed curriculum* (i.e., the tested content of assessments). Alignment refers to the extent to which the contents of these curricula overlap and serve in conjunction with one another to promote student achievement (Webb, 2002). In a well-aligned system, the content of teachers’ instructional activities covers the content intended by the standards, which in turn are measured (or rather sampled) by the tested content of state assessments. Several methods for the quantitative and qualitative analysis of alignment between these curricula are available (see Martone & Sireci, 2009; Roach, Niebling, & Kurz, 2008).

Building on prior work (Petty & Green, 2007; Porter, 2006), Kurz and Elliott (2011) extended the traditional three-part curriculum framework

to explicate issues of access and OTL for students with disabilities at the system, teacher, and student level. In the context of their intended curriculum model (ICM), the authors defined OTL as being concerned with “students’ opportunity to learn the intended curriculum” (p. 38). The present version of the ICM has been revised to clarify several questions that were prompted by the original model related to the conceptual relevance of OTL: Are the concepts of *alignment* and OTL interchangeable? Are the concepts of *access* to the general curriculum and OTL interchangeable? Is the *intended curriculum* the same for students with and without disabilities? Consequently, a general education and special education version of the model were specified. Figure 6.1 presents the ICM for general education.

The ICM for General Education

At the system level, the ICM posits the intended curriculum as the primary target of schooling. The intended curriculum hereby represents a collection of educational objectives, which in their entirety encompass the intended purposes of schooling (i.e., what students are expected to know and be able to do). Ideally, the intended curriculum identifies all valued and expected outcomes via operationally defined and measurable objectives at different levels of aggregation such as subject and grade. Under the NCLB Act (2001), states were required to develop challenging academic content and performance standards that specify “what” and “how much” is expected of students in mathematics, reading/language arts, and science (Linn, 2008). This federal mandate was intended to compel states

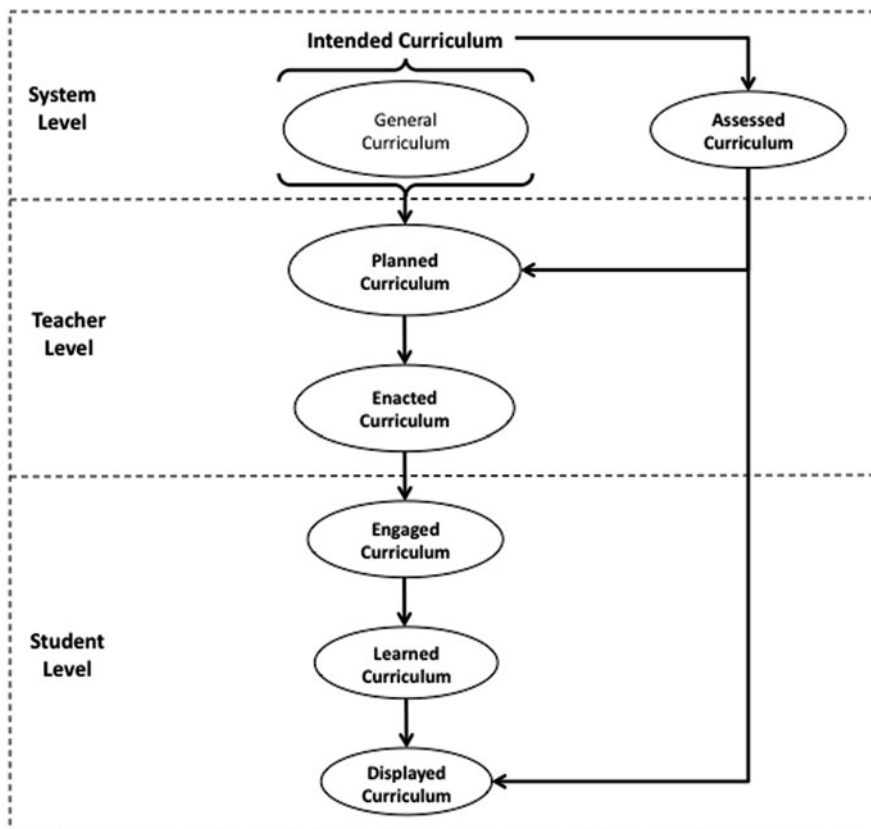


Fig. 6.1 The intended curriculum model for general education

to define and improve the so-called “general curriculum” (Karger, 2005). NCLB further described the general curriculum as applicable to *all* students – hence the term “general.” The statute’s implementing regulations, for example, stated that NCLB requires “each State to develop grade-level academic content and achievement standards that [NCLB] expects all students – including students with disabilities – to meet” (67 F.R. 71710, 71741). Additional legislative mandates that circumscribe or augment this general curriculum are not available for students without disabilities. The academic content and performance standards that comprise the general curriculum at the system level thus signal the entirety of the intended curriculum for students without disabilities. In other words, the general curriculum *is* the intended curriculum in the context of general education. For students with disabilities, however, the intended curriculum is not under the exclusive purview of the general curriculum – as will be discussed shortly.

The assessed curriculum for accountability purposes is designed at the system level in alignment with the intended curriculum. That is, the tested content of a state’s large-scale assessment is used to sample exclusively across the various content domains of the intended curriculum to permit a valid test score inference about the extent to which students have achieved the intended curriculum. It would be unreasonable to expect state tests to cover all skills prescribed by the intended curriculum due to test length and time constraints. Figure 6.1 therefore displays the assessed curriculum as being slightly smaller than the intended (general) curriculum. Under the NCLB Act (2001), all states are required to document alignment between the intended and assessed curriculum (Linn, 2008). Alignment methodologies such as the Surveys of the Enacted Curriculum (SEC; Porter & Smithson, 2001) and the Webb method (Webb, 1997) allow stakeholders to provide evidence of alignment beyond a simple match of content topics using additional indices such as content emphasis and match of cognitive process expectations (see Martone & Sireci, 2009; Roach et al., 2008). Lastly, it is important to note that the uniform description of

the intended curriculum via the general curriculum results in only *one* assessed curriculum for accountability purposes: the annual state achievement test.

At the teacher level, the ICM posits the planned curriculum as the first proxy of the intended curriculum. The planned curriculum represents a teacher’s cumulative plans for covering the content prescribed by the intended curriculum. Although the intended curriculum informs what content should be covered for a particular subject and grade, a teacher’s planned curriculum is likely to be constrained as a function of the teacher’s subject matter knowledge or familiarity with the intended curriculum. For example, a teacher may deliberately plan to emphasize certain content domains and omit others, while a different teacher may simply be unable to plan for comprehensive coverage of the intended curriculum due to missing content expertise and professional development experiences. To date, the content of teachers’ planned curriculum and its alignment with the intended curriculum have received limited research attention. As part of their alignment study, Kurz et al. (2010) adapted the SEC methodology to examine alignment between teachers’ planned curriculum and the state’s intended curriculum for 18 general and special education teachers. Results via the SEC’s alignment index (AI), which represents content alignment along two dimensions (i.e., topics and cognitive demand) on a continuum from 0 to 1, indicated that approximately 10% of teachers’ self-reported planned curriculum (for the first half of the school year) was aligned with the intended curriculum. Although more research is needed, the planned curriculum represents a viable target for professional development, because a teacher’s planned curriculum directly informs and potentially constrains his or her enacted curriculum. In the Kurz et al. study, for example, alignment between the planned and enacted curriculum was significantly greater (about 45%) than between the intended and enacted curriculum (about 10%). That is, teachers appear to adhere first and foremost to their own planned curriculum (rather than the intended curriculum). Lastly, the model

indicates that the planned curriculum is informed by both the intended and assessed curriculum. In the context of test-based accountability, the content of the assessed curriculum exerts a strong influence on what teachers plan to cover and ultimately implement. Under the NCLB Act (2001), the intended and assessed curriculum have to be aligned, which should allow teachers to focus their planning and teaching efforts on the intended curriculum. Misalignment, however, may pressure teachers to focus on the assessed curriculum because inferences about their effectiveness are made on the basis of test scores – in short, teachers may “teach to the test” rather than the broader intended curriculum.

The next proxy of the intended curriculum at the teacher level is the enacted curriculum, which largely comprised of the content of classroom instruction and its accompanying materials (e.g., textbooks). Teachers also make pedagogical decisions about the delivery of this content including instructional practices, activities, cognitive demands, and time emphases related to the teaching of certain topics and skills. The enacted curriculum plays a central role in our definition and measurement of OTL (i.e., students' opportunity to learn the intended curriculum) because it is primarily through the teacher's enacted curriculum that students access the intended curriculum. The enacted curriculum consequently represents one of the key intervention targets for increasing OTL. As can be seen in Fig. 6.1, the model again illustrates the potentially attenuated uptake of the intended curriculum by each subordinate curriculum. At this level, the day-to-day realities of school instruction may prevent teachers from enacting their entire planned curriculum in response to students' rate of learning, school assemblies, absences, and so on. The extent to which students have the opportunity to learn the intended curriculum via the teacher's enacted curriculum, however, is critical to their performance on achievement tests, even after controlling for other factors (e.g., Cooley & Leinhardt, 1980; Porter, Kirst, Osthoff, Smithson, & Schneider, 1993; Stedman, 1997). Moreover, providing students with the opportunity to learn the content that they are expected

to know represents a basic aspect of fairness in testing, particularly under high-stakes conditions (see American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). OTL also plays a role in the validity of certain test score inferences such as those that interpret assessment results as a function of teacher instruction or that explain mean test score differences between subgroups of examinees: “OTL provides a necessary context for interpreting test scores including inferences about the possible reasons underlying student achievement (e.g., teacher performance, student disability) and suggestions for remedial actions (e.g., assignment of PD training, referral to special education)” (Kurz & Elliott, p. 39).

At the student level, the engaged curriculum represents those portions of content coverage during which the student is engaged in the teacher's enacted curriculum. Considering data from the 2006 High School Survey of Student Engagement, on which 28% of over 80,000 students reported being unengaged in school (Yazzie-Mintz, 2007), it seems reasonable to suggest that some students are unlikely to engage in a teacher's entire enacted curriculum as it unfolds across the school year. Moreover, a student's engaged curriculum is likely to constrain his or her learned curriculum. That is, a student will presumably learn only those portions of the enacted curriculum during which he or she is engaged. The ICM thus indicates the potential for further attenuation as the intended curriculum reaches the student level via the teacher's enacted curriculum. At the end of the intended curriculum chain, the model posits the displayed curriculum, which represents the content of the intended curriculum that a student is able to demonstrate via classroom tasks, assignments, and/or assessments. A student's displayed curriculum may not reveal the entirety of his or her learned curriculum due to various factors including interactions between test-taker characteristics and features of the test that do not permit the student to fully demonstrate his or her knowledge of the target construct (see Chapter 9, this volume), test

anxiety, or constraints directly linked to the assessed curriculum. The latter concern is essentially an issue of alignment: A student can only “display” his or her achievement of the intended curriculum to the extent to which the assessed curriculum is aligned with the intended curriculum.

So far, we have discussed how the intended curriculum unfolds across the system, teacher, and student level in *general education*. It is within this educational context that most states use the general curriculum (i.e., the academic content and performance standards applicable to all students) to define their students’ intended curriculum. As such, it is not surprising that researchers have failed to see the need to distinguish between the general and intended curriculum at the system level: In the context of general education both curricula are indeed synonymous. However, the uncritical adoption of traditional curriculum

models in the context of special education can blur important distinctions among curricula that determine the intended outcomes of schooling for students with disabilities (i.e., what students are expected to know and be able to do). In fact, an ongoing debate in special education centers around the perceived tension between two federal policies relevant to standards-based reform and questions about the extent to which the newly established standards should circumscribe the intended and assessed curriculum for students with disabilities: “There is increasing recognition of a fundamental tension between the prevailing K-12 educational policy of universal standards, assessments, and accountability as defined through [NCLB] and the entitlement to a Free Appropriate Public Education (FAPE) within IDEA” (McLaughlin, 2010, p. 265). Figure 6.2 presents the ICM for special education.

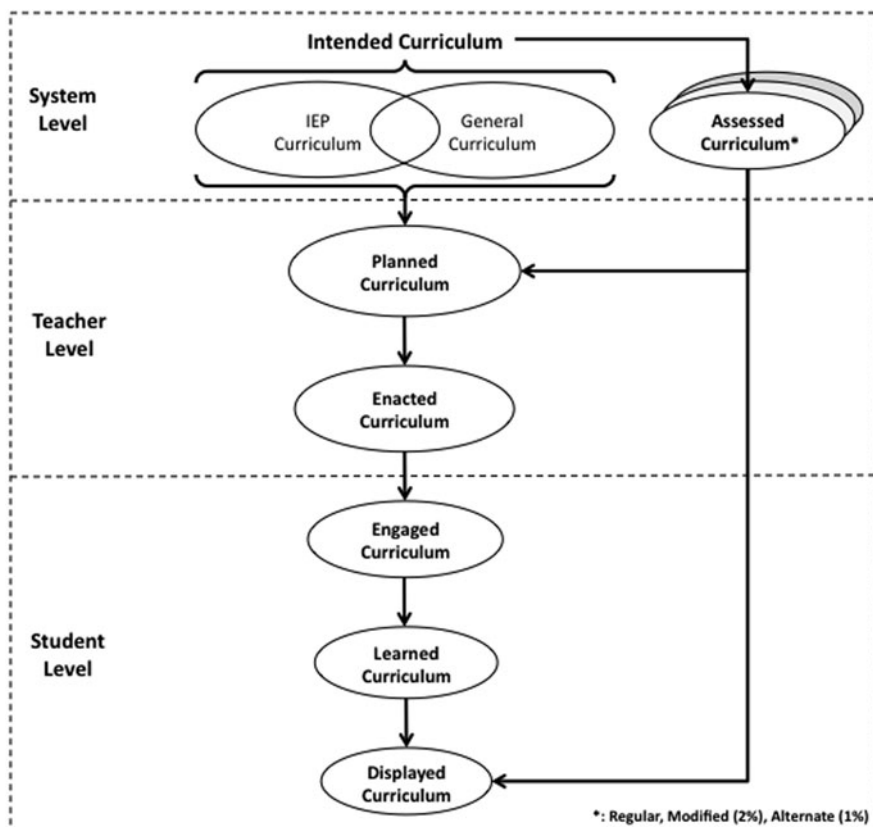


Fig. 6.2 The intended curriculum model for special education

The ICM for Special Education

In the context of special education, the ICM posits the intended curriculum as dually determined by both the general curriculum and the student's Individualized Education Program (IEP) curriculum. That is, neither the IEP curriculum nor the general curriculum exclusively represents the intended curriculum for a student with a disability. The implementing regulations for the reauthorization of IDEA in 1997 identified the intended purposes of special education as follows: "To [(a)] address the unique needs of the child that result from the child's disability; and [(b)] to ensure access of the child to the general curriculum, so that he or she can meet the educational standards within the jurisdiction of the public agency that apply to all children" (34 C.F.R. § 300.26(b)(3)). Both reauthorizations of IDEA in 1997 and 2004 further emphasized the IEP as the central mechanism for detailing a student's access, involvement, and progress in the general (education) curriculum (Karger, 2005). The IEP is used further to document educational objectives relevant to the student's present levels of performance as well as accommodations and modifications that facilitate the student's access to the enacted and assessed curriculum (Chapter 7, this volume). The IEP curriculum can thus include content that goes beyond the knowledge and skills put forth in the general curriculum. A student's IEP, for example, can include social and behavioral goals or other functional goals that are not part of subject- and grade-specific academic standards. The requirement to document a student's access, involvement, and progress in the general curriculum also has promoted the development of so-called "standards-based IEPs," which refers to the practice that links IEP objectives to a state's grade-level standards and assessments (Ahearn, 2006). As such, a student's IEP may include specific objectives that come directly from the general curriculum of his or her peers. In short, the IEP curriculum delineates the extent to which the general curriculum is part of the student's intended curriculum and includes a set of specific (intended) educational objectives, which, depending on the student's

unique disability-related needs, may fall within or outside the general curriculum. To this end, the ICM depicts overlap between the IEP curriculum and the general curriculum. The degree to which both curricula overlap is specified in each individual student's IEP and thus varies from student to student. Consequently, there is no uniform intended curriculum in the context of special education: The intended curriculum for students with disabilities is *student specific* by law.

The possibility of individualized intended curricula has direct implications for the remaining curricula within the ICM framework. Most importantly, the notion of only one assessed curriculum fully aligned with the general curriculum and applicable to all students is no longer tenable. For purposes of the assessed curriculum, the ICM therefore reflects the three assessment options currently available to students with disabilities: the *regular* state assessment, the alternate assessment based on *modified* academic achievement standards (AA-MAS), and the alternate assessment based on *alternate* academic achievement standards (AA-AAS). According to the model, the varying degrees of overlap between the IEP curriculum and the general curriculum can be grouped into three broad categories of the intended curriculum that directly correspond to the three assessed curricula: regular, modified, and alternate.

For students with disabilities whose IEP curriculum largely overlaps with the general curriculum, the intended curriculum should lead to planned and enacted curricula that offer students the opportunity to learn grade-level subject matter content and progress toward predetermined NCLB achievement goals. The content of the *regular* achievement test thus represents the appropriate assessed curriculum. As for students without disabilities, the resulting displayed curriculum would be used to monitor educational progress.

For students with disabilities whose IEP curriculum moderately overlaps with the general curriculum, the intended curriculum should lead to planned and enacted curricula that continue to offer students the opportunity to learn grade-level subject matter content and progress toward

predetermined NCLB achievement goals. However, we would expect the non-overlapping portions of the IEP curriculum to include modified outcomes for some general curriculum objectives, a set of non-academic educational objectives (e.g., social and behavioral goals), as well as more intensive and specialized accommodations and related services that support OTL. The content of the *modified* achievement test thus represents a more appropriate assessed curriculum. Progress monitoring via the resulting displayed curriculum would be benchmarked to modified and regular achievement standards.

For students with disabilities whose IEP curriculum barely overlaps with the general curriculum, the intended curriculum should lead to highly individualized planned and enacted curricula that offer students the opportunity to learn subject matter content that is linked to a limited and not fully representative sample of grade-level content. We would expect the non-overlapping portions of the IEP curriculum to represent alternate outcomes for most general curriculum objectives, a large set of non-academic educational objectives (e.g., social, behavioral, and functional goals), intensive and specialized accommodations and modifications, and several related services that support OTL. The content of the *alternate* achievement test thus represents a more appropriate assessed curriculum. Progress monitoring via the displayed curriculum would occur against highly differentiated outcomes related to functional independence and self-sufficiency.

As for students without disabilities, the intended curriculum is subject to change on an annual basis as students advance from one grade to the other. Besides subject- and grade-specific changes in the general curriculum, students with disabilities also experience an annual review and update of their IEP. Additional changes in the IEP curriculum are therefore very likely. Ongoing feedback loops from the displayed curriculum to the curricula at the teacher level (i.e., planned and enacted) and system level (i.e., intended) should further permit formative changes in the content of the intended curriculum and the planning and implementation of classroom instruction. Lastly, it should be noted that the discussed intended

curriculum categories serve illustrative purposes and do not suggest separate “tracks” of intended special education curricula.

At the teachers and student level, the intended curriculum unfolds in much the same way as described previously in the general education context. However, the student-specific nature of the IEP curriculum implies that the content of a teacher’s planned and enacted curricula ought to reflect each student’s unique intended curriculum. Differentiated instruction according to the specific needs and abilities of each student, of course, represents the very essence of special education and summarizes much of the teacher training content for special educators. The sources of instruction for students with disabilities responsible for implementing their intended curriculum, however, are rarely comprised of only special education teachers. In most cases, general and special education teachers share the responsibility of providing a student with the opportunity to learn his or her intended curriculum, supported by paraprofessionals, teacher consultants and specialists, and other related services providers. The fragmentation of OTL sources therefore presents a significant measurement challenge – as will be discussed in a later section of this chapter.

The Relevance of OTL

Up until now, our general definition of OTL was used to establish the “who” (i.e., students) and “what” (i.e., the intended curriculum) of OTL and to pinpoint “how” this opportunity to learn the intended curriculum presents itself to students, namely through the teacher’s enacted curriculum. Not surprisingly, researchers interested in measuring OTL have concentrated their efforts on the enacted curriculum (e.g., Pianta, Belsky, Houts, Morrison, & National Institute of Child Health and Human Development [NICHD], 2007; Rowan, Camburn, & Correnti, 2004; Smithson, Porter, & Blank, 1995). Within the context of the ICM, the focus has been on the *content* of the various curricula at the system, teacher, and student level. However,

the content of the enacted curriculum (i.e., the content of teacher instruction) represents only one dimension of OTL examined in the literature. At the enacted curriculum level, two additional dimensions related to the concept of OTL – *time* on instruction and *quality* of instruction – have received sustained research attention for nearly five decades. Before tracing the three dimensions of OTL discussed in the literature, it is helpful to briefly review the “why” of OTL in light of the ICM framework.

Conceptual Relevance

The importance of OTL can be separated into its conceptual and substantive relevance. The conceptual relevance of OTL lies in the fact that it intersects and informs the concepts of alignment and access to the general curriculum. To that end, three questions were introduced earlier: Are the concepts of *alignment* and OTL interchangeable? Are the concepts of *access* to the general curriculum and OTL interchangeable? Is the *intended curriculum* the same for students with and without disabilities? The last question already has been answered via the two versions of the ICM. For students without disabilities, the intended curriculum is exclusively comprised of the general curriculum, which is the *same* for all students. For students with disabilities, the intended curriculum is comprised of a student's unique IEP curriculum and (to varying degrees) the general curriculum, which results in a *different* intended curriculum for each student. The answer to the second question is also “no.” Access to the general curriculum is a policy mandate under IDEA (1997, 2004) and applies only to students with disabilities. As defined in this chapter, the concept of OTL is concerned with a student's intended curriculum and hence is applicable to both students with and without disabilities. According to the ICM, however, the general curriculum is always part of the intended curriculum (at least to some degree), which implies that documentation of OTL can also serve as an indicator of access to the general curriculum. In this sense, access to the general curriculum and OTL are related but not interchangeable concepts.

The first question requires further clarification before it can be answered. Alignment is always established between two curricula. With respect to OTL, researchers have suggested alignment between the enacted and intended curriculum as a possible proxy (e.g., Kurz et al., 2010; Porter, 2002; Smithson et al., 1995). The revised question thus reads as follows: Are the concepts of *alignment between the enacted and intended curriculum* and *students' opportunity to learn the intended curriculum* interchangeable? Unfortunately, an answer to this question depends on the constraints of the alignment method used. First, current alignment methodologies do not account for the IEP curriculum as part of the intended curriculum (see Martone & Sireci, 2009; Roach et al., 2008). The overlap between the content of classroom instruction and academic standards could thus represent a narrow aspect of students' opportunity to learn the intended curriculum. Second, interchangeable use of both concepts would imply that one considers the content dimension of the enacted curriculum (i.e., the degree to which its content is aligned with state content standards) as a sufficient indicator of OTL – let alone the assumption that the teachers' enacted curriculum represents the critical juncture at which to measure OTL.

Substantive Relevance

The substantive relevance of OTL underscores the need to document OTL because of its impact on four related issues: (a) *validity* of test score inferences (e.g., Burstein & Winters, 1994; Kurz & Elliott, 2010; Wang, 1998); (b) *equity* in terms of educational opportunity to learn the intended curriculum and tested content (e.g., Pullin, 2008; Yoon & Resnick, 1998); (c) *compliance* with federal policy for students with disabilities (e.g., Kurz et al., 2010; McLaughlin, 2010); and (d) *student achievement* (e.g., Gamoran, Porter, Smithson, & White, 1997; Kurz et al., 2010; Wang, 1998). First, documenting OTL can assist stakeholders in the current test-based accountability system to draw valid inferences about possible reasons for differential subgroup

achievement or the extent to which students have learned subject matter content as a function of teacher instruction. Second, OTL in the context of the ICM has highlighted that equal opportunity to learn tested content for all students represents a limited view of OTL because it is not targeted toward students' intended curriculum. Moreover, the intended curriculum for students with disabilities is legally augmented via their IEP curriculum, which establishes an *individualized* intended curriculum reflective of their unique abilities and needs. Equal OTL across all student groups could therefore lead to unequal outcomes for students with disabilities. OTL as defined within the ICM highlights *equitable* OTL in the context of special education. That is, opportunity to learn the intended curriculum should not be equal across students due to the student-specific nature of the intended curriculum in special education (as attested by special education practices such as modified instructional content, additional time on task, or differentiated instruction). In short, students with disabilities should receive equitable OTL according to their individual abilities and needs. Third, documenting OTL can provide evidence of access to general curriculum for students with disabilities as mandated under IDEA (1997, 2004) and the NCLB Act (2001), as well as access to tested content for all students as supported by court rulings (for purposes of graduation), such as *Debra P. v. Turlington* (1981). Fourth, researchers have used the concept of OTL and its measurement dimensions, especially at the enacted curriculum level, to identify significant predictors of student achievement. Documenting OTL, and ultimately increasing OTL, thus represents an important avenue to improve student outcomes. As mentioned earlier, research on OTL at the enacted curriculum level has focused on several different instructional dimensions, all of which were identified as contributors to student achievement.

In summary, OTL is an important concept that warrants measurement. Prior to discussing different measurement options, it is necessary to establish (a) at what level of the educational environment and via which curriculum to document OTL and (b) the respective curriculum

dimensions to be measured. So far, we have used the curriculum framework of the ICM to substantiate the teacher's enacted curriculum as students' key access point to the intended curriculum. The next section thus examines measurement dimensions of OTL at the enacted curriculum level as emphasized in the literature.

Instructional Dimensions of OTL

For nearly five decades, researchers from a variety of disciplines have focused on different process indicators at the system (e.g., per-pupil expenditures), teacher (e.g., content overlap), and student level (e.g., engagement) to (a) obtain descriptions of the educational opportunities provided to students; (b) monitor the effects of school reform efforts; and (c) understand and improve students' academic achievement (Kurz & Elliott, 2011). Arguably the most proximal variable to the instructional lives of students and their opportunity to learn the intended curriculum is the teacher's enacted curriculum: "[S]tudents' opportunities to learn specific topics in the school curriculum are both the central feature of instruction and a critical determinant of student learning. The importance of curricular content to student learning has led researchers to become increasingly interested in measuring the 'enacted curriculum' . . ." (Rowan et al., 2004, pp. 75–76). Starting in the 1960s, separate OTL research strands started to form around three different instructional dimensions of the enacted curriculum: time on instruction (e.g., Carroll, 1963), content of instruction (e.g., Husén, 1967), and quality of instruction (e.g., Brophy & Good, 1986). Each strand is briefly reviewed next.

Time on Instruction

The first research strand emerged with John Carroll (1963), who introduced the concept of OTL as part of his model of school learning: "*Opportunity to learn* is defined as the amount of time allowed for learning, for example by

a school schedule or program” (Carroll, 1989, p. 26). Carroll included OTL as one of five variables in a mathematical formula, which he used to express a student’s degree of learning (i.e., ratio of the time spent on a task to the total amount of time needed for learning the task). Subsequent research on time and school learning began to refine this OTL conceptualization from *allocated time* into “engaged time,” “instructional time,” and “academic learning time” (see Borg, 1980; Gettinger & Seibert, 2002). The concept of academic learning time (ALT) introduced by Fisher and colleagues (1980), for example, considers the amount of time a student is engaged in an academic task that he or she can perform with high success. The amount of time dedicated to instruction has received substantial empirical support in predicting student achievement (e.g., Carroll, 1989; Denham & Lieberman, 1980; Fisher & Berliner, 1985; Walberg, 1988). In a research synthesis on teaching, Walberg (1986) identified 31 studies that examined the “quantity of instruction” and its relation to student achievement. Walberg reported a median (partial) correlation of 0.35 controlling for other variables such as student ability and socioeconomic status. In a meta-analysis on educational effectiveness, Scheerens and Bosker (1997) examined the effect of (allocated) time on student achievement using 21 studies with a total of 56 replications across studies. The average Cohen’s *d* effect size for time was 0.39 (as cited in Marzano, 2000). Both research reviews, however, provided insufficient information about the extent to which time usage was reported by special education teachers and failed to disaggregate the relation between time and student achievement for students with and without disabilities. Considering that time usage related to instruction represents one of the best documented predictors of student achievement across schools, classes, student abilities, grade levels, and subject areas (Vannest & Parker, 2010), it is not surprising that research regarding time on instruction continues across the system (i.e., allocated time), teacher (i.e., instructional time), and student level (i.e., ALT) of the ICM.

Despite the fact that NCLB has posited increased time on instruction as an important

avenue for improving the academic achievement of *all* students (Metzker, 2003), little is known about the extent to which special education teachers spend time on instruction (Vannest & Hagan-Burke, 2010). Special education has been marked by significant changes in teacher roles, settings, and instructional arrangements over the last few decades, which have increased the number of activities that require substantial amounts of teacher time such as paperwork, consultation, collaboration, assessment, and behavior management (e.g., Conderman & Katsiyannis, 2002; Kersaint, Lewis, Potter, & Meisels, 2007; Mastropieri & Scruggs, 2002). The majority of available data, however, are anecdotal. In one of the first studies that used teacher self-reports in conjunction with continuous and interval direct observation data, Vannest and Hagan-Burke (2010) reported on the results of 2,200 hours of data from 36 special education teachers. Two findings are noteworthy: (a) Time use for 12 different activities ranged from 2.9 to 15.6%, which indicates that no single activity took up the majority of the hours of the day; (b) academic instruction, instructional support, and paperwork occupied large percentages of time with 15.6, 14.6, and 12.1%, respectively. Vannest and Hagan-Burke concluded that “the sheer number of activities in which [special education] teachers engage is perhaps more of an issue than any one type of activity, although paperwork (12%) certainly reflects a rather disastrously large quantity of noninstructional time in a day” (p. 14).

In summary, time on instruction represents an important instructional dimension of the enacted curriculum and has received substantial empirical support as a strong contributor to student achievement. Unfortunately, research data on time usage for special education teachers are virtually non-existent, especially in relation to student achievement. Moreover, the limited research available for special education teachers indicates that large percentages of time are occupied by non-instructional activities, which raises concerns about the total amount of time a special education teacher can dedicate to instruction (see Vannest & Hagan-Burke, 2010).

Content of Instruction

The second research strand emerged with studies that focused on the *content overlap* between the enacted and assessed curriculum (e.g., Comber & Keeves, 1973; Husén, 1967). Husén, one of the key investigators for several international studies of student achievement, developed an item-based OTL measure that required teachers to report on the instructional content coverage for each assessment item via a three-point Likert scale: “Thus opportunity to learn from the Husén perspective is best understood as the match between what is taught and what is tested” (Anderson, 1986, p. 3682). To date, the International Association for the Evaluation of Educational Achievement (IEA) has conducted six comparative studies of international student achievement, the results of which have supported students’ opportunity to learn the assessed curriculum as a significant predictor of systematic differences in student performance. This content overlap conceptualization of OTL remained dominant in several other research studies during the 1970s and 1980s, all of which focused on general education teachers (e.g., Borg, 1979; Mehrens & Phillips, 1986; Walker & Schaffarzick, 1974; Winfield, 1987). For their meta-analysis, Scheerens and Bosker (1997) reviewed 19 studies focused on teachers’ content coverage of tested content and reported an average Cohen’s *d* effect size of .18 (as cited in Marzano, 2000).

Another line of research on content overlap focused on students’ opportunity to learn important content objectives (e.g., Armbuster et al., 1977; Jenkins & Pany, 1978; Porter et al., 1978). Porter et al., for instance, developed a basic taxonomy for classifying content included in mathematics curricula and measured whether different standardized mathematics achievement tests covered the same objectives delineated in the taxonomy. Porter continued his research on measuring the content of the enacted curriculum during the advent of standards-based reform (e.g., Gamoran et al., 1997; Porter, Kirst, Osthoff, Smithson, & Schneider, 1993) and developed a survey-based measure that examined the content of instruction

along two dimensions: *topics* and *categories of cognitive demand* (Porter & Smithson, 2001; Porter, 2002). The findings of Gamoran et al. indicated that alignment between instruction and a test of student achievement in high school mathematics accounted for 25% of the variance among teachers. None of the mentioned studies, however, considered students’ opportunity to learn the assessed or intended curriculum and academic achievement in the context of special education.

Porter’s measure, now called the SEC, is presently the only method that can assess alignment among various enacted, intended, and assessed curricula via a content translation of each curriculum into individual content matrices along two dimensions (i.e., topics, cognitive demands). For purposes of the SEC, Porter (2002) developed an AI to determine the content overlap between two matrices at the intersection of topic and cognitive demand. Researchers have utilized this continuous alignment variable as an independent variable in correlational studies predicting student achievement (Smithson & Collares, 2007; Kurz et al., 2010).

Smithson and Collares (2007) used simple correlations, multiple regression, and hierarchical linear modeling to examine the relation between alignment (using the SEC’s AI) and student achievement. The average correlation between alignment (of the enacted to the intended curriculum) and student achievement was 0.34 ($p < 0.01$). Smithson and Collares subsequently used multiple regression analyses to control for the effects of prior achievement, grade level, and socioeconomic status (SES). The results supported alignment (of the enacted to the intended curriculum) as a significant predictor of achievement with adjusted R^2 ranging between 0.41 and 0.70. Smithson and Collares further noted that the results of the multi-level analysis supported alignment as significant predictor of achievement at the classroom level (Level 2) controlling for grade level and SES as well as controlling for prior achievement at the student level (Level 1). Herman and Abedi (2004) conducted analyses similar to Smithson and Collares’s (2007), using their own item-based OTL measure

(i.e., asking students and teachers about the extent to which tested content was covered). As such, the OTL construct related to the content of instruction was aimed at the content overlap between the teacher's instruction on 28 Algebra I content domains and an aligned mathematics assessment. The correlation between student-reported OTL (at the class level) and class achievement was 0.72 ($p < 0.01$), and the correlation between teacher-reported OTL (at the class level) and class achievement was 0.53 ($p < 0.01$). Their multi-level analyses further indicated that the proportion of English language learners in a class and OTL have significant effects on student achievement, even after controlling for students' prior achievement and background.

Research data on the relation between OTL and student achievement in the context of special education are presently very limited. Roach and Elliott (2006) used student grade level, teacher reports of students' curricular access, percentage of academic-focused IEP goals, and time spent in general education settings as predictors of academic performance on a state's alternate assessment. Results indicated the model accounted for 41% of the variance in student achievement. Teacher-reported coverage of general curriculum content was the best predictor in the model accounting for 23% in the variance in student performance. Kurz et al. (2010) used the SEC alignment methodology to examine the relation between OTL (i.e., alignment between the enacted and intended curriculum) and student achievement averages for general and special education teachers. The content of instruction delivered by general and special education teachers as measured by the SEC did not indicate significantly different alignment indices between the two groups. The correlation between OTL and (class averages of) student achievement was 0.64 ($p < 0.05$). When general and special education teachers were examined separately, the correlation between alignment and achievement remained significant only for the special education group with 0.77 ($p < 0.05$). Unfortunately, these findings cannot be generalized due to the study's limited sample size. A multi-level (re)analysis of the Kurz et al.

data via hierarchical linear modeling allowed for variance decomposition of students' end-of-year achievement using predictors at the student level (i.e., prior achievement) and classroom level (i.e., classroom type, classroom alignment). The intra-class correlation coefficient (ICC) was $\hat{\rho} = 0.34$ (i.e., 34% of variance in students' end-of-year achievement was between classrooms). The final (main effects) model predicted individual student achievement as a function of overall mean classroom achievement, main effect for classroom type (i.e., general education, special education), main effect for classroom alignment, prior achievement as a covariate, and random error. All four fixed effects were significant ($p < 0.001$), while the random effects were not significant ($p > 0.05$). The results of the reanalysis thus supported classroom type and classroom alignment as significant predictors of individual student achievement even after controlling for prior achievement at the student level. In addition, both classroom type and classroom alignment accounted for virtually all variance in student achievement that was between classrooms.

In summary, the available data support an empirical association between the content of instruction and student achievement. The quality of the data, however, is limited, which makes it difficult to generalize the findings and develop interventions. First, the measures of students' opportunity to learn instructional content vary across studies. Researchers have repeatedly employed two approaches for collecting OTL data on the content of instruction: (a) item-based OTL measures, which teachers use to report on the relative content coverage related to each test item (e.g., Herman & Abedi, 2004; Husén, 1967; Winfield, 1993); and (b) taxonomic OTL measures that provide an exhaustive list of subject-specific content topics, which teachers use to report on the relative emphases of enacted content according to different dimensions (e.g., Porter, 2002; Rowan & Correnti, 2009). Second, the quality of achievement measures used across studies is unclear. That is, little information is available on the reliability of achievement test scores and the test's alignment to the intended curriculum. The latter concern is about the

extent to which the achievement test in question measured the content that teachers were supposed to teach (i.e., the content prescribed by the standards). That is, alignment between the enacted and intended curriculum cannot be expected to correlate highly with student achievement, if the test fails to be aligned with the respective content standards. In addition, the instructional sensitivity of assessments used to detect the influence of OTL on achievement typically remains an unexamined assumption among researcher (D'Agostino, Welsh, & Corson, 2007; Polikoff, 2010). Another limitation in the presently available data on OTL related to the content of instruction is the paucity of research involving special education teachers and students with disabilities.

Quality of Instruction

The third and most diverse research strand related to an instructional dimension of OTL can be traced back to several models of school learning (e.g., Bloom, 1976; Carroll, 1963; Gagné, 1977; Harnischfeger & Wiley, 1976). Both Carroll's model of school learning and Walberg's (1980) model of educational productivity, for example, featured quality of instruction alongside quantity of instruction. The operationalization of instructional quality for purposes of measurement, however, resulted in a much larger set of independent variables related to student achievement than instructional time. In his research synthesis on teaching, Walberg (1986) reviewed 91 studies that examined the effect of quality indicators on student achievement, such as frequency of praise statements, corrective feedback, classroom climate, and instructional groupings. Walberg reported the highest mean effect sizes for (positive) reinforcement and corrective feedback with 1.17 and 0.97, respectively. Brophy and Good's (1986) seminal review of the process-product literature identified aspects of giving information (e.g., pacing), questioning students (e.g., cognitive level), and providing feedback as important instructional quality variables with consistent empirical support. Additional meta-analyses focusing on specific subjects and student

subgroups are also available (e.g., Gersten et al., 2009; Vaughn, Gersten, & Chard, 2000). Gersten et al. (2009), for example, examined various instructional practices that enhanced the mathematics proficiency of students with learning disabilities. Gersten and colleagues hereby identified two instructional practices that provided practically and statistically important increases in effect size: teaching students the use of heuristics (i.e., general problem-solving strategy) and explicit instruction.

OTL research related to the quality of instruction also has considered teacher expectations for the enacted curriculum (i.e., cognitive demands) and instructional resources such as access to textbooks, calculators, and computers (e.g., Boscardin, Aguirre-Muñoz, Chinen, Leon, & Shin, 2004; Herman et al., 2000; Porter, 2002; Wang, 1998). Wang provided one of the first multi-level OTL studies that examined the quality of instruction alongside three other content variables (i.e., coverage, exposure, and emphasis). Wang's findings supported students' attendance rate, content coverage, content exposure, and quality of instruction as significant predictors of student achievement (controlling for ability, gender, and race). Wang further noted that content exposure (i.e., indicator of time on instruction via periods allocated to instruction) was the most significant predictor of written test scores, while quality of instruction (i.e., lesson plan completion, equipment use, textbook availability, material adequacy) was the most significant predictor of hands-on test scores. Although Wang considered the multi-dimensional nature of OTL, she did not include time on instruction and used an unconventional measure of content coverage, namely the teachers' predicted pass rate for students on each test item. The latter measure of instructional content, however, is difficult to interpret without knowing the extent to which the test covered the teachers' enacted curriculum. Moreover, questions that ask teachers to predict students' pass rates on items are likely to be confounded by their estimates of student ability. This caveat notwithstanding, Wang demonstrated that quality of instruction can serve as a significant predictor of student test scores even with other

key OTL variables in the model. The empirical relation between quality of instruction and student achievement, however, is mostly based on the reports of general education teachers and the academic achievement of students without disabilities.

In summary, many researchers interested in OTL have started to consider the dimension of instructional quality. Herman et al. (2000) identified two broad categories of interest in this instructional dimension related to (a) *instructional resources* such as equipment use and (b) *instructional practices* such as working in small groups. However, as discussed earlier, numerous other indicators of quality associated with student achievement are found in the literature including teacher expectations for student learning, progress monitoring, and corrective feedback (e.g., Brophy & Good, 1986, Porter, 2002). The wide range of instructional quality variables available underscores the importance for researchers to provide a theoretical and empirical rationale for their particular operationalization of instructional quality. Lastly, it should be acknowledged that all quality indicators used in OTL research based on teacher self-report are limited to information about frequency or emphasis. While such information is clearly useful, it cannot indicate the extent to which certain practices were implemented correctly.

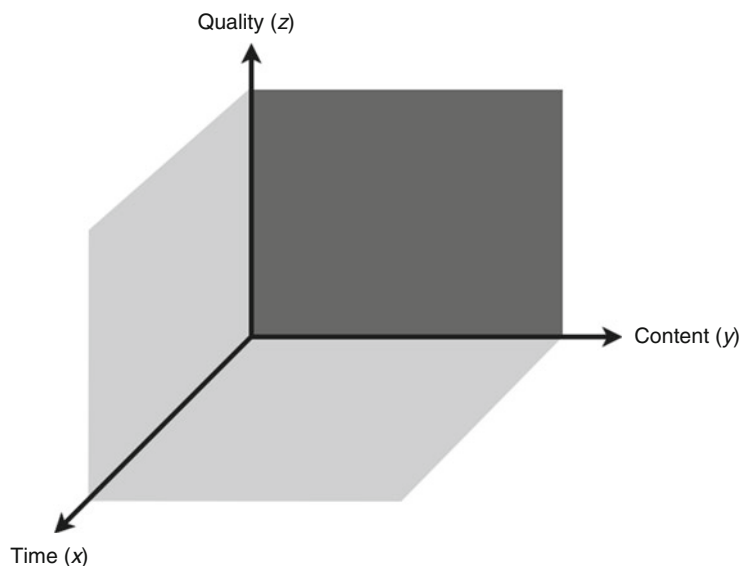
The Unfolding of Instruction

So far, we have redefined the concept of OTL in the context of the current test-based accountability system as *students' opportunity to learn the intended curriculum* and considered the respective implications for both students with and without disabilities. We further posited the enacted curriculum as students' main access point to the intended curriculum. The subsequent literature review was used to highlight three broad research strands related to OTL that focused on different instructional dimensions of the enacted curriculum: time on instruction, content of instruction, and quality of instruction. Anderson first articulated a merger of the various

OTL conceptualizations in 1986: "A single conceptualization of opportunity to learn coupled with the inclusion of the variable[s] in classroom instructional research . . . could have a profound effect on our understanding of life in classrooms" (p. 3686). In 1993, Stevens and Grymes established the first unified conceptual framework of OTL to investigate "students' access to the core curriculum" along four dimensions: *content coverage* (i.e., content of the general and assessed curriculum), *content exposure* (i.e., time spent on the general curriculum), *content emphasis* (i.e., skill emphasis, cognitive process expectations, instructional groupings), and *quality of instructional delivery* (i.e., instructional strategies). Unfortunately, Stevens and Grymes did not develop an empirical program of research on the basis of this framework. Nonetheless, Stevens (1996) identified a critical gap in the OTL literature, namely that most "opportunity to learn studies looked at [only] one variable at a time" (p. 4).

The scarcity of research studies used to investigate all three instructional dimensions of OTL is unfortunate because neither aspect of OTL can occur in isolation for all practical intents and purposes. That is, instructional content enacted by a teacher always unfolds along (at least) two additional dimensions: time and quality. For example, a teacher's instruction is not adequately captured by referring solely to the content of instruction such as *solving algebraic equations*. In actuality, the teacher may have asked students to *apply* strategies related to *solving algebraic equations* for *15 min* in *small groups* while providing *guided feedback*. The different sets of italicized words refer to various aspects of OTL – time, content, and quality of instruction – that have to occur in conjunction with one another whenever instruction is enacted by a teacher. Figure 6.3 depicts a model of the three instructional dimensions of OTL. The *x*-axis represents *time on instruction* (i.e., the amount of time spent on teaching the intended curriculum); the *y*-axis represents the *content of instruction* (i.e., the extent to which the content of instruction is aligned with the intended curriculum); and the *z*-axis represents the *quality of instruction*

Fig. 6.3 The instructional dimensions of OTL



(i.e., the extent to which a teacher’s instructional practices are evidence based). At the enacted curriculum level, almost all instruction should unfold along all three dimensions. Instruction restricted to the *time–content plane* would indicate that the teacher delivered content of the intended curriculum without relying on any evidence-based instructional practices. Instruction restricted to the *time–quality plane* would indicate that the teacher used evidence-based instructional practices to deliver content, yet the enacted content failed to be aligned with the student’s intended curriculum. Lastly, the *content–quality plane* would only occur at the planned curriculum level. Although the suggested continuums for the three different dimensions of OTL are consistent with the aforementioned literature, the model is, at least for now, intended to serve illustrative purposes. That is, the model can be used as a framework to categorize measurement options related to each instructional dimension of OTL.

Measurement of OTL

The measurement of OTL at the enacted curriculum level has historically relied on three basic methods: direct observation, teacher report,

and document analysis. That is, the instructional dimensions of OTL related to time, content, and quality can be operationalized and subsequently documented using (a) *observers* who conduct classroom observations or code videotaped lessons, (b) *teachers* who self-report on their classroom instruction via annual surveys or daily logs, or (c) *experts* who are trained to review classroom documents such as textbooks, assessments, and other student products. Third-party observations and teacher surveys are by far the most frequently used methods, each with a unique set of advantages and challenges (Rowan & Correnti, 2009).

Third-party observations are often considered the “gold standard” for classroom research, but the high costs associated with this method limit its large-scale application outside well-funded studies for purposes of documenting OTL. Moreover, the complexity and variability of classroom instruction across the school year (Jackson, 1990; Rogosa et al., 1984) raise the question of *generalizability*: How many observations are necessary to generalize to a teacher’s entire enacted curriculum? Annual surveys, on the other hand, are relatively inexpensive but rely exclusively on teacher *memory* for the accurate recall of the enacted curriculum. To address these measurement challenges, Rowan and colleagues

suggested a third alternative, namely the use of frequently administered teacher logs (see Rowan et al., 2004). Teacher logs are intended to (a) reduce a teacher's response burden by focusing on a discreet set of behaviors, (b) increase accuracy of teacher recall by focusing on a recent time period (e.g., today's lesson), and (c) increase generalizability through frequent (cost-effective) administrations across the school year.

As part of their *Reform Up Close* study, Porter et al. (1993) used a variety of methods to collect data on teachers' enacted curriculum including daily logs, weekly surveys, classroom observations, and questionnaires. The agreement between classroom observations and teacher log data (calculated on each observation pair and averaged over all pairs) along four dimensions – broad content area (A), subskills within broad content area (AB), delivery of content (C), and cognitive demand (D) – was 0.78, 0.68, 0.67, and 0.59, respectively. Porter et al. also noted significant correlations between log data and questionnaire data on dimension (A) of 0.50 to 0.93 in mathematics and of 0.61 to 0.88 in science (as cited in Smithson & Porter, 1994). In 2002, Porter argued that a number of studies investigating the validity of survey data have confirmed that “survey data is excellent for describing quantity – for example, what content is taught and for how long – but not as good for describing quality – for example, how well particular content is taught” (p. 9). For purposes of validating teacher logs, Camburn and Barnes (2004) discussed the challenges related to reaching (inter-rater) agreement as one of their validation strategies including rater background, type of instructional content, level of detail (e.g., subskills) associated with content, and frequency of occurrence. On the basis of their statistical results, Camburn and Barnes expressed confidence in teacher logs to measure instruction at grosser levels of detail and for activities that occurred more frequently. Rowan and Correnti (2009) concluded that teacher logs are (a) “far more trustworthy” than annual surveys to determine the frequency with which particular content and instructional practices are enacted and (b) yield “nearly equivalent” data to what would be gathered via

trained observers. That being said, classroom observations are presently unrivaled in determining aspects of child-instruction or teacher-child interactions (e.g., Connor, Morrison, et al., 2009; Pianta & Hamre, 2009).

The measurement of the enacted curriculum has attracted much research attention in recent years, as evidenced by two special issues dedicated to “opening up the black box” of classroom instruction: the September 2004 issue of the *Elementary School Journal* and the March 2009 issue of *Educational Researcher*. A thorough review of available measurement tools, however, is beyond the scope of this chapter. Instead, I situate and discuss four guiding questions originally posed by Douglas (2009) for purposes of measuring classroom instruction in the context of OTL: What should we measure in classroom instruction? How can we best analyze data on classroom instruction? At what level should we measure classroom instruction? What tools can we use to measure classroom instruction?

The first question challenges researchers to provide a (theoretical and/or empirical) framework for selecting measurement variables of interest and for understanding their relation to the overall construct in question. With respect to OTL, the argument presented in this chapter suggests three instructional dimensions at the enacted curriculum level for purposes of documenting students' opportunity to learn the intended curriculum. The ICM framework, a review of three distinct research strands related to OTL at the enacted curriculum level, and a subsequent instructional dimensions model provided the theoretical and empirical underpinnings for this argument. The general answer to “what” should be measured for purposes documenting OTL is thus: time on instruction, content of instruction, and quality of instruction. Depending on the specific framework, research context, and questions asked, stakeholders may define these instructional dimensions differently. The model introduced in this chapter, for example, posited three different continua: *time on instruction* as the amount of time spent on teaching the intended curriculum; *content of instruction* as the extent to which the content of instruction is aligned

with the intended curriculum; and *quality of instruction* as the extent to which a teacher's instructional practices are evidence based.

The second question points to the nesting of classroom instruction and the importance of variance decomposition models in evaluating the effects of classroom instruction on student achievement. Scheerens and Bosker's (1997) review of the literature indicated that variance in student achievement status (without controlling for prior achievement and SES) can be decomposed as follows: About 15–20% of the variance lies among *schools*; another 15–20% of the variance lies among *classrooms* within schools; and about 60–70% of the variance lies among *students* within classroom within schools. Scheerens and Bosker, however, used an unconditional model (i.e., no independent variables were used to predict student achievement). For their analyses of achievement data, Rowan, Correnti, and Miller (2002) also used a three-level hierarchical linear model but included covariates at each level (i.e., prior achievement, home and social background, social composition of schools). Their results indicated that about 4–16% of the variance in students' reading achievement and about 8–18% of students' mathematics achievement lie among *classrooms* (depending on grade level). Although these studies support the methodological appropriateness of using multi-level models in the measurement of OTL, which is ultimately a teacher effect, several analysts have challenged the adequacy of covariate adjustment models to model changes in student achievement (Rogosa, 1995; Stoolmiller & Bank, 1995). The evaluation of teacher effects on students' academic *growth* via gain score as the outcome variable, however, has its own set of unique challenges especially when differences among students on academic growth are rather small (see Rowan et al., 2002). Nonetheless, researchers can select from many options within multi-level modeling that can account for the unique nesting of the enacted curriculum and its relation to student achievement. A cross-classified random effects model (Raudenbush & Bryk, 2002), for example, can account for a situation (common in special education) in which lower-level units are

cross-classified by two or more higher-level units (e.g., a students' sources of OTL can come from different teachers nested within different classrooms). In short, multi-level analysis is an invaluable tool for evaluating the effects of OTL on student achievement by portioning true variance from error variance and for modeling interactions across time, students, classrooms, and schools (Douglas, 2009).

The third question is also related to the nested nature of OTL and asks researchers to consider how to locate and sample for OTL. One of the first challenges is to decide the number of measurement points for purposes of documenting OTL at the enacted curriculum level. Rowan and Correnti (2009), who used daily teacher logs to measure different aspects of a teacher's enacted curriculum, decomposed variance in time spent on reading/language arts instruction into three levels: time on instruction on a given *day* (Level 1), days nested within *teachers* (Level 2), and teachers nested within *schools* (Level 3). Their results on the basis of about 2,000 teachers, who logged approximately 75,000 days, indicated that approximately 72% of the variance in instructional time lies among *days*, about 23% lies among *teachers* within schools, and about 5% lies among *schools*. In other words, time on instruction can vary significantly from day to day: "the average teacher in the [study] provided students with about 80 min of reading/language arts instruction per day, but the standard deviation of instructional time across days for a given teacher was 45 min, with 15% of all days including 0 min of reading/language arts instruction!" (Rowan & Correnti, 2009, p. 123). This wide variability of classroom instruction around key instructional dimensions of OTL seems to suggest a fairly large number of measurement points for purposes of reliably discriminating among teachers. Rowan and Correnti suggested that about 20 logs per year are optimal (with diminishing returns thereafter), if the measurement goal is to reliably discriminate among teachers.





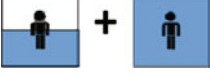



In addition to day-to-day variability, Connor, Morrison, et al. (2009), who used an observational measure, reported that different students nested within the same class may be experiencing

different amounts and types of instruction. This issue points to the appropriate measurement level of OTL: Should it be documented at the student level or the classroom level? Most of the teacher-report measures mentioned in the previous literature review were used to collect information on the enacted curriculum at the class level. Given empirical evidence of significant variation along key instructional dimensions of OTL for students *within* the same class and the theoretical model of the ICM, measurement of OTL restricted to the classroom level does not appear to be sufficient. That is, data on classroom-level OTL cannot be generalized to individual students and, in the case of students with disabilities, cannot yield information on the extent to which students had the opportunity to learn their specific intended curriculum (which presumably varies from student to student).

Croninger and Valli (2009) identified additional challenges related to the variability of instruction, namely the sources and boundaries of (reading) instruction. Results from their five-year longitudinal study of teaching in schools

of poverty indicated that only one-third of students experienced no shared instruction. That is, the majority of students received reading instruction from multiple sources in one or more locations. Croninger and Valli noted that “the most prevalent form experienced by students was simultaneous instruction involving an instructional assistant (30%), student teacher (17%), staff developer/resource teachers (15%), and/or in-class help assigned specifically to them (8%)” (p. 105). Moreover, nearly 20% of students received additional reading instruction outside classrooms. Croninger and Valli further noted that many students experienced more reading instruction outside their scheduled reading class than during their scheduled lesson. These findings underscore an important measurement challenge, namely to account for all sources of instruction that contribute to a student’s opportunity to learn his or her intended curriculum. This issue is particularly relevant for students with disabilities who are likely to share multiple sources of (reading or mathematics) instruction such as a special education teacher, a general education

Table 6.1 Taxonomy of instructional sources and scenarios for students with disabilities

Source of instruction	Instructional scenario	
 GenED	Target student receives instruction <i>exclusively</i> from a GenED teacher. ■ <i>Full inclusion</i>	
 SPED	Target student receives instruction <i>exclusively</i> from a SPED teacher. ■ <i>Pull-out, resource</i>	
 GenED/SPED	Target student receives instruction from a GenED <i>and</i> SPED teacher. ■ <i>Co-teaching, collaboration</i>	
 GenED + SPED	Target student receives instruction <i>separately</i> from a GenED teacher <i>and</i> a SPED teacher. ■ <i>Full inclusion plus pull-out/resource</i>	
 GenED/SPED + SPED	Target student receives instruction from a GenED <i>and</i> SPED teacher <i>and additionally</i> from a SPED teacher. ■ <i>Co-teaching/collaboration plus pull-out/resource</i>	
 Instruction by general education teacher (GenED)	 Instruction by special education teacher (SPED)	 Target student teacher (SPED)

teacher, a teaching assistant, and/or related services providers. Table 6.1 provides a taxonomy that illustrates several instructional sources and scenarios for students with disabilities. Clearly, measurement of a student's opportunity to learn the intended curriculum is unlikely to be adequately captured via one teacher's instruction during one class period at the class level. With all the aforementioned measurement challenges in mind, we are now ready to explore some measurement options for the three instructional dimensions of OTL.

Options for Measurement

A selection of available measurement tools with information on their stated purpose, general method, and emphasized instructional dimension(s) of OTL are highlighted in Table 6.2. Additional information includes theoretical grounding, instructional target (i.e., class level, student level), psychometric support, and cost (i.e., high, medium, or low investment of time and funds). This table is intended to assist readers in their initial search for potential measures of OTL along several key criteria. The description of each measure is anchored in one main literature source; readers are therefore advised to consult the entire body of literature related to each measure for purposes of making a final selection. Each measure will be discussed briefly.

The *Individualizing Student Instruction Observation and Coding System* (ISI; Connor, Morrison, et al., 2009) is an observational measure designed to capture literacy instruction (K-3) and the extent to which teachers are matching their instruction to students' assessed skill levels. This observational measure is connected to a larger program of intervention research focused on Child x Instruction interactions and the effects of timing of instructional activities (e.g., Connor, Piasta, et al., 2009). Connor and colleagues have developed a comprehensive model of the learning environment that incorporates student characteristics and skills, multiple dimensions of the classroom environment (i.e., teacher

characteristics), and multiple dimensions of instruction (see Connor, Morrison, et al., 2009). As indicated in the table, Connor et al. identified four, simultaneously operating, dimensions of instruction relevant to Child x Instruction interactions: (a) *management* of students' attention, (b) *context* of the activity, (c) *content* of the activity, and (d) *duration* of the activity. Within the context of OTL, it is important to note that the measure is restricted to literacy content at specific grade levels (K-3). Although the content codes are presumably expandable, they are currently not set up to reflect student-specific intended curricula (including state or district-specific general curricula). The ISI is targeted at the individual student level, yet context codes can account for instruction at the classroom level.

The *Classroom Assessment Scoring System* (CLASS; Pianta & Hamre, 2009) is a standardized observation system to assess global classroom quality via indicators in three broad domains: (a) *emotional supports* (e.g., affect, negativity, responsiveness, flexibility); (b) *classroom organization* (e.g., clear expectations, maximized time use, variety); and (c) *instructional supports* (e.g., creativity, feedback loops, conversation). The CLASS is focused on teacher-child interactions at the indicator level, which are operationalized for purposes of rating-specific behaviors and interaction patterns within a 30-min time frame. Although the CLASS can only provide information on the quality dimension of OTL, it does so across a wide grade spectrum and without being restricted to a specific subject. However, the CLASS cannot be used to discern differences in instructional quality at the individual student level.

The *Study of Instructional Improvement* (SII) logs (Rowan et al., 2004) are self-report measures that require teachers to report on the time, content, and quality of their instruction for a particular student and day. The logs are available for the content areas of language arts and mathematics at the elementary school level. For the *time* dimension, teachers are asked to report on the number of minutes a particular student has received instruction including in another classroom or by another teacher. For the

Table 6.2 Measurement options for the instructional dimensions of OTL

Measure	Purpose	Method	Time	Content	Quality	Theoretical grounding	Instr. target	Reliability/validity evidence	Cost
Individualizing Student Instruction (ISI) observation and coding system (e.g., Connor et al., 2009)	<ul style="list-style-type: none"> ■ To capture literacy instruction (K-3) ■ To examine whether teachers are matching the content of the literacy instruction to students' assessed skill levels 	<ul style="list-style-type: none"> ■ Direct observation to complete field notes ■ Videotaping to code instruction ■ 4–8 full-day observations /tapings 	<ul style="list-style-type: none"> ■ Duration (sec) 	<ul style="list-style-type: none"> ■ Content (i.e., code focused^a, meaning focused^b) 	<ul style="list-style-type: none"> ■ Management (i.e., teacher/child managed, peer/child managed) ■ Context (i.e., whole class, small group/pair, individual) 	<ul style="list-style-type: none"> ■ Classroom environment model^c ■ Ecological model^d 	<ul style="list-style-type: none"> ■ Student 	<ul style="list-style-type: none"> ■ Interrater reliability (e.g., Kappa) ■ Predictive validity (e.g., literacy skill gains) 	<ul style="list-style-type: none"> High (i.e., trained observers, two video cameras, time for coding)
Classroom Assessment Scoring System (CLASS; e.g., Pianta, La Paro, & Hamre, 2008)	<ul style="list-style-type: none"> ■ To measure global classroom quality (pre-K-12) 	<ul style="list-style-type: none"> ■ Direct observation (i.e., time-sampled codes, global ratings) ■ Approx. 6 observations (30 min each) 	<ul style="list-style-type: none"> N/A 	<ul style="list-style-type: none"> N/A 	<ul style="list-style-type: none"> ■ Emotional supports (i.e., positive classroom climate, teacher sensitivity, regard for student perspectives) ■ Classroom organization (i.e., effective behavior management, productivity, instructional learning formats) ■ Instructional supports (i.e., concept development, quality of feedback, language modeling) 	<ul style="list-style-type: none"> ■ Literature on classroom teaching and educational effectiveness 	<ul style="list-style-type: none"> ■ Class 	<ul style="list-style-type: none"> ■ Interrater reliability ■ Construct validity (e.g., factor analysis) ■ Criterion validity ■ Predictive validity (e.g., achievement gains) 	<ul style="list-style-type: none"> High (i.e., trained observers)

Table 6.2 (continued)

Measure	Purpose	Method	Time	Content	Quality	Theoretical grounding	Instr. target	Reliability/validity evidence	Cost
Study of Instructional Improvement (SI) logs (e.g., Rowan, Camburn, & Correnti, 2004)	<ul style="list-style-type: none"> To measure frequencies of instructional activities (1–5) To evaluate school reform efforts 	<ul style="list-style-type: none"> Teacher log (about 100 items for approx. 45 min) Approx. 20 logs 	<ul style="list-style-type: none"> Duration (min) 	<ul style="list-style-type: none"> Broad content strands with subskills/activities 	<ul style="list-style-type: none"> Cognitive demand Materials, student responses 	OTL literature	<ul style="list-style-type: none"> Student 	<ul style="list-style-type: none"> Interrater reliability Construct validity Predictive validity (e.g., achievement gains) 	Low (i.e., training, teacher incentives, scoring)
Surveys of the Enacted Curriculum (SEC; Porter, 2002)	<ul style="list-style-type: none"> To measure topics and cognitive demand of instruction To evaluate alignment between enacted, intended, and assessed curricula 	<ul style="list-style-type: none"> Teacher survey End-of-year survey 90–120 min 	N/A	<ul style="list-style-type: none"> Broad content strands with subskills 	<ul style="list-style-type: none"> Cognitive demand Instructional practices survey (e.g., homework, instructional activities, technology) 	OTL literature	<ul style="list-style-type: none"> Class 	<ul style="list-style-type: none"> Interrater reliability Construct validity Predictive validity (e.g., achievement gains) 	Low (i.e., training, teacher incentives)
My Instructional Learning Opportunities System (MyILOGS; Kurz, Elliott, & Shrago, 2009)	<ul style="list-style-type: none"> To measure opportunity to learn the intended curriculum To evaluate alignment between planned, enacted, intended, and assessed curricula To evaluate differentiation instruction 	<ul style="list-style-type: none"> Teacher log Daily calendar (2–5 min) Approx. 20 instruction details (2–5 min) 	<ul style="list-style-type: none"> Duration (min) 	<ul style="list-style-type: none"> Broad content strands with subskills Custom curricula 	<ul style="list-style-type: none"> Cognitive demand Instructional practices Context Engagement Goal attainment 	<ul style="list-style-type: none"> OTL literature Intended curriculum model 	<ul style="list-style-type: none"> Student Class 	<ul style="list-style-type: none"> N/A (<i>In planning</i>) 	Low (i.e., training)

^aCode-related literacy instruction (e.g., phonological awareness, phonics, letter and word fluency)

^bActive extraction and construction of meaning from text (e.g., oral language, vocabulary, comprehension, text fluency, writing)

^cThe classroom environment model purports that students' reading outcomes are multiply determined by student characteristics (e.g., language, self-regulation, home support), classroom characteristics (e.g., warmth/responsiveness, teacher's content area knowledge), instructional characteristics (i.e., duration, content, management, context)

^dThe ecological model purports that children's development is affected by both proximal (e.g., teacher) and distal sources of influence (e.g., community) and that these influences and their associations with development may change over time (Morrison, Bachman, & Connor, 2005)

content dimension, teachers are asked to indicate their content foci along several broad content topics (i.e., major focus, minor focus, touched on briefly, not taught). The *quality* dimension is addressed along specific enacted language arts (i.e., comprehension, writing, word analysis) and mathematics topics (i.e., number concepts, operations, patterns/functions/algebra) via questions related to cognitive demands, materials, and student responses. With regard to theory, the development of the SII logs is grounded in the aforementioned OTL literature, which explains their good match to the desired OTL dimension. It should be noted, however, that the content dimension of the SII logs is not reflective of individual intended curricula. In fact, their content selection is more appropriately described as a core content selection within elementary school language arts and mathematics (see Rowan & Correnti, 2009). Lastly, the cost of administering SII logs is relatively low. Rowan and Correnti mentioned a less than \$30 per log cost (not including training).

The *Surveys of the Enacted Curriculum* (SEC; Porter, 2002) are annual teacher surveys designed to provide information on the alignment between intended, enacted, and assessed curricula. The SEC method relies on content translations by teachers (for purposes of the enacted curriculum) or curriculum experts (for purposes of the intended and assessed curriculum) who code a particular curriculum (i.e., classroom instruction, state test, or state standards) into a content framework that features a comprehensive list of topics in mathematics, English/language arts, and science (K-12). The topic list is exhaustive to accommodate different levels of content specificity. For the enacted curriculum, teachers indicate content coverage and cognitive demand for approximately 200 topics. Content coverage is indicated via relative time emphases (i.e., none, slight, moderate, sustained). Time on instruction is thus not assessed directly via the SEC. To gather additional information on instructional quality (besides cognitive demand), teachers are required to complete an additional survey on instructional practices (i.e., activities, homework, technology use, assessments, teacher opinions, and other characteristics). The SEC is available

online (www.seconline.org), which allows teachers who complete the survey online to review graphical representations of their content coverage. In addition, the SEC website can calculate content overlap between two curricula via an AI (see Porter, 2002). The SEC can thus provide information on the content and quality dimensions of OTL. This information, however, is limited to the classroom level. The graphical reports further allow teachers to review their self-reported content coverage, which holds potential for formative feedback. Figure 6.4 displays the online SEC's graphical output of two content matrices taken from the Kurz et al. (2010) study.

The content map on the left represents a teacher's enacted curriculum, and the content map to the right represents a state's eight-grade general curriculum in mathematics. The AI between the two content maps is 0.05 (see Kurz et al. for details). A review of these content maps can allow teachers to pinpoint differences between emphasized content topics and cognitive demands of the state's general curriculum and their own enacted curriculum. This teacher, for example, emphasized the topics of *Number Sense* and *Basic Algebra*. However, these topics were not taught at the intended category of cognitive demand (i.e., demonstrate understanding). Moreover, the teacher did not place the intended instructional emphases on the topics of *Measurement*, *Geometric Concepts*, and *Data Displays*. Although these charts are capable of delivering formative feedback, the SEC is typically completed at the end of the school year, which limits its formative benefit to teachers.

My Instructional Learning Opportunities Guidance System (MyiLOGS; Kurz, Elliott, & Shrago, 2009) is an online teacher tool (www.myilogs.com) designed to assist teachers with the planning and implementation of intended curricula at the class and individual student level. The development of the measure was grounded in the literature and models discussed in this chapter and thus can be used to document all three instructional dimension of OTL. The tool features state-specific general curricula for various subjects and additional customizable skills that allow special education teachers and other

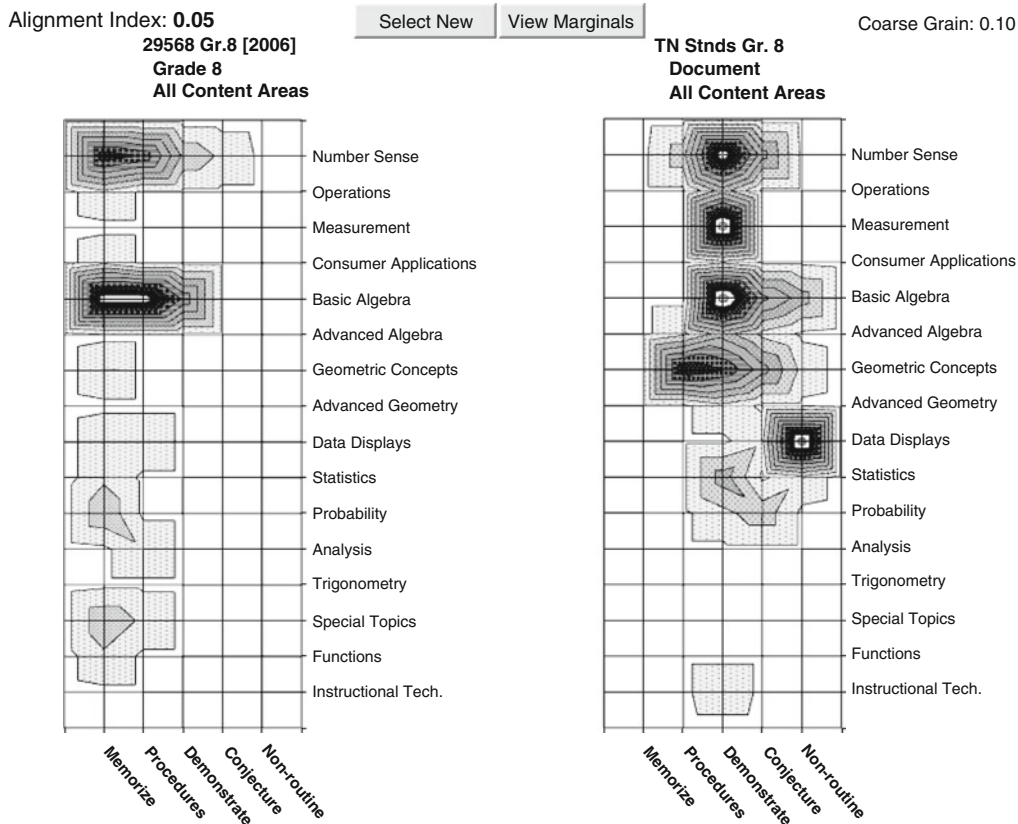


Fig. 6.4 Content map example of an enacted and general curriculum matrix from the SEC

related services providers to add student-specific objectives. The tool therefore permits teachers to document the extent to which their classroom instruction covers individualized intended curricula. To this end, MyiLOGS provides teachers with an instructional calendar that features an expandable sidebar, which lists the skills that comprise the intended curriculum. Teachers then drag and drop planned skills onto the respective calendar days and indicate the number of minutes attached to each skill. After the lesson, teachers confirm enacted skills and respective times at the class level. On a user-specified sample of days, teachers are further asked to report on additional time emphases (in minutes) related to the (planned and enacted) skills listed on the calendar including cognitive demand, instructional groupings, use of evidence-based instructional practices (e.g., use of heuristics, explicit instruction), engagement, and goal attainment. This

detailed reporting occurs at the classroom level *and* student level. Teachers can further review a range of graphic reports and tables that provide detailed information on their enacted curriculum and its relation to the intended curriculum. Reports are available for the entire class and individual students. Psychometric evidence, such as predictive and criterion-related validity, is currently being collected as part of a three-state field test. Figure 6.5 provides an example of a time allocation report. This time allocation report is considered cumulative because it is based on the total amount of minutes dedicated to instruction as reported by the teacher on a daily basis.

The top graphic in Fig. 6.5 displays a teacher's time allocation across all five mathematics content strands prescribed by the state of Pennsylvania. In addition, the top graphic displays the amount of instructional time dedicated

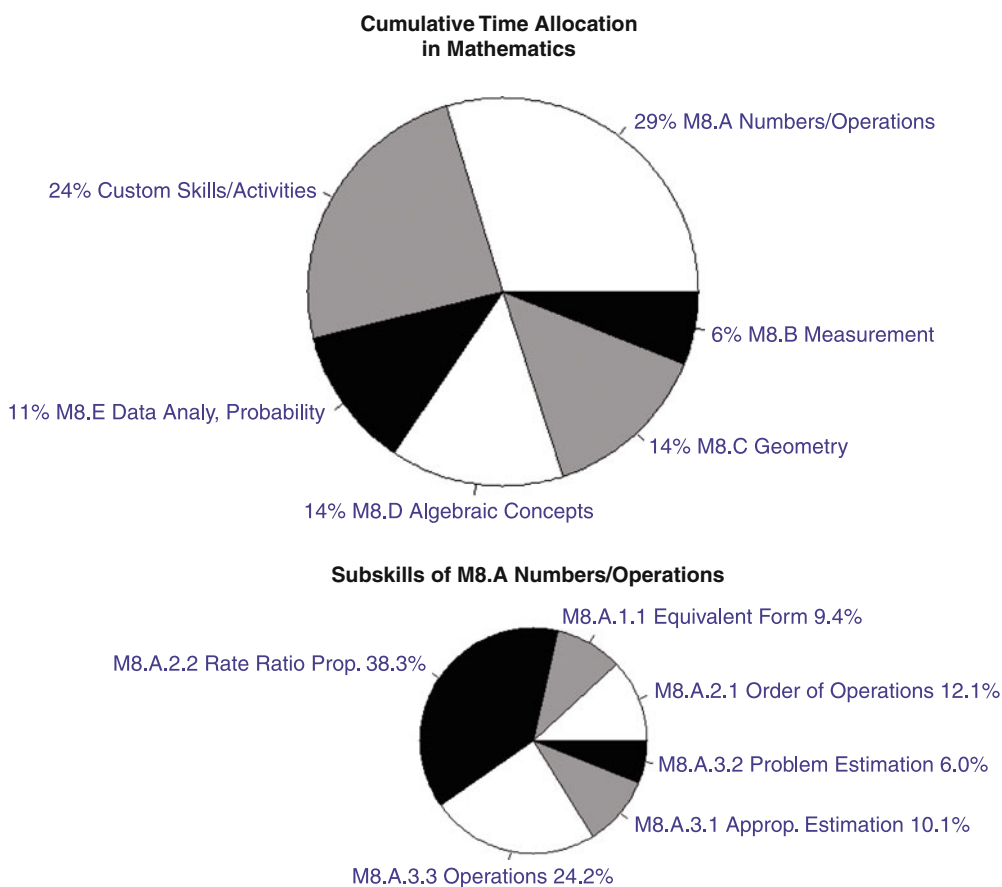


Fig. 6.5 Example of a time allocation report from MyiLOGS

to *Custom Skills/Activities*, which allows teachers to document skills related to a student's IEP curriculum as well as skills or activities that fall outside the state's general curriculum for mathematics. This teacher, for example, dedicated over half of his or her instructional time to *Numbers/Operations* and *Custom Skills/Activities*. The bottom graphic of Fig. 6.5 displays the breakdown for the *Numbers/Operations* strand into its respective subskills. An example of a cognitive process dimensions report is shown in Fig. 6.6.

This report is based a sample of 77 days on which the teacher decided to report on instructional details at class and student level. The top and bottom graphics allow a teacher to review (a) his or her time emphases on different cognitive process dimensions at the class level and

(b) the extent to which differentiated instruction took place for two target students. In this example, the teacher's expectations largely focused on the *Understand/Apply* dimension at the class level. Kayla received very comparable instruction. James, on the other hand, received a much larger emphasis on the *Remember* dimension. In addition, James experienced much less time on instruction (i.e., about 25% were *not* available for instruction) than the rest of the class. Together with additional tables and reports, MyiLOGS is intended to provide teachers with formative feedback that can enhance instructional focus, alignment, and access to the general curriculum.

The present selection of measurement tools related to the three instructional dimensions of OTL at the enacted curriculum level indicates a diverse set of options, each method with

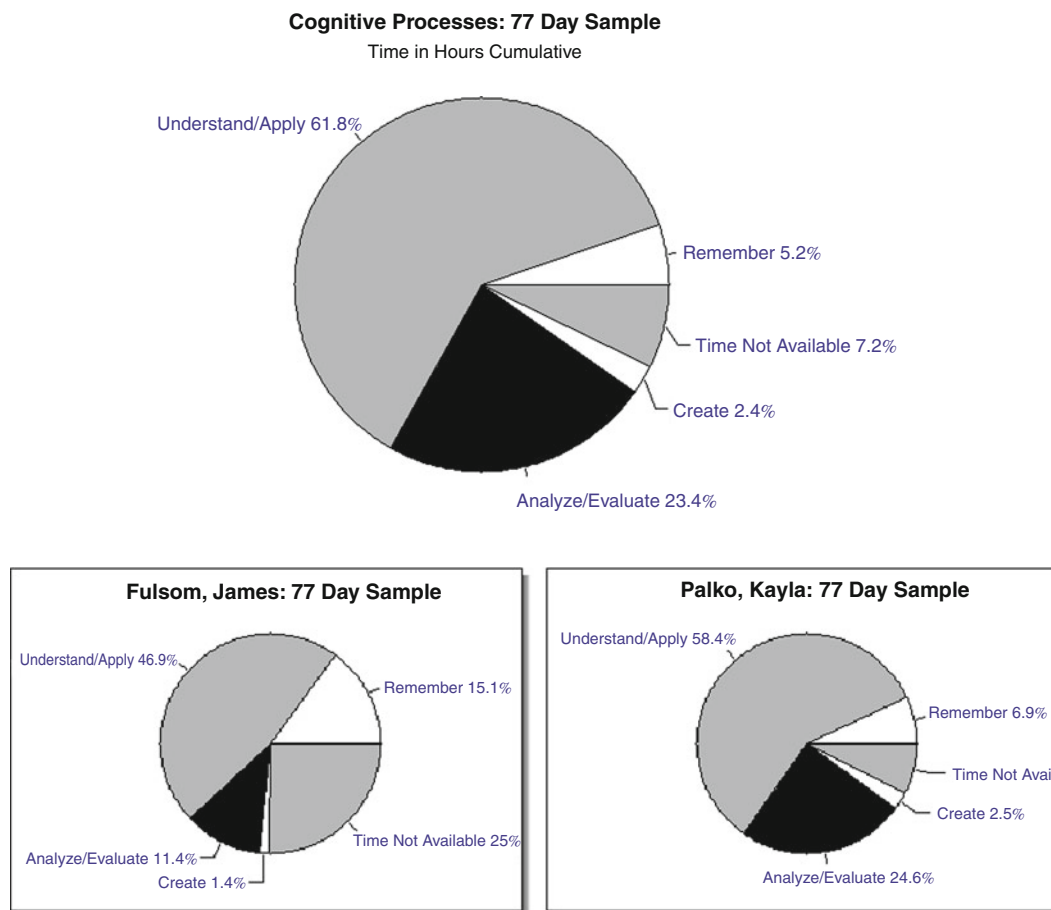


Fig. 6.6 Example of cognitive process dimensions report from MyiLOGS

its unique benefits and drawbacks. The relative weights of issues such as observer reliability, reactivity, teacher memory, and cost can only be assigned after the proper research context and questions have been established. However, even after those issues are settled, it is important to remember that “all methods are to some extent limited in scope with regard to measuring the multifaceted nature of the classroom” (Pianta & Hamre, 2009).

The Future of OTL

In 1995, Schmidt and McKnight declared that “there *is* a story to be told . . . a story about children and also about curricula – curricula

transforming national visions and aims into intentions that shape children’s opportunities for learning through schooling” (p. 346). This chapter provided a new framework that delineated these curricula and their mediated relations to the intended curriculum. The concept of OTL was redefined within that framework, which posited that the transformative powers of OTL are the greatest at the enacted curriculum level where “intentions” become measurable teacher actions. The measurement of students’ opportunity to learn the intended curriculum via three instructional dimensions of the enacted curriculum, however, cannot be the sole direction for future research and work on OTL. Although the conceptual and substantive relevance of OTL may be sufficient to sustain the concept and its measurement for some

time, the prospects of OTL lie in the usefulness of OTL data to assist stakeholders in developing interventions and making meaningful changes in instruction that increase learning opportunities for all students.

Teacher actions at the enacted curriculum level determine the extent to which students have the opportunity to learn the intended curriculum. Empirical evidence indicates that teacher actions along key instructional dimensions – time on instruction, content of instruction, and quality of instruction – impact student outcomes. Documenting these actions thus should not be an end in itself. Rather than “admiring the problem,” measurement of OTL should function as *a means to an end*, namely to use the gathered information as feedback to modify teaching and learning activities with the ultimate goal of increasing student outcomes. In short, the formative benefits of measuring OTL represent an exciting and promising direction for future research and work on OTL.

Shepard, Hammerness, Darling-Hammond, and Rust (2005) argued that an assessment becomes formative when “[it] is carried out during the instructional process for the purpose of improving teaching and learning” (p. 275). The extent to which measurement of OTL at the enacted curriculum can be integrated into the instructional process and improve teaching and learning remains an open empirical question. Some researchers have already begun to answer this question by using measures highlighted in the previous section in formative ways. Connor and colleagues, for example, have developed a web-based software tool that incorporates algorithms, which prescribe amount and types of instruction based on previously observed interactions via the ISI (e.g., Connor, Morrison, & Katch, 2004; Connor, Morrison, et al., 2009; Connor, Morrison, & Underwood, 2007). Porter and colleagues have used the graphical reports of the SEC for purposes of professional development, which resulted in greater alignment (between the enacted and intended curriculum) for the treatment group with an effect size of 0.36 (Porter, Smithson, Blank, & Zeidner, 2007). Lastly, Kurz and colleagues have developed an OTL

measurement tool that is to be used as an ongoing part of the instructional process with continuous teacher feedback to assist with targeted instructional changes at the classroom and student level (Kurz, Elliott, & Shrago, 2009). Enhancing access to “what should be learned and will be tested” has been the main topic of this chapter. To this end, OTL and its measurement hold great promise. And yet, the future of OTL ultimately lies in our ability to *deliver* this promise for all students.

References

- Abedi, J., Leon, S., & Kao, J. C. (2008). *Examining differential item functioning in reading assessments for students with disabilities* (CRESST Report No. 744). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Ahearn, E. (2006). *Standards-based IEPs: Implementation in selected states*. Alexandria, VA: National Association of State Directors of Special Education.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American with Disabilities Act, 42 U.S.C. §§ 12101 *et seq.* (1990).
- Anderson, L. W. (1986). Opportunity to learn. In T. Husén & T. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies*. (pp. 3682–3686). Oxford, UK: Pergamon.
- Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory into Practice*, 41(4), 255–260.
- Armbruster, B. B., Stevens, R. J., & Rosenshine, B. (1977). *Analyzing content coverage and emphasis: A study of three curricula and two tests* (Technical Report No. 26). Urbana, IL: Center for the Study of Reading, University of Illinois.
- Bloom, B. S. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.
- Borg, W. R. (1979). Teacher coverage of academic content and pupil achievement. *Journal of Educational Psychology*, 71(5), 635–645.
- Borg, W. R. (1980). Time and school learning. In C. Denham & A. Lieberman (Eds.), *Time to learn* (pp. 33–72). Washington, DC: National Institute of Education.
- Boscardin, C. K., Aguirre-Muñoz, Z., Chinen, M., Leon, S., & Shin, H. S. (2004). *Consequences and validity of performance assessment for English learners: Assessing opportunity to learn (OTL) in grade 6*

- language arts (CSE Report No. 635). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillian.
- Burstein, L., & Winters, L. (1994, June). *Models for collecting and using data on opportunity to learn at the state level: OTL options for the CCSSO SCASS science assessment*. Presented at the CCSSO National Conference on Large-scale Assessment, Albuquerque, NM.
- Camburn, E., & Barnes, C. A. (2004). Assessing the validity of a language arts instruction log through triangulation. *Elementary School Journal*, 105(1), 49–73.
- Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, 64(8), 723–733.
- Carroll, J. B. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher*, 18(1), 26–31.
- Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries*. New York: Halsted Press.
- Conderman, G., & Katsiyannis, A. (2002). Instructional issues and practices in secondary special education. *Remedial and Special Education*, 23(3), 169–179.
- Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, 8(4), 305–336.
- Connor, C. M., Morrison, F. J., & Underwood, P. S. (2007). A second chance in second grade: The independent and cumulative impact of first- and second-grade reading instruction and students' letter-word reading skill growth. *Scientific Studies of Reading*, 11(3), 199–233.
- Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., et al. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher*, 38(2), 85–99.
- Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., et al. (2009). Individualizing student instruction precisely: Effects of child \times instruction interactions on first graders' literacy development. *Child Development*, 80(1), 77–100.
- Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis*, 2(1), 7–25.
- Croninger, R. G., & Valli, L. (2009). "Where is the action?" Challenges to studying the teaching of reading in elementary classrooms. *Educational Researcher*, 38(2), 100–108.
- D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment*, 12(1), 1–22.
- Denham, C., & Lieberman, A. (Eds.). (1980). *Time to learn*. Washington, DC: National Institute for Education.
- Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational Researcher*, 38(7), 518–521.
- Elliott, S. N., Kurz, A., & Neergaard, L. (in press). Large-scale assessment for educational accountability. In S. Graham, A. Bus, S. Major, & L. Swanson (Eds.), *The handbook of educational psychology: Application of educational psychology to learning and teaching* (Vol. 3). Washington, DC: American Psychological Association.
- Fisher, C. W., & Berliner, D. C. (Eds.). (1985). *Perspectives on instructional time*. New York: Longman.
- Fisher, C. W., Berliner, D. C., Filby, N. N., Marliave, R., Cahen, L. S., & Dishaw, M. M. (1980). Teaching behaviors, academic learning time, and student achievement: An overview. In C. Denham & A. Lieberman (Eds.), *Time to learn* (pp. 7–32). Washington, DC: National Institute of Education.
- Fuchs, D., Fuchs, L. S., & Stecker, P. M. (2010). The "blurring" of special education in a new continuum of general education placements and services. *Exceptional Children*, 76(3), 301–323.
- Gagné, R. M. (1977). *The conditions of learning*. Chicago: Holt, Rinehart & Winston.
- Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, 19(4), 325–338.
- Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, 79(3), 1202–1242.
- Gettinger, M., & Seibert, J. K. (2002). Best practices in increasing academic learning time. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (Vol. 1, pp. 773–787). Bethesda, MD: National Association of School Psychologists.
- Harnischfeger, A., & Wiley, D. E. (1976). The teaching-learning process in elementary schools: A synoptic view. *Curriculum Inquiry*, 6(1), 5–43.
- Herman, J. L., & Abedi, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics* (CSE Report No. 633). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.
- Herman, J. L., Klein, D. C., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice*, 19(4), 16–24.
- Heubert, J. P. (2004). High-stakes testing in a changing environment: Disparate impact, opportunity to learn, and current legal protections. In S. H. Fuhrman &

- R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press (pp. 220–242).
- Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries*. New York: Wiley.
- Individuals with Disabilities Education Act Amendments, 20 U.S.C. §§ 1400 *et seq.* (1997).
- Individuals with Disabilities Education Improvement Act, Amending 20 U.S.C. §§ 1400 *et seq.* (2004).
- Jackson, P. W. (1990). *Life in classrooms*. New York: Teachers College Press.
- Jenkins, J. R., & Pany, D. (1978). Curriculum biases in reading achievement tests. *Journal of Reading Behavior*, 10(4), 345–357.
- Karger, J. (2005). *Access to the general education curriculum for students with disabilities: A discussion of the interrelationship between IDEA and NCLB*. Wakefield, MA: National Center on Accessing the General Curriculum.
- Kersaint, G., Lewis, J., Potter, R., & Meisels, G. (2007). Why teachers leave: Factors that influence retention and resignation. *Teaching and Teacher Education*, 23(6), 775–794.
- Kurz, A., & Elliott, S. N. (2011). Overcoming barriers to access for students with disabilities: Testing accommodations and beyond. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques* (pp. 31–58). Charlotte, NC: Information Age Publishing.
- Kurz, A., Elliott, S. N., & Shrago, J. S. (2009). *MyiLOGS: My instructional learning opportunities guidance system*. Nashville, TN: Vanderbilt University.
- Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *Journal of Special Education*, 44(3), 1–20.
- Linn, R. L. (2008). Educational accountability systems. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 3–24). New York: Routledge.
- Malmgren, K. W., McLaughlin, M. J., & Nolet, V. (2005). Accounting for the performance of students with disabilities on statewide assessments. *The Journal of Special Education*, 39(2), 86–96.
- Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, 79(4), 1332–1361.
- Marzano, R. J. (2002). *A new era of school reform: Going where the research takes us* (REL No. #RJ96006101). Aurora, CO: Mid-continent Research for Education and Learning.
- Mastropieri, M. A., & Scruggs, T. E. (2002). *Effective instruction for special education* (3rd ed.). Austin, TX: PRO-ED.
- McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis*, 17(3), 305–322.
- McDonnell, L. M., McLaughlin, M. J., & Morison, P. (1997). *Reform for one and all: Standards-based reform and students with disabilities*. Washington, DC: National Academy of Sciences Press.
- McLaughlin, M. J. (1999). Access to the general education curriculum: Paperwork and procedures or redefining “special education.” *Journal of Special Education Leadership*, 12(1), 9–14.
- McLaughlin, M. J. (2010). Evolving interpretations of educational equity and students with disabilities. *Exceptional Children*, 76(3), 265–278.
- Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement*, 23(3), 185–196.
- Metzker, B. (2003). *Time and learning* (ED474260). Eugene, OR: University of Oregon.
- No Child Left Behind Act, 20 U.S.C. §§ 6301 *et seq.* (2001).
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119.
- Pianta, R. C., Belsky, J., Houts, R., Morrison, F., & the National Institute of Child Health and Human Development. (2007). Teaching: Opportunities to learn in America’s elementary classrooms. *Science*, 315, 1795–1796.
- Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, 29(4), 3–14.
- Porter, A. C. (1991). Creating a system of school process indicators. *Educational Evaluation and Policy Analysis*, 13(1), 13–29.
- Porter, A. C. (1993). School delivery standards. *Educational Researcher*, 22(5), 24–30.
- Porter, A. C. (1995). The uses and misuses of opportunity-to-learn standards. *Educational Researcher*, 24(1), 21–27.
- Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31(7), 3–14.
- Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–159). Mahwah, NJ: Lawrence Erlbaum.
- Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. Fuhrman (Ed.), *From the Capitol to the classroom: Standards-based reform in the states. One Hundredth Yearbook of the National Society for the Study of Education* (pp. 60–80). Chicago: University of Chicago Press.
- Porter, A. C., Kirst, M. W., Osthoff, E. J., Smithson, J. L., & Schneider, S. A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms* (Final Report). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

- Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. (1978). *Impact on what? The importance of content covered* (Research Series No. 2). East Lansing, MI: Michigan State University, Institute for Research on Teaching.
- Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*(1), 27–51.
- Pullin, D. C. (2008). Assessment, equity, and opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn*. New York: Cambridge University Press.
- Pullin, D. C., & Haertel, E. H. (2008). Assessment through the lens of “opportunity to learn”. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 17–41). New York: Cambridge University Press.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Roach, A. T., & Elliott, S. N. (2006). The influence of access to general education curriculum on alternate assessment performance of students with significant cognitive disabilities. *Educational Evaluation and Policy Analysis, 28*(2), 181–194.
- Roach, A. T., Chilungu, E. N., LaSalle, T. P., Talapatra, D., Vignieri, M. J., & Kurz, A. (2009). Opportunities and options for facilitating and evaluating access to the general curriculum for students with disabilities. *Peabody Journal of Education, 84*(4), 511–528.
- Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools, 45*(2), 158–176.
- Rogosa, D. (1995). Myths and methods: Myths about longitudinal research plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Lawrence Erlbaum.
- Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology, 76*(6), 1000–1027.
- Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the Study of Instructional Improvement. *Educational Researcher, 38*(2), 120–131.
- Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *Elementary School Journal, 105*(1), 75–101.
- Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*(8), 1525–1567.
- Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. New York: Pergamon.
- Schmidt, W. H., & McKnight, C. C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis, 17*(3), 337–353.
- Shepard, K., Hammerness, K., Darling-Hammond, L., & Rust, F. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275–326). San Francisco: Jossey-Bass.
- Smithson, J. L., & Collares, A. C. (2007). *Alignment as a predictor of student achievement gains*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.
- Smithson, J. L., & Porter, A. C. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum* (CPRE Research Report No. 31). New Brunswick, NJ: Consortium for Policy Research in Education.
- Smithson, J. L., Porter, A. C., & Blank, R. K. (1995). *Describing the enacted curriculum: Development and dissemination of opportunity to learn indicators in science education* (ED385430). Washington, DC: Council of Chief State School Officers.
- Stedman, L. C. (1997). International achievement differences: An assessment of a new perspective. *Educational Researcher, 26*(3), 4–15.
- Stevens, F. I. (1996, April). *The need to expand the opportunity to learn conceptual framework: Should students, parents, and school resources be included?* Paper presented at the annual meeting of the American Educational Research Association, New York City, NY.
- Stevens, F. I., & Grymes, J. (1993). *Opportunity to learn: Issues of equity for poor and minority students* (NCES No. 93-232). Washington, DC: National Center for Education Statistics.
- Stoolmiller, M., & Bank, L. (1995). Autoregressive effects in structural equation models: We see some problems. In J. M. Gottman (Ed.), *The analysis of change* (pp. 261–276). Mahwah, NJ: Lawrence Erlbaum.
- Vannest, K. J., & Hagan-Burke, S. (2010). Teacher time use in special education. *Remedial and Special Education, 31*(2), 126–142.
- Vannest, K. J., & Parker, R. I. (2010). Measuring time: The stability of special education teacher time use. *Journal of Special Education, 44*(2), 94–106.
- Vaughn, S., Gersten, R., & Chard, D. J. (2000). The underlying message in LD intervention research: Findings from research syntheses. *Exceptional Children, 67*(1), 99–114.
- Walberg, H. J. (1980). A psychological theory of educational productivity. In F. H. Farley & N. Gordon (Eds.), *Psychology and education* (pp. 81–110). Berkeley, CA: McCutchan.
- Walberg, H. J. (1986). Syntheses of research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 214–229). New York: Macmillan Publishing Company.
- Walberg, H. J. (1988). Synthesis of research on time and learning. *Educational Leadership, 45*(6), 76–85.

- Walker, D. F., & Schaffarzick, J. (1974). Comparing curricula. *Review of Educational Research, 44*(1), 83–111.
- Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*(3), 137–156.
- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.
- Winfield, L. F. (1987). Teachers' estimates of test content covered in class and first-grade students' reading achievement. *Elementary School Journal, 87*(4), 437–454.
- Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *Journal of Negro Education, 62*(3), 288–310.
- Yazzie-Mintz, E. (2007). *Voices of students on engagement: A report on the 2006 high school survey of student engagement*. Bloomington, IN: Indiana University, Center for Evaluation and Education Policy.
- Yoon, B., & Resnick, L. B. (1998). Instructional validity, opportunity to learn and equity: New standards examinations for the California mathematics renaissance (CSE Technical Report No. 484). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.
- Ysseldyke, J., Thurlow, M., Langenfeld, K., Nelson, J. R., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Technical Report No. 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Instructional Adaptations: Accommodations and Modifications That Support Accessible Instruction

Leanne R. Ketterlin-Geller and Elisa M. Jamgochian

Introduction

The purpose of this book is to describe how educational environments can be made maximally accessible for students with disabilities. In this chapter we situate the discussion of accessibility within the classroom instructional setting. We describe how the use of instructional accommodations and instructional modifications can help eliminate barriers inherent in many instructional products, environments, or systems that prevent maximum engagement by students with disabilities. Implementing these instructional adaptations may increase the accessibility of the learning environment, thereby permitting equal access for all students.

Recent National Assessment of Educational Progress (NCES, 2009) results in reading and mathematics indicate that students with disabilities are not reaching expected proficiency levels. Specifically, the 2009 reading scores identify that 12% of fourth-grade and 8% of eighth-grade students with disabilities are performing at or above proficiency as compared to 34% and 33% of their peers without disabilities, respectively. Similarly, the 2009 mathematics results show 19% of fourth-grade and 9% of eighth-grade students with disabilities are performing at or above

proficiency as compared to 41% and 35% of the general education population, respectively. One possible cause of these results is inaccessible instruction.

Instructional accessibility is influenced by the interaction between individual student characteristics and features of instructional products, environments, or systems. Accessibility can be enhanced by intentionally designing and delivering instruction that aligns with the student's needs. Just as testing adaptations improve the accessibility of assessments, instructional adaptations support students' interactions through changes in the *presentation*, *setting*, *timing* or *scheduling*, and *response mode* of instruction. The purpose of this chapter is to describe the role of instructional adaptations in improving the accessibility of instruction.

As with testing adaptations, instructional adaptations are divided into two general classes: accommodations and modifications. *Instructional accommodations* are adaptations to the design or delivery of instruction and associated materials that do not change the breadth of content coverage and depth of knowledge of the grade-level content standards. The use of instructional accommodations has no impact on the performance expectations implied in the general education learning objectives. As such, most students using instructional accommodations are expected to learn the same material to the same level of proficiency as students who are not using instructional accommodations. Conversely, *instructional modifications* are adaptations to the

L.R. Ketterlin-Geller (✉)
Southern Methodist University, Dallas, TX 75252, USA
e-mail: lkgteller@smu.edu

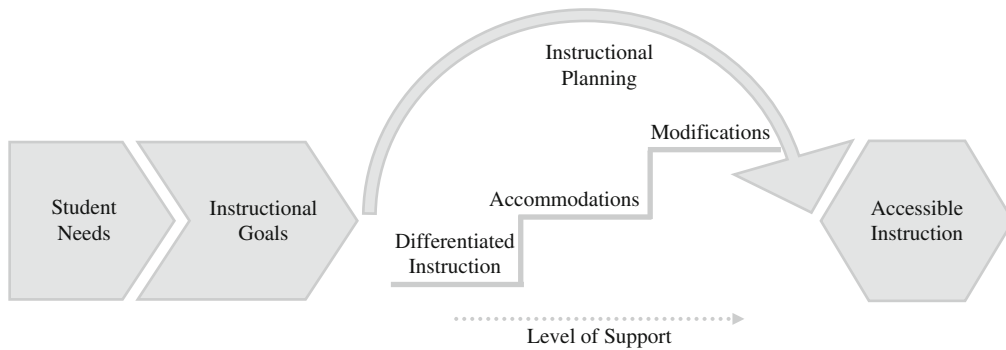


Fig. 7.1 Instructional planning to support access

design or delivery of instruction and associated materials that result in a reduction of the breadth of content coverage and/or depth of knowledge of the grade-level content standards. Instructional modifications may also result in a reduction of the performance expectations established for the general education student population. As such, students receiving instructional modifications are not expected to learn the same material to the same level of proficiency as students in the general education classroom.

In this chapter, we highlight assumptions and implications related to the content and performance expectations of classroom instruction when making instructional adaptations. Further, we specify the distinctions underlying the use of instructional accommodations as opposed to instructional modifications and present examples to illustrate the similarities and differences in these adaptations. Additionally, we describe implications for implementing instructional adaptations including potential consequences of misalignment between adaptations used in instruction and those used in assessment. Figure 7.1 provides an advanced organizer that identifies and illustrates the interaction between the instructional planning components to support students' access to instruction. Students' needs and instructional goals are considered during instructional planning. Support features including differentiated instruction, accommodations, and modifications are implemented depending on the level of support individuals need to access instruction.

Instructional Adaptations: Important Distinctions

To understand the role of instructional adaptations in improving the accessibility of instructional environments, two important distinctions are necessary. First, at the broadest level, instructional adaptations designed to improve access for students with disabilities are differentiated from instructional planning approaches applied to classroom instruction that promote learning for all students. Second, the defining characteristics that differentiate instructional accommodations from instructional modifications are fully articulated. Specifically, the degree of alignment with grade-level content standards and the consistency across performance expectations are discussed to distinguish between these instructional adaptations.

Distinguishing Between Instructional Adaptations and Differentiated Instruction

To begin the discussion on instructional adaptations, it is important to distinguish these practices from the process of differentiating instruction. Similar to instructional adaptations, *differentiated instruction* is an instructional planning approach that intentionally adapts instructional design and delivery methods to support student learning of the instructional objectives. When differentiating

instruction, the teacher considers students' background knowledge, readiness for the instructional objective, language skills and abilities, preferences, and interests of all students in his/her classroom (Hall, 2002). These variables are matched to different instructional methods such as varied supports from the teacher and/or peers, modeling and the use of manipulatives, coaching, use of different media sources, self-reflection, and alternate review activities (Tomlinson, 1999). The process of aligning student characteristics with instructional approaches may be formally articulated in teachers' lesson plans or may be generated informally through teachers' interactions with students. Ultimately, by incorporating student variables in the instructional planning process, the teacher can adjust the learning experiences accordingly to facilitate student learning of the instructional objective.

Because both instructional adaptations and differentiated instruction focus on adapting the design and delivery of instruction to support student learning, many teachers intertwine these instructional methodologies. However, several key factors distinguish instructional adaptations from differentiated instruction. Specifically, Individualized Education Program (IEP) teams assign instructional adaptations based on the persistent needs of students with disabilities. As with other IEP decisions, the student's characteristics are considered along with performance data to identify the student's strengths and limitations. Instructional adaptations are identified that align with the student's needs in meeting the most appropriate learning objectives. In some cases, IEP teams may evaluate the effectiveness of differentiated instructional approaches on the student's learning prior to systematically assigning the approaches as instructional adaptations on the student's IEP. Because the IEP is a legally binding document between the educational service providers and the student (Individuals with Disabilities Education Act [IDEA], 2004), the instructional adaptations listed on the IEP must be systematically applied across the learning environment. Even though some instructional adaptations may mirror differentiated instructional strategies (e.g., use of a graphic organizer,

additional practice opportunities), when specified on an IEP, the teacher must regularly incorporate them into the student's instructional opportunities. Consequently, the flexibility in instructional planning proffered by differentiated instruction is not afforded to teachers when implementing instructional adaptations.

It is important to note that the use of instructional adaptations does not supplant the use of differentiated instruction. In fact, students with disabilities often benefit from differentiated instruction in addition to the instructional adaptations identified on their IEP. For example, an instructional adaptation for a student with working memory limitations might include using a graphic organizer to document story elements when reading fictional text. To increase the student's motivation, a differentiated instructional strategy might involve allowing him/her to select the text based on personal interests. The graphic organizer would be formally identified as an instructional adaptation on the student's IEP; however, it is not necessary to document variations in story context. Researchers and policy makers identify the value of differentiating general education instruction as a mechanism for making the general education classroom more accessible for students with disabilities (e.g., Gersten et al., 2008). As such, these instructional approaches can be used conjunctively to support student learning.

Distinguishing Between Instructional Accommodations and Modifications

The primary distinctions between instructional accommodations and instructional modifications are based on the degree of alignment with grade-level content standards and the consistency across performance expectations. As previously introduced, most students using instructional accommodations are expected to learn the same material to the same level of performance as other students in the general education classroom. Conversely, students receiving instructional modifications are not necessarily expected to learn the same material to the same level of performance as other

students. Although these distinguishing characteristics often help differentiate between accommodations and modifications, they are not as easily separated. For example, a student with significant cognitive disabilities might receive both instructional accommodations (e.g., large print or verbatim oral presentation of direction) at the same time he or she is receiving instructional modifications (e.g., using off-grade-level reading material). As such, it is often useful to consider the impact instructional adaptations have on the outcomes of instruction when classifying them as accommodations or modifications.

Alignment of Instructional Adaptations with Grade-Level Content Standards

As noted elsewhere in this volume, IDEA (2004) requires IEP teams to align students' instructional goals with grade-level content standards that guide learning objectives for all students in the state. Because instructional goals are enacted through the design and delivery of instruction, the adapted content resulting from the use of instructional accommodations and modifications should be evaluated for alignment with grade-level content standards. When referencing alignment to grade-level content standards, two dimensions are typically considered: content coverage and depth of knowledge (also referred to as cognitive complexity) (La Marca, 2001). Alignment to grade-level content standards is evaluated by considering the degree of consistency of content coverage and depth of knowledge between the content standards and students' instructional goals as stated on their IEP.

Content coverage refers to the knowledge and skills represented in the content standards, such as facts, concepts, principles, and procedures. Federal legislation (NCLB, 2001; IDEA, 2004) requires that all students receive instruction on grade-level content, regardless of their educational classification. A key distinction between instructional accommodations and instructional modifications is the extent to which the content covered aligns with the grade-level content standards. Because instructional accommodations are

not intended to change the breadth or depth of the grade-level content standards, instructional accommodations should maintain a one-to-one correspondence in the content coverage between the instructional goals and the grade-level content standards. This does not preclude accommodations such as providing extra time for learning the content or rearranging the sequence of content presented, as long as the content coverage remains consistent. On the other hand, for some students with disabilities, it may not be feasible to learn the breadth of content stated in the grade-level content standards. For these students, the IEP team may identify instructional modifications that reduce the content covered within the year. For example, the IEP team may recommend changing the scope of instruction to focus on the "big ideas" presented in the content standards. As such, these instructional modifications reduce the degree of alignment between the instructional design and delivery and the grade-level content standards.

Depth of knowledge, or cognitive complexity, refers to the level of cognitive engagement at which students are expected to interact with the content. Taxonomies of depth of knowledge (e.g., Webb's taxonomy (Webb, 1999), Bloom's taxonomy (Krathwohl, 2002)) classify engagement into categories depending on the type of interaction expected from the student, such as reiteration of facts, summarization of a concept, or application of a procedure to a novel situation. Instructional accommodations differ from instructional modifications in the degree to which they alter the cognitive complexity of the grade-level content standards. Instructional accommodations designed to support students' engagement with the content should not alter the intended cognitive complexity of the grade-level content standard. Instead, these accommodations may change the sequence of instruction to provide additional practice in prerequisite knowledge and skills, prime background knowledge, or scaffold instruction to gradually increase the level of cognitive complexity in which students interact with content. Still, instruction should culminate in the depth of knowledge that is consistent with the grade-level content standards. However,

instructional modifications may reduce the cognitive complexity of instruction.

Federal legislation recognizes that the depth of knowledge in the grade-level content standards may not be appropriate for some students with disabilities (NCLB, 2001; IDEA, 2004). For these students, IEP teams may recommend instructional modifications that reduce level of cognitive complexity implied by the grade-level content standards. For example, consider the mathematics content standard that expects students to classify and sort two-dimensional figures based on their attributes. An instructional modification may instead require students to identify two-dimensional figures. By reducing the depth of knowledge of the content standard, instructional modifications improve accessibility to instruction and associated materials for some students with disabilities. However, because instructional modifications alter alignment between grade-level content standards and instruction, students receiving instructional modifications do not have the same opportunity to learn the breadth of content and/or depth of knowledge as other students in the general education classroom.

Consistency of Performance Expectations with General Education

Performance standards or expectations are most typically expressed in terms of the degree of mastery of the instructional objectives students are expected to reach. Sometimes this is manifested in the criterion established for decision making (e.g., students must complete 80% of the homework items correctly); other times it is evident in the characteristics of the tasks students are expected to complete (e.g., students answer 20 homework problems with graduated difficulty). As previously introduced, the consistency of these performance expectations with the general education instructional objectives distinguishes instructional accommodations from instructional modifications.

Instructional accommodations are typically intended for students who are expected to

learn the same material to the same level of performance expectations as students who are not receiving instructional accommodations. As such, for students with disabilities who may need additional instructional support to reach the general education performance expectations, IEP teams should consider instructional accommodations that scaffold students' development of content knowledge and skills to the expected level of proficiency, as opposed to instructional adaptations that reduce these expectations. Such instructional accommodations may include providing more guided practice on instructional activities to reach the expected performance level on independent class work or extending the instructional time in which students are expected to reach the performance level. Although these accommodations may alter the schedule for meeting the general education performance expectations, they result in no changes to the instructional objectives themselves.

Modifications to the performance expectations should be considered for students with disabilities who are not likely to reach the performance expectations established for the general education population within the academic year. Although it is often difficult to determine with certainty who these students are, their IEP teams may recommend reducing the level of proficiency expected on independent work (e.g., students must complete 60% as opposed to 80% of the homework items correctly) or reducing the number or difficulty of the tasks to be completed (e.g., students answer 10 homework problems as opposed to 20). This instructional modification may also be implemented in tandem with a reduction in the content coverage and/or depth of knowledge of the grade-level content standards. In these instances, students might be completing fewer instructional tasks that cover a narrower focus of content and involve less difficult cognitive processing skills.

In summary, instructional accommodations and instructional modifications serve the purpose of improving accessibility of the general education curriculum by changing the design and delivery of instruction and associated materials. These adaptations differ based on the extent to

which these changes alter the alignment with the content coverage and/or depth of knowledge of the grade-level content standards and/or the consistency across performance expectations. In the next section, we describe possible consequences IEP teams should consider prior to assigning instructional adaptations.

Consequences of Implementing Instructional Adaptations

Prior to assigning instructional accommodations and instructional modifications, IEP teams must be fully aware of the likely consequences of these adaptations. Specifically, implementing instructional adaptations should not preclude instruction in the knowledge and skill areas that are being accommodated or modified. Additionally, implementing instructional modifications can significantly change students' immediate and subsequent instructional programming, as well as assessment participation options. Further, assigning instructional adaptations may influence the adaptations students use during testing. IEP teams should not interpret these implications as potential reasons for not assigning needed instructional adaptations; instead, IEP teams should be aware of these issues to facilitate decision making and ensure fidelity of implementation.

Consequences of Overuse of Instructional Adaptations

Instructional adaptations have the potential of being overused in classroom practices at the expense of instruction on the knowledge and skills that are being supported (Phillips, 1999). Although most often unintended, some teachers may inadvertently omit instruction on the knowledge and skills that are being accommodated or modified from the student's educational programming. For example, students struggling to master computational fluency may be allowed to use a calculator to solve situated problems. However, these students need continued instruction and

practice to develop fluency with computational algorithms. As such, the use of instructional adaptations should be paired with sufficient and focused instruction on the knowledge and skills that are being supported. By developing students' knowledge and skills in these areas, students may be able to gradually reduce their use and dependency on the instructional adaptation.

Related to overuse of instructional adaptations is misuse or misclassification of instructional changes that result in a reduction of content or expectations. For example, in a survey of 176 general and special education teachers' instruction and accommodations practices in mathematics, Maccini and Gagnon (2006) found 41% of special education teachers and 19% of general education teachers assigned fewer practice problems as an instructional strategy and 33% of special education teachers and 17% of general education teachers assigned fewer homework problems during instruction on computational skills. Similar results were observed when the content focused on solving multi-step problems. This adaptation was perceived as an instructional strategy; however, because the expectation of performance was reduced from that of the general education instructional objectives, it should be classified as a modification. Similarly, in this same study, Maccini and Gagnon (2006) found that 54% of special education teachers and 40% of general education teachers allowed students to use calculators during instruction on basic skills or computational tasks. Because the construct of instruction was closely aligned with the knowledge and skills that were being supported through the use of a calculator, this instructional adaptation should be viewed as a modification as opposed to an instructional strategy. These results mirror those of other researchers examining teachers' classification of test adaptations (e.g., Hollenbeck, Tindal, & Almond, 1998; Ketterlin-Geller, Alonzo, Braun-Monegan, & Tindal, 2007). Assignment of instructional adaptations is beyond the scope of this chapter; however, results of these studies indicate that teachers may need additional support in distinguishing between instructional strategies, accommodations, and modifications.

Consequences of Instructional Modifications

Regardless of the type of instructional modification(s) assigned to a student, IEP teams must be aware of the implications of these adaptations. First, modifications alter the alignment between grade-level content standards and classroom instruction. Although this statement is obvious, the implications are significant: The altered alignment inherently reduces students' opportunity to learn grade-level content standards to the level of proficiency expected of the general education population. Consequently, subsequent learning experiences will be influenced by the (lack of) content coverage, (reduced) depth of knowledge, and/or (lower) expectations of proficiency. For example, consider an instructional modification for a fourth-grade student in mathematics that reduces the geometry content covered to focus only on two-dimensional and not three-dimensional figures. Subsequent instructional objectives that rely on the student's understanding of three-dimensional figures, such as understanding and calculating volume, will also need to be modified to account for this student's lack of understanding of three-dimensional figures. Similarly, consider an instructional modification that allows a seventh-grade student studying life science to use a textbook written at the fourth-grade reading level. Inherent in this instructional modification is a reduction in the student's exposure to grade-level vocabulary, which may impact subsequent instructional objectives. As such, whenever possible, instructional modifications should be limited in favor of providing accommodations that maintain alignment with the grade-level content standards.

A second implication of assigning instructional modifications is the (mis)alignment between grade-level content standards and large-scale assessments. Because grade-level content standards form the basis for most large-scale state and district achievement tests, it follows that these assessments are not appropriate for students who are not receiving instruction aligned to these standards (Browder, Spooner, Wakeman,

Trela, & Baker, 2006). As such, IEP teams need to consider alternate participation options on large-scale state and district achievement assessments that might be most appropriate for students receiving instructional modifications.

Federal legislation allows for two alternate participation options on state accountability tests: alternate assessments based on alternate achievement standards (AA-AAS) and alternate assessments based on modified achievement standards (AA-MAS). Specific guidelines govern the selection of students participating in AA-AAS (e.g., students with the most significant cognitive disabilities), and states are currently in the process of providing guidance for selecting students who may participate in the AA-MAS. Federal regulations, however, stipulate that scores from only a small portion of students with disabilities (10% for the AA-AAS and 20% for the AA-MAS, which translates to 1% and 2% of the overall student population, respectively) can be included in states' reporting of proficiency for Adequate Yearly Progress (AYP) under these participation options. Consequently, these participation options should be selected only for those students for whom the general education assessment is inappropriate. As such, prior to assigning instructional modifications, IEP teams need to consider the lasting and immediate implications for students' instructional programming and assessment participation.

Interdependence Between Accessible Instruction and Accessible Assessments

To support the educational needs of students with disabilities, accessible instruction is inseparable from accessible assessments. Consider, for example, an accessible instructional environment coupled with inaccessible assessments. In this setting, all students have equal opportunities to learn instructional content to specific expectations; however, as a result of inaccessible assessments, students with disabilities may not be able to demonstrate what they know and can do on classroom-based and large-scale assessments.

Consequently, poor student performance may be (mis)interpreted as a lack of progress in learning the instructional content to specific performance expectations. Because accurate representation of students' abilities was obstructed by inaccessible assessments, decisions based on inaccessible assessments are invalid. Issues about test-score validity are discussed in [Chapters 10, 12, and 13](#) of this book.

Conversely, consider a situation in which instruction is inaccessible but the assessments are accessible. In this case, all students do not have an equal opportunity to learn the content of instruction. Although the assessments may accurately represent students' (lack of) knowledge and skills, the representation may not adequately reflect students' potential abilities. However, if interpreted appropriately, these results might highlight the lack of opportunities to learn the content. As these scenarios point out, to support student learning, accessibility must be viewed across multiple dimensions of the educational environment.

Because of the interdependence between accessible instruction and accessible assessments, instructional adaptations should form the basis for testing adaptations. As such, instructional accommodations and instructional modifications should be seamlessly integrated into classroom-based and large-scale assessment practices. In fact, in 2007, 47 states required that the use of instructional accommodations be considered when making decisions about testing accommodations (Christensen, Lazarus, Crone, & Thurlow, 2008). However, some researchers point out the discrepancy between these practices. In a study examining the IEPs for 280 students with disabilities, Ysseldyke et al. (2001) found that approximately 86% of the participants had either instructional or testing accommodations listed on their IEPs. Of these students, 84% had instructional accommodations that matched testing accommodations. The rate of agreement varied based on the prevalence of the disability: Instructional and testing accommodations matched in 98% of the IEPs for students with low-prevalence disabilities as compared to 71% for students with moderate-prevalence and

84% for high-prevalence disabilities. Similarly, DeStefano, Shriner, and Lloyd (2001) conducted a study on the accommodation practices of 100 urban educators. In their findings, they noted little association between accommodations used in instruction and those provided on tests; most frequently, teachers used more accommodations during testing than instruction. However, after teachers received training on accommodation assignment procedures, results indicated that teachers used more accommodations during instruction and fewer during assessments. Ultimately, students should have access to the same adaptations in instruction and assessment.

Instructional Adaptations That Support Accessible Learning Environments

In this chapter, we highlighted characteristics of instructional accommodations that distinguish these instructional adaptations from instructional modifications. This differentiation provided a basis for discussing the implications of assigning instructional adaptations on current and future instructional programming and assessment practices. The remainder of this chapter will focus on practical suggestions for implementing these instructional adaptations in classroom instruction. We describe instructional accommodations followed by instructional modifications in four categories: presentation, setting, timing or scheduling, and response modes. Although we identify possible adaptations, it is important to note that we do not discuss assignment criteria for IEP teams to consider when making instructional adaptation decisions.

Table 7.1 provides a summary of types of instructional adaptations, student characteristics to describe those who may benefit from each type of adaptation, and examples of accommodations and modifications to succinctly illustrate the differences between these instructional adaptations. It is important to note that allowable accommodations may differ from state to state, and differences may exist between state and federal policies. Because of this, teachers and other

Table 7.1 Summary of instructional adaptations

Adaptation type	Benefits students. . .	Accommodation examples	Modification examples
Presentation	With physical, sensory, or cognitive difficulties that affect perception of visual or auditory stimulus	Visual – Braille, large print font, copies of written material, magnification, electronic texts/materials Auditory – read aloud, audio-/video-recorded presentations, screen readers, amplification devices, peer-assisted learning, cross-grade tutoring Tactile – tactile prompts, physical guidance, raised-line paper, raised graphics	Abridged version of text or novel, reduced content coverage (fewer concepts), reduced depth of knowledge, read aloud (as a modification)
Setting	Who are distractible Whose adaptation may be distractible Who need alternate physical access	Preferential seating/grouping (near teacher, away from doors and windows), individual or small group instruction, headphones, study carrel, separate location	Independent work on group task, work with partner on individual task
Timing and scheduling	Who process information slowly With physical disabilities that affect task completion Who use adaptations that require additional time	Extended time, multiple breaks, breaking task into smaller parts, on-task reminders (e.g., timer), daily schedule posted	Extended time/frequent breaks on timed tasks
Response	With expressive language difficulties or motoric impairments Who need additional support with organization or problem solving	Write, type, or word-process responses, point or sign to indicate response, communication devices, text-to-speech, audio record responses, scribe, oral response, manipulatives, scratch paper, respond directly on assignment, visual/graphic organizers, calculator	Reducing number of items to complete or paragraphs to write, providing fewer answer choices, reducing depth of knowledge required, dictionary or thesaurus (as a modification)

instructional decision makers need to be well informed of these guidelines when assigning student accommodations.

Instructional Accommodations

Instructional accommodations are adaptations to the design or delivery of instruction and associated materials that do not change the breadth of content coverage and depth of knowledge of the grade-level content standards. Additionally, the use of instructional accommodations has no impact on the level of performance students are expected to reach. As such, most students using instructional accommodations are expected to learn the same material to the same level of expectation as students who are not

using instructional accommodations, although the learning process may take more time. Below is a summary of instructional accommodations categorized by type: presentation, setting, timing or scheduling, and response modes. IEP teams should reference state accommodations policies for allowable accommodations prior to assignment.

Changes in Presentation

Presentation accommodations provide students with alternate ways to access instructional material. Students with difficulties perceiving visual or auditory stimulus may benefit from visual, tactile, and/or auditory changes in presentation (Thompson, 2005). To increase the accessibility

for all students, general guidelines suggest that material presented visually should include type in a standard, legible font that is no less than 12-point size for printed materials (24-point for projected material); high contrast between text and background; sufficient space between letter, words, and lines; and text that is left-aligned (with staggered right margins). Also, images and graphics should be relevant to the content of the text, have dark lines, and sufficient contrast (Thompson, Johnstone, Anderson, & Miller, 2005). Material presented auditorily should provide students with appropriate cues for understanding, including expression and stance (e.g., facing class when speaking), repetition of questions, and summary of discussion or material covered and important points (University of Kansas, 2005). Specific examples of visual, auditory, and tactile changes in presentation are detailed below.

Visual presentation accommodations benefit students with physical, sensory, or cognitive disabilities that affect their ability to visually read standard print. Common accommodations include Braille, large print font, magnification, and providing copies of written materials. Large-print editions of textbooks and other instructional materials are often available through the publishing company. Photocopying at a size greater than 100 percent can enlarge regular print materials. Some students may use a magnifying device (handheld, eyeglass mounted, freestanding) to view materials in a larger format. Copies of presented materials (e.g., presentation slides, notes, outlines) also support students in understanding and accessing visually presented material. Some students may also benefit from using a guide to help track while reading (e.g., bookmark, index card, or ruler), colored overlays to improve contrast (e.g., Wilkins, Lewis, Smith, Rowland, & Tweedie, 2001), or revised documents with less information and/or fewer items per page.

Technology may also support students' access. As the availability of electronic texts and materials increases, students have greater control over the visual presentation of information, including font, font size, color/contrast, graphic/image captions, and can visually track their progress through the material (e.g., Rose & Dalton, 2009).

Teachers can also make word-processed documents, presentations, or other materials available to students on a computer. By having access to the source documents, students have flexibility to individualize the material by changing visual features (e.g., font size/color, line spacing, contrast) to customize access.

Auditory changes in presentation benefit students with visual or hearing impairments or difficulty processing print or auditory information. Such changes in presentation include read-aloud accommodations, audio recordings, or video presentations of written material, amplification devices, and screen readers. Read-aloud accommodations involve oral presentation of text to students to provide access to instructional content (Tindal, 1998). A teacher or educational assistant may provide this accommodation, text may be recorded and played back for the student, text may be read aloud on a computer, or the student may read aloud on his/her own. Additionally, students may work with each other to provide this support, either through peer-assisted learning strategies (e.g., Fuchs, Fuchs, Mathes, & Simmons, 1997; Fuchs, Fuchs, Hamlett, Phillips, & Karns, 1995) or cross-grade tutoring, or classroom volunteers can provide read-aloud support. It is important to note that this is an *accommodation* as long as the intended construct remains unchanged; for example, if the instructional objective involves understanding and using new vocabulary from a story or requires a student to formulate and propose a solution to a mathematics story problem, a read-aloud accommodation maintains the goals and expectations of instruction. In contrast, if the instructional objective requires students to use letter-sound correspondence to sound out novel words or to read aloud grade-level text fluently and accurately with appropriate intonation and expression, providing a read-aloud accommodation alters the instructional goals and expectations.

Similar to the read-aloud accommodation, providing students with audio recordings or video presentations of written material can help to improve access to instruction. An additional benefit of using recorded material is that a student

has the ability to replay the recording, providing multiple exposures to the content. Also, as noted in the description of visual changes in presentation, electronic texts embed flexibility, providing students with access to auditory support, including the option to hear text spoken. Access to common technology tools may help facilitate this accommodation: Many students have cell phones or MP3 players (e.g., iPod, Zune) that support various audio files. Audio recordings of written material may be created directly on such devices or may be created, saved, and transferred from a computer-based application. In addition, many books for children are available on CDs or as downloadable audio files. Other auditory accommodations may include various assistive technologies, such as amplification devices and screen readers.

Tactile presentation accommodations benefit students with visual impairments, as well as students needing alternate representations of information. Tactile accommodations provide information to students through touch and may include tactile prompts or physical guidance (e.g., positioning a student's fingers on a keyboard, mouse, or pencil), manipulatives, writing paper with raised lines, or graphic material presented in a raised format (e.g., maps, charts, and graphs with raised lines). Tactile accommodations often require additional information, such as a verbal description, for students to fully access the content presented and understand performance expectations (O'Connell, Lieberman, & Petersen, 2006). Practically, teachers can outline maps, charts, and graphs on student worksheets in glue the day prior to their use in a lesson. When dry, the raised lines provide tactile information to help students identify the graphic.

Changes in Setting

Setting accommodations are changes to the conditions or location of an instructional setting. Students who are distractible, who receive accommodations that may distract others, or who need alternate physical access may benefit from changes in the instructional setting. To promote

access for all students, the classroom setting should be well lit, well ventilated, and have a comfortable temperature, with chairs and tables that are appropriate in height and have sufficient space for students to work, and materials and equipment should be in good condition (University of Kansas, 2005).

May involve a change of location either within or outside of the classroom. Within the classroom, a distractible student may be seated near the teacher, away from doors and windows, and/or with a group of students who are not distracting. In addition, students may receive instruction individually or in a small group, use headphones to buffer noise, or a study carrel to diminish visual distractions. If a student receives accommodations that may distract other students, such as a scribe, multiple breaks, or read-aloud, he or she may receive them in a separate location. In addition, some students may require a change in setting to improve physical access or to access equipment that is not readily available in the classroom.

Changes in Timing or Scheduling

Timing and scheduling accommodations alter the amount of time allotted to complete a test, assignment, or activity, or change the way instructional time is organized or allocated (Thompson, 2005). Changes in timing and scheduling may benefit students who need additional time, frequent breaks, or multiple exposures to information, including students who process information slowly, students with physical disabilities that may impact their task completion rate, and students who use accommodations or equipment that require additional time. Timing and scheduling accommodations may also benefit students with difficulty focusing and attending to tasks for extended periods of time, students who may become easily frustrated or anxious, or those taking medication or who have health-related disabilities that affect their ability to attend.

Possible accommodations include extending the time for task completion, providing multiple breaks during a given task, breaking a larger

task into smaller parts, and/or setting reminders to remain on task. For example, a teacher may set a timer for a student during independent class work to provide a visual cue that indicates expected time for engagement in a particular activity. Additionally, teachers can post the schedule for daily activities in a consistent place to provide students with structured guidelines for class work.

Changes in Response Mode

Response format accommodations change the way students express their knowledge and skills. Changes in response mode may benefit students with expressive language difficulties or motoric impairments or students who need additional support with organization or problem solving. Students who have difficulty with verbal expression may be allowed to write or word process their answers, point or sign to indicate their response, or use a communication device or text-to-speech program. Students with motor difficulties may respond orally, tape/digitally record responses, and/or have a scribe record their responses. Students who need support with organization or problem solving may benefit from the use of manipulatives to model problems, scratch paper to plan responses, writing their answers directly on a test or assignment rather than transferring to a separate paper, visual/graphic organizers, using a calculator, or highlighting important information within a given task.

Technology already in place in the classroom may facilitate implementation of many response accommodations. For example, teachers can provide classroom-based assignments electronically that include embedded mnemonic devices or scaffolded instructional prompts that structure students' responses. Additionally, students might type their responses to assignments or respond to multiple-choice questions by highlighting or making their answers in bold. Students can use software such as Inspiration[®] or Kidspiration[®] to help create an outline or visual map to organize their ideas prior to responding to assignments or as a method for

highlighting relationships presented in a lesson. In addition, many classrooms have electronic input devices (commonly referred to as "clickers"), which capture student responses and can be incorporated into Microsoft PowerPoint[®] presentations.

Depending on the instructional objectives, some of the accommodations described above may become modifications if they are used in a way that changes the intended construct. For example, if a student is provided with a read-aloud accommodation on a decoding or reading comprehension task, the construct may change from reading comprehension to listening comprehension.

Instructional Modifications

Instructional modifications are adaptations to the design or delivery of instruction and associated materials that result in a reduction of the breadth of content coverage and/or depth of knowledge of the grade-level content standards. Instructional modifications may also result in a reduction of the performance expectations established for the general education student population. As such, students receiving instructional modifications are not necessarily expected to learn the same material to the same level of expectation as students in the general education classroom. Below is a summary of instructional modifications categorized by type: presentation, setting, timing or scheduling, and response modes (see Table 7.1).

Changes in Presentation

Presentation modifications change instructional delivery in a way that reduces the breadth of content coverage or depth of knowledge, and in turn, alters the construct of instruction. Students who continue to demonstrate difficulty perceiving or understanding visual or auditory presentations of instructional content, given appropriate accommodations, may benefit from modifications to presentation. For these students, the cognitive

processing expectations of the learning environment may exceed their processing capacities, thereby causing cognitive overload (Mayer & Moreno, 2003). Specifically, students' cognitive processing might be burdened by the amount of new information presented during instruction that is associated with the learning goal. Alternatively, some students might find the context of the instructional environment excessively complex, thereby taxing their cognitive processing. Additional theories of students' cognitive processing are likely plausible explanations for cognitive overload. For students experiencing cognitive overload, instructional modifications may improve the accessibility of instruction by reducing the cognitive processing demands.

Presentation modifications may reduce the cognitive processing load in several ways. For example, reducing the complexity of information by simplifying the level of instructional material may decrease the cognitive processing demands thereby allowing the student to focus on learning essential knowledge and skills (Elliott, Kurz, Beddow, & Frey, 2009). For example, a student might benefit from instructional materials, such as a social studies textbook written at a lower grade level that covers the same content as the general education textbooks. Similarly, a student may meet the learning objective by reading an abridged version of a novel. These modifications provide students access to grade-level material and content but lower performance expectations. Another way to modify instructional delivery to support students' cognitive processing is to reduce content coverage; that is, students may be instructed on a limited number of concepts that represent the essence of the grade-level content standards. By eliminating non-essential attributes of instruction, students' cognitive processing can be diverted from extraneous information to essential learning (Mayer & Moreno, 2003). Additionally, the depth of knowledge intended by the grade-level content standard might be reduced. For example, rather than presenting an analysis of historical events, the teacher may compare and contrast events to simplify the material and concepts. As noted previously, these modifications have significant

implications for the student's subsequent instructional experiences.

Changes in Setting

Setting modifications change the conditions or location of the instructional setting that are warranted due to the need for specialized instructional supports, materials, or equipment. As often as possible, students should receive instruction in the general education classroom. However, for some students to meet their instructional objectives, modifications to the setting of instruction may be needed. Possible setting modifications that may be implemented in the general education classroom include allowing the student to work independently on group tasks or, conversely, allowing a student to work with a partner on an independent task. These changes are classified as modifications if they alter the content or performance expectations of the instructional objectives.

Changes in Timing or Scheduling

A modification to the timing or scheduling of instruction changes the amount or organization of time in a way that affects the goals of the instructional activity. These changes may be needed for students who require additional support beyond those provided through accommodations. Such students may have difficulty processing information, specific schedule requirements, or may have a physical disability that significantly impacts the time needed to complete tasks. An example of a scheduling accommodation may include changing the student's daily schedule while at school. A modification to the timing of a task might include providing extended time or frequent breaks on an activity or test involving fluency. Because fluency activities typically involve a rate of performance, changing the allowed time or providing multiple breaks may change the instructional objective. For example, consider a task designed to build students' fluency in mathematical computation by providing students

with 2 minutes to complete as many addition problems as possible. Allowing a student additional time on this task changes grade-level proficiency expectations, but still provides the student an opportunity to demonstrate computational understanding. Another timing modification would be to eliminate the time requirement for a given task or limit timed tasks to mastered skills. By providing such timing modifications, the expectations of proficiency or fluency and completion rate changes.

Changes in Response Mode

Modifications to response mode change a task in a way that reduces student expectations or depth of knowledge required to complete the task. Students with expressive language or information processing difficulties or motoric impairments, or those with difficulty organizing information, may require more significant changes in response mode than those offered through accommodations to demonstrate their knowledge and skills. Examples of response mode modifications include reducing the number of items a student is expected to complete or the number of answer choices per item on an assignment thus reducing proficiency expectations. A modification to response mode that changes the depth of knowledge required for an activity might include changing a task that requires students to analyze and synthesize information to one that requires them to identify the cause and effect of an event. The use of a calculator on a mathematics task that is designed to gauge students' computation skills is a modification that alters student expectations is another example of a response mode modification. Other similar examples would be the use of a dictionary or thesaurus for a vocabulary task or the use of spelling or grammar aids for a writing task that targets or measures these skills. A modification to a writing assignment that requires a five-paragraph essay might reduce the amount of writing a student is expected to do to a paragraph with three-to-five sentences supplemented with images to demonstrate content and sequence.

Integrating Instructional Adaptations Based on Students' Needs

Students' needs provide the foundation for making instructional planning decisions. In some cases, students with disabilities will be appropriately supported in the general education curriculum through differentiated instructional strategies. In other cases, students may need instructional accommodations and/or instructional modifications to reach their learning potential. In this chapter, we have described these approaches to supporting students with disabilities in isolation. The case example presented in Fig. 7.2, however, illustrates how these approaches can be integrated to provide a maximally accessible learning environment based on an individual's needs.

Conclusions

Accessibility of instruction can be enhanced by designing and delivering instruction that supports the interaction between an individual's characteristics and the format, sequencing, delivery method, and other features of instruction and the tasks or activities in which students engage. In this chapter, we described instructional adaptations as a method for increasing the accessibility of instruction for students with disabilities to promote student learning. We highlighted the important distinctions between instructional accommodations and instructional modifications and provided examples of how each adaptation may be implemented in the classroom. Moreover, we discussed possible consequences of assigning instructional adaptations to students' current and future educational programming and assessment participation options. This information was intended to provide special education service providers, administrators, and policy makers with a clear understanding of the role of these instructional adaptations for improving accessibility of educational environments.

It is important to note, however, that instructional adaptations should be viewed in the larger context of educational programming

Case example	
<p>Maria is a 15 year-old, female student, in the eighth grade. Maria repeated first grade because she struggled to learn basic concepts of reading. Maria continues to have difficulty with decoding and math computation, but she excels on comprehension and problem-solving tasks. Maria's poor decoding skills often impact her work in other classes, as decoding difficulties make it hard for her to understand her assignments. In addition, her teachers report that she is often off task, doesn't complete assignments, and is getting further behind in her studies. Maria works well with the teacher in the learning center, but has trouble following directions from other school staff. In addition, Maria has recently been identified as having a behavior disorder.</p>	
Possible adaptations to instruction	
Differentiated instruction strategies	<p>Provide opportunities for flexible grouping (whole class, small groups, pairs, etc.)</p> <p>Explicitly state learning goals and objectives for each lesson</p> <p>Support and incorporate background knowledge</p> <p>Maintain clear and consistent classroom rules and expectations</p> <p>Provide opportunities for student choice (materials, assignments, partners, etc.)</p>
Accommodations	<p>Peer-assisted learning during math problem-solving tasks; reading partner during independent reading tasks</p> <p>Read aloud the directions and expectations for class assignments</p> <p>Highlight key words or ideas</p> <p>Provide notes or outlines for class lectures or presentations; allow student to audio-record presentations</p> <p>Break large assignments into smaller tasks</p> <p>Teach self-monitoring strategies for on-task behavior; scaffold with reminders</p>
Modifications	<p>Read aloud (for decoding/comprehension tasks)</p> <p>Calculator (for computation tasks)</p> <p>Allow student breaks with learning center teacher (as needed/requested)</p>

Fig. 7.2 Case example illustrating integrated instructional supports

available to students with disabilities. Instructional adaptations can improve access to grade-level content but must be considered as an integral part of the students' educational opportunities. Instructional adaptations are neither a "silver bullet" that should be viewed as the only solution to improving access nor should they be cast aside as supplemental resources that are only used by the special education professional.

References

- Browder, D. M., Spooner, F., Wakeman, S., Trela, K., & Baker, J. (2006). Aligning instruction with academic content standards: Finding the link. *Research and Practice for Persons with Severe Disabilities*, 31(4), 309–321.
- Christensen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- DeStefano, L., Shriner, J. G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessment. *Exceptional Children*, 68(1), 7–22.
- Elliott, S. N., Kurz, A., Beddow, P., & Frey, J. (2009). *Cognitive load theory: Instruction-based research with applications for designing tests*. Presentation at the annual convention of the National Association of School Psychologists, Boston, MA.
- Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. C. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity.

- American Educational Research Journal*, 34, 174–206.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., Phillips, N. B., & Karns, K. (1995). General educators' specialized adaptation for students with learning disabilities. *Exceptional Children*, 61, 440–459.
- Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., et al. (2008). *Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades. A practice guide* (NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <http://ies.ed.gov/ncee/wwc/>
- Hall, T. (2002). *Differentiated instruction*. Wakefield, MA: National Center on Accessing the General Curriculum. Retrieved on February 1, 2009 from http://www.cast.org/publications/ncac/ncac_diffinstruc.html
- Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *The Journal of Special Education*, 32(3), 175–183.
- Ketterlin-Geller, L. R., Alonzo, J., Braun-Monegan, J., & Tindal, G. (2007). Recommendations for accommodations: Implications of (in)consistency. *Remedial and Special Education*, 28(4), 194–206.
- Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice*, 42, 212–218.
- La Marca, Paul M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7(21). Retrieved on January 31, 2010 from <http://PAREonline.net/getvn.asp?v=7&n=21>
- Maccini, P., & Gagnon, J. C. (2006). Mathematics instructional practices and assessment accommodations by secondary special and general educators. *Exceptional Children*, 72(2), 217–234.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38, 43–52.
- National Center for Education Statistics. (2009). *The Nation's Report Card: Mathematics 2009 (NCES 2010-451)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- O'Connell, M., Lieberman, L., & Petersen, S. (2006). The use of tactile modeling and physical guidance as instructional strategies in physical activity for children who are blind. *Journal of Visual Impairment & Blindness*, 100(8), 471–477.
- Phillips, S. (1999, June). *Needs, wants, access, success: The stone soup of accommodations*. Presentation at the 29th annual Council of Chief State School Officers National Conference on Large-Scale Assessment, Snowbird, UT.
- Rose, D., & Dalton, B. (2009). Learning to read in the digital age. *Mind, Brain, and Education*, 3(2), 74–83.
- Thompson, S. J. (2005). An introduction to instructional accommodations. *Special Connections*. Retrieved on May 1, 2006 from <http://www.specialconnections.ku.edu/cgi-bin/cgiwrap/speconn/main.php?cat=instruction§ion=main&subsection=ia/main>
- Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved on November 1, 2006 from <http://education.umn.edu/NCEO/OnlinePubs/Technical42.htm>
- Tindal, G. (1998). *Models for understanding task comparability in accommodated testing*. Washington, D. C.: Council of Chief State School Officers.
- Tomlinson, C. A. (1999). *How to differentiate instruction in mixed-ability classrooms*. Alexandria, VA: ASCD.
- University of Kansas. (2005). *Special Connections: Connecting Teachers to Strategies that Help Students with Special Needs Successfully Access the General Education Curriculum*. Retrieved on February 1, 2010 from <http://www.specialconnections.ku.edu/cgi-bin/cgiwrap/speconn/index.php>
- U.S. Department of Education, Office of Elementary and Secondary Education. (2001). *No Child Left Behind*. Washington, DC: [AUTHOR].
- U.S. Department of Education, Office of Special Education and Rehabilitative Services. (2004). *Individuals with Disabilities Education Act*. Washington, DC: [AUTHOR].
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states: Research monograph no. 18*. Madison, WI: National Institute for Science Education, University of Wisconsin-Madison.
- Wilkins, A. J., Lewis, E., Smith, F., Rowland, E., & Tweedie, W. (2001). Coloured overlays and their benefit for reading. *Journal of Research in Reading*, 24, 41–64.
- Ysseldyke, J., Thurlow, M., Bielinski, J., House, A., Moody, M., & Haigh, J. (2001). The relationship between instructional and assessment accommodations in an inclusive state accountability system. *Journal of Learning Disabilities*, 34, 212–220.

Test-Taking Skills and Their Impact on Accessibility for All Students

8

Ryan J. Kettler, Jeffery P. Braden,
and Peter A. Beddow

Do you remember the first time in school that a peer or teacher advised you to use some type of strategy, which seemed to have nothing to do with the subject being tested, but would likely improve your score? This advice, which may have been to answer all of the easy questions first and then go back to the harder questions, to select *false* whenever the word *always* or the word *never* appeared in a true or false question, or to choose the longest answer of the options on a multiple-choice question, was ostensibly intended to help you attain a higher score on the test. Another outcome of this advice was that it increased your wisdom in the domain of taking tests, so that you would be better able to show what you knew and could do in content areas like reading and mathematics. This chapter explains the relationship between access in educational testing, test-wiseness, and validity; provides a brief history of frameworks and findings related to test-taking skills; explores the interaction between instruction of these skills and more recently embraced methods for increasing access to tests [e.g., accommodations, modifications, computer-based tests (CBTs)]; and discusses implications of this relationship for test development and instruction.

R.J. Kettler (✉)
Department of Special Education, Peabody College
of Vanderbilt University, Nashville, TN 37067, USA
e-mail: ryan.j.kettler@vanderbilt.edu

Access and Test-Wiseness

Access in the context of educational testing refers to the opportunity for a student to demonstrate what she or he knows on a test. It is the opportunity to determine how proficient the student is on the target construct (e.g., reading, mathematics, science) and can be conceptualized as an interaction between the individual and the test (Kettler, Elliott, & Beddow, 2009). While it is often easier to think about access in terms of the barriers that block it (e.g., confusing visuals on a reading test, high reading load on a mathematics test, information spread across multiple pages on a test of anything other than working memory), increased access allows tests to be more reliable and have greater construct validity. Reliability is the consistency of scores yielded by a measure, and construct validity is the degree to which the scores represent the characteristics or traits that the test is designed to measure. Both reliability and construct validity are higher when the proportion of variance within a set of test scores that is construct relevant is maximized. Construct-relevant variance is based on variance in the actual characteristic or trait being measured. The remainder of the variance in test scores is construct-irrelevant variance, which is variance that emanates from other sources, including the aforementioned barriers to access. Reducing barriers to access decreases construct-irrelevant variance, thereby increasing the reliability and construct validity of the scores.

One way to reduce barriers to access is to develop tests or testing situations that minimize the need for access skills. Access skills are those which are necessary for a student to take a test but which are not a part of the construct being tested. If the format of a mathematics test includes lengthy word problems, the ability to read is an access skill for that test. Having the test read aloud to a student would remove reading ability as an access skill, would decrease the variance based on reading ability that would be reflected in scores from the test, and would increase the reliability and construct validity of the test scores.

Another way to reduce barriers to access is to train students in access skills. This training is intended to decrease the degree to which students vary in their knowledge of access skills, therefore reducing the contribution of that variance to the test scores. Test-wiseness is an access skill for tests designed to measure academic achievement in a variety of content areas, and teaching test-taking skills to increase this wiseness is intended to reduce the differences among students on this construct that are reflected in scores intended to indicate abilities in reading, mathematics, and so on.

Millman, Bishop, and Ebel (1965) define *test-wiseness* as a respondent's "capacity to utilize the characteristics and formats of the test and/or test-taking situation to receive a higher score" (p. 707). Test-taking skills (or strategies) are taught with the intent that they will improve test-wiseness, resulting in higher scores, but more importantly resulting in scores from which more valid inferences can be drawn. Without explicit instruction, students naturally vary in their degrees of test-wiseness. While some have a very high capacity to use the test and the testing situation to show what they know, others are confused by the characteristics (e.g., timed versus untimed, penalties for guessing versus no penalties) and format of the test (e.g., multiple choice versus constructed response, online versus paper and pencil), and cannot understand how to show what they know. Instruction in test-taking skills may be able to reduce this variance, or at least reduce the degree to which this variance affects the test scores, by increasing test-wiseness

until most or all respondents have reached a theoretical threshold necessary to access the test. Variance in the access skill above this threshold would not contribute to variance in the test scores. This threshold hypothesis requires further examination.

Threshold Hypothesis

To clarify the aforementioned point about a threshold ability level for an access skill, we return to the example of reading as an access skill for a mathematics test. A mathematics test with word problems written at a sixth-grade reading level and with sixth-grade mathematics content would be difficult to access for a student who reads at a fourth-grade level. A very powerful intervention that allows the student to read at an eighth-grade level would in turn allow the student to better show what she knows on the mathematics test. However, a second intervention that improves the student's reading ability from an eighth-grade level to a twelfth-grade level would not further improve her score on the math test, because she would already be beyond the threshold ability level necessary to access the test. If the student did not know how to solve the mathematics problems, no increase in reading ability beyond the threshold would help. Table 8.1 depicts the relations between reading ability and the construct being measured on the test in this hypothetical example.

Test-wiseness is theoretically analogous to reading ability from the example, in that training students beyond a certain threshold would reduce its influence over scores on an achievement test, and therefore increase the degree to which scores on the test reflect the intended construct (e.g., reading, math, science). However, there is insufficient research to support this threshold hypothesis, and so it is not clear whether test-taking

Table 8.1 Relationship between reading ability and construct indicated by math test scores in example

Reading ability	Construct indicated
Above threshold	Mathematics
Below threshold	Reading

skill training would be viewed as “leveling the playing field” for students with and without disabilities, or might actually increase gaps in test performance between them. Some research using students with learning disabilities (e.g., Dunn, 1981) found that test-wiseness training moderately improved scores on reading tests and that those improvements were retained over a 2-week period, but the design of the study was unable to establish whether training “leveled the playing field” relative to students without disabilities.

Whether test-taking skills influence test performance uniformly across all levels of ability, or if there is a threshold after which skills no longer improve performance, affects how one views teaching test-taking skills to students with disabilities as a method of providing access. For example, eyeglasses are widely viewed as a fair accommodation for individuals with visual acuity difficulties, because the eyeglasses provide a threshold of access equal to that enjoyed by students without visual acuity difficulties. This is because additional correction beyond the threshold needed to decode text does not provide an advantage to users; put another way, students without visual acuity problems would be welcome to use eyeglasses – but would not benefit from their use. In contrast, extra time is a controversial accommodation because there is no clear threshold after which additional time does not enhance performance. Consequently, students with and without disabilities alike benefit from extra time on a test. In fact, students without disabilities may benefit more from extra time than their peers with disabilities (Elliott & Marquart, 2004).

The question of whether test-taking skills function as a threshold or a continuous influence on performance is relevant. If test-taking skills have a threshold beyond which additional strategy application or training does not improve test performance, training students to that level of test-taking strategy proficiency would be viewed as both effective and fair. It would be effective, in that it would eliminate the influence of test-taking skills on scores for students with (and without) disabilities, thus improving the validity of the assessment for both groups. It would be

fair, in that students who already have test-taking skills would not benefit from additional training once the threshold is reached, whereas those who lacked the skills would gain equal access through training.

Conversely, if test-taking skills have a continuous effect across the performance domain, the effectiveness and fairness of teaching test-taking strategies is called into question. First, teaching test-taking skills may not be effective, as students with more or better strategies may benefit more from training than students with fewer or poorer strategies (i.e., training could widen, rather than decrease, test performance differences between more and less test-wise students). Second, teaching test-taking strategies could be unethical, because there would be no clear point at which the playing field was level for students with and without disabilities (e.g., would one withhold test-taking skill instruction from some students but not others, and if so, how would one decide at which point students should or should not receive the training?).

A third possibility regarding test-taking skills training is that it might improve test scores for students with disabilities, but nonetheless produce larger lost opportunity costs for students without disabilities, compared to those with disabilities. In other words, the time taken to teach students with disabilities test-taking skills may improve their scores, but it may not improve them as much as devoting the same amount of time to instruction in test content. Some evidence (Ehrmann, 2001) suggests that is the case for students without disabilities – that is, training using an established test-taking program *lowered* scores on some subject matter tests, but it appeared to improve scores on other subject matter tests. Therefore, careful consideration of interactions between subject matter knowledge, test design, and test-taking skills should be considered in deciding whether to provide test-taking skills training to any students – especially those with disabilities.

The literature on test-taking skills suggests the answer to the threshold question is dependent on three factors: the design of the test, the specific test-taking skills, and the test-taker’s knowledge

of tested material. In general, the better designed the test, the less that performance is influenced by test-taking skills. For example, scores on multiple-choice tests that use only plausible distractors (incorrect response options), simple item stems, and positive wording are generally resistant to test-taking skills; performance on tests that do not incorporate these strong design features is more susceptible to test-taking skills (Rogers & Harley, 1999; Scruggs & Mastropieri, 1988). Although there is little direct research on the subject, it is logical to infer that large-scale, professionally developed tests are more likely to incorporate strong design features, meaning the influence of test-taking skills would be limited to approaches such as time management or scanning the test for easy items before tackling harder items, and therefore would be more likely to have a threshold beyond which additional improvements in strategies would be unlikely to improve test performance.

The threshold hypothesis has received limited research but remains an important issue in the instruction of test-taking skills. Next we review other important historical frameworks and findings in the area of test-taking skills and test-wiseness.

Frameworks and Findings

Millman et al. (1965) developed the most commonly cited theoretical framework for empirical studies of test-wiseness. The authors characterized their framework as narrow in that it did not include factors related to mental or motivational state, and it was restricted to objective tests of achievement and aptitude. There were no published empirical studies of test-wiseness prior to the development of this framework. Millman et al. (1965) based their framework on principles of test construction, advice for taking tests, and the theory that test-wiseness was one source of variability in test scores that cannot be attributed to item content or random error.

Millman et al. (1965) divided their framework of test-wiseness principles into two main categories. The first category, elements which

are independent of the test constructor or test purpose, included four subdivisions: (a) time-using strategy, (b) error-avoidance strategy, (c) guessing strategy, and (d) deductive reasoning strategy. The time-using strategies are specific to timed tests and are intended to avoid losing points because of poor use of time, rather than lack of knowledge of the tested content. The error-avoidance strategies are also focused on avoiding the loss of points for reasons other than lack of knowledge, with the source in this case being carelessness. The guessing strategies are intended to help respondents gain points for responses made completely randomly, and the deductive reasoning strategies are focused on helping gain points when the respondent has part of the necessary knowledge, but does not know the correct answer. Table 8.2 lists Millman et al.'s (1965) elements which are independent of the test constructor or test purpose.

Millman et al.'s (1965) second category, elements dependent upon the test constructor or purpose, includes strategies that require some knowledge of the specific test, constructor, or purpose. The two subdivisions of this category are (a) intent consideration strategy and (b) cue-using strategy. While the intent consideration strategies allow the respondent to avoid being penalized for anything other than a lack of content knowledge, the cue-using strategies are focused on the respondent gaining points when a specific answer is not known. Table 8.3 lists Millman et al.'s (1965) elements dependent upon the test constructor or test purpose.

The effect of coaching on test performance has received more attention than has test-wiseness. Bangert-Drowns, Kulik, and Kulik (1983) published a meta-analysis of 30 controlled studies of the effects of coaching programs on achievement test results. Coaching programs were defined as those in which respondents "are told how to answer test questions and are given hints on how to improve their test performance" (p. 573). The authors found an average effect size of 0.25 in favor of respondents who were coached versus respondents from a control group. The average effect size for short programs ($M = 3.8$ h) was 0.17, compared to an average effect

Table 8.2 Millman et al.'s (1965) elements independent of test constructor or test purpose

A. Time-using strategy	
1.	Bring to work as rapidly as possible with reasonable assurance or accuracy
2.	Set up a schedule for progress through the test
3.	Omit or guess at items (see I.C. and II.B.) which resist a quick response
4.	Mark omitted items, or items which could use further consideration, to assure easy relocation
5.	Use time remaining after completion of the test to reconsider answers
<hr/>	
B. Error-avoidance strategy	
1.	Pay careful attention to directions, determining clearly the nature of the task and the intended basis for response
2.	Pay careful attention to the items, determining clearly the nature of the question
3.	Ask examiner for clarification when necessary, if it is permitted
4.	Check all answers
<hr/>	
C. Guessing strategy	
1.	Always guess if right answers only are scored
2.	Always guess if the correction for guessing is less severe than a "correction for guessing" formula that gives an expected score of zero for random responding
3.	Always guess even if the usual correction or a more severe penalty for guessing is employed, whenever elimination of options provides sufficient chance of profiting
<hr/>	
D. Deductive reasoning strategy	
1.	Eliminate options which are known to be incorrect and choose from among the remaining options
2.	Choose neither or both of two options which imply the correctness of each other
3.	Choose neither or one (but not both) of two statements, one of which, if correct, would imply the incorrectness of the other
4.	Restrict choice to those options which encompass all of two or more given statements known to be correct
5.	Utilize relevant content information in other test items and options

Note: Adapted from Millman et al. (1965, pp. 711–712). Reprinted by Permission of SAGE Publications

Table 8.3 Millman et al.'s (1965) elements dependent on the test constructor or purpose

A. Intent consideration strategy	
1.	Interpret and answer questions in view of previous idiosyncratic emphases of the test constructor or in view of the test purpose
2.	Answer items as the test constructor intended
3.	Adopt the level of sophistication that is expected
4.	Consider the relevance of specific detail
<hr/>	
B. Cue-using strategy	
1.	Recognize and make use of any consistent idiosyncrasies of the test constructor which distinguish the correct answer from incorrect options
a.	She or he makes it longer (shorter) than the incorrect options
b.	She or he qualifies it more carefully, or makes it represent a higher degree of generalization
c.	She or he includes more false (true) statements
d.	She or he places it in certain physical positions among the options (such as in the middle)
e.	She or he places it in a certain logical position among an ordered set of options (such as the middle of the sequence)
f.	She or he includes (does not include) it among similar statements, or makes (does not make) it one of a pair of diametrically opposite statements
g.	She or he composes (does not compose) it of familiar or stereotyped phraseology
h.	She or he does not make it grammatically inconsistent with the stem
2.	Consider the relevancy of specific detail when answering a given item
3.	Recognize and make use of specific determiners
4.	Recognize and make use of resemblances between the options and an aspect of the stem.

Note: Adapted from Millman et al. (1965, p. 712). Reprinted by Permission of SAGE Publications

size of 0.43 for longer programs ($M = 15.1$ h). Bangert-Drowns et al. (1983) also found that a logarithmic relationship between length of coaching program and effect size was a better fit than the linear relationship. This finding indicated that although longer programs were more effective, the returns would be diminished when adding time to programs that were already fairly long.

Teaching of test-wisness strategies was viewed as a subset of coaching in Bangert-Drowns et al. (1983). No significant differences were found between the average effects of the 21 studies that included a focus on test-wisness and the 9 studies that did not. Elements that were identified as more important included drill and practice on items and direct teaching of content.

The clear impact of coaching, which often includes the teaching of test-taking skills, makes relevant any concerns about the ethical implications of spending classroom time attempting to increase test-wisness. One such concern is that when the stakes related to a test are high, teachers are likely to focus on teaching to the test, rather than teaching the broader content area (Mehrens, 1989). Mehrens and Kaminski (1989) identified a seven-point continuum between the ideal of solely teaching content and the alternative of focusing instruction entirely on improving test performance. Figure 8.1 depicts this test preparation continuum.

Point 1 on Mehrens and Kaminski's (1989) continuum, which involves giving general instruction on district objectives without referring to the test, is always considered ethical. Teaching of test-taking skills is also typically considered ethical, and Mehrens (1989) indicated that doing so is worthwhile in that it takes little instructional time and can keep students who know the content from making errors. The line between ethical and unethical practice is typically considered to be somewhere among Points 3 (providing instruction based on objectives a variety of tests measure), 4 (providing instruction specifically on the tested objectives), and 5 (providing instruction on tested objectives and in tested format). Points 6 and 7 refer to practice on parallel forms of a test or on the test itself and are never considered to be ethical

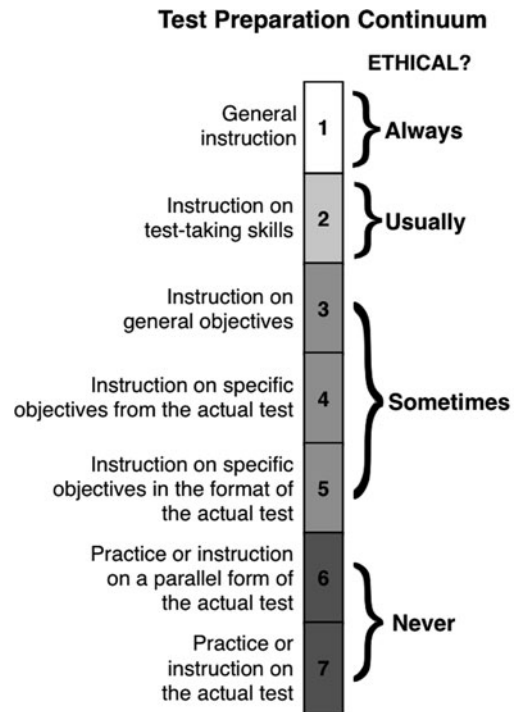


Fig. 8.1 A depiction of Mehrens and Kaminski's (1989) continuum of test preparation

practice. The danger in teaching to the test is that it limits legitimate inferences from those that are based on a broad domain to those that are based on performance on a specific test, and it is rare that a test is given for the sake of the latter form of inferences. Mehrens (1989) wrote that for the level of inference typically desired, the line between ethical and unethical test preparation is between Point 3 and Point 4.

In a comprehensive review of the issues surrounding test preparation, Crocker (2005) identified an increasingly mobile society and generalizability to real-world skills as reasons that test-taking skills have become increasingly important. With students moving across state and national borders to attend college or find employment, a common metric such as achievement tests from which valid inferences can be made is critical. Also, these tests have incorporated short essays and performance tasks, and the access skills for these types of items can be helpful in other situations. Crocker (2005) indicates that test preparation "encompasses not only study of content

from the domain of knowledge sampled by the assessment, but also practicing the skills that will allow students to demonstrate their knowledge on various types of assessment exercises” (p. 161). The author recommends teaching *for* assessment, rather than teaching to the test, and identified four essential elements of this practice: (a) challenging core curriculum, (b) comprehensive instruction in the curriculum, (c) developing test-taking skills, and (d) adherence to ethical guidelines. The challenging core curriculum includes all of the content that is intended to be learned, rather than the subset that is likely to be sampled on the assessment. Comprehensive instructional practices involve spending most of the time teaching the content and trusting that student achievement will be reflected by the assessment. Developing test-taking skills includes many of the strategies identified by Millman et al. (1965), incorporated into year-long classroom instruction.

Crocker (2005) outlined four broad criteria for meeting the fourth essential element of teaching for assessment: adherence to ethical guidelines. The first of these criteria is academic ethics, the idea that test preparation should be consistent with the ethics of education, including that cheating and stealing the work of others is inappropriate. The second criterion is validity, indicating that students should be able to show what they know and are able to do through the test. The third of these criteria is transferability, the idea that test preparation should be in those domains that are applicable to a broad range of tests. The fourth criterion is educational value, meaning that test preparation that leads to increases in scores should also lead to increases in student mastery of the content. A test preparation practice that meets these four criteria is likely to be an appropriate method of improving test-wiseness and assessment.

Following the suggestions of Crocker (2005), as well as the strategies, frameworks, and findings that influenced this work, provides students with greater access to show their knowledge in a testing situation. Next we examine how test-wiseness and test-taking skills interact with other methods of improving access to appropriate assessment.

Test-Taking Skills and Other Methods of Increasing Access

Since the passage of the Individuals with Disabilities Education Act (IDEA) in 1997, mandating that all students be included to the greatest extent possible within large-scale accountability systems, assessment accommodations have been the most common method for helping students with disabilities overcome access barriers inherent in achievement tests. *Assessment accommodations* are changes in the standard assessment process made because an individual’s disability requires changes for the test to be a measure from which valid inferences can be drawn (Elliott, Kratochwill, & Schulte, 1999; Kettler & Elliott, 2010). More recently the final regulations of the No Child Left Behind (NCLB) Act of 2001 have provided for changes to the content of a test, made for a group of persistently low-performing students with disabilities, rather than on an individual basis (U.S. Department of Education, 2007a, 2007b). *Modifications* are the result of “a process by which a test developer starts with a pool of existing test items with known psychometric properties, and makes changes to the items, creating a new test with enhanced accessibility for the target population” (Kettler et al., in press). Both assessment accommodations and modifications can interact with test-taking skills in ways that should be considered by test developers.

Test-Taking Skills and Accommodations

Although there exists no research addressing interactions between test-wiseness and assessment accommodations, it would be wise to anticipate whether some common accommodations might have the unintended consequence of inhibiting appropriate use of test-taking skills. For example, the accommodation of extra time is intended to give respondents who work slowly the time they need to access test content. However, additional time might inadvertently encourage use of dysfunctional strategies (e.g., excessive correction or second-guessing

of answers), or inhibit use of functional strategies (e.g., students might not manage their time well because they lack a sense of urgency or they fail to deploy strategies because they become physically fatigued). Likewise, changes in test item presentation (e.g., large print, reading test items aloud to students, magnification) may encourage linear progression through the test and may discourage adaptive strategies, such as scanning the test for easy items, skipping hard problems, or returning to parts of the test that were skipped. Likewise, having a mathematics or science test read aloud to a student might discourage spending more time on more difficult sections of a passage, underlining the most important parts, or returning to the passage to find an answer. In these ways, assessment accommodations may interact with test-taking skills; therefore, test-wiseness training should be deployed, practiced, and evaluated well in advance of the actual assessment session so that unintended consequences can be recognized and corrected.

Test-Taking Skills and Modifications

Many common modifications are designed to reduce the impact of variance in test-taking skills on variance reflected in scores from achievement tests. For example, Millman et al. (1965) suggested eliminating any options which are known to be incorrect, prior to answering a multiple-choice question. One common modification technique is elimination of the least-plausible distractor, so that the total number of answer choices typically changes from four to three. Because the distractor that is removed via this modification is usually one that is eliminated by most wise test-takers, the field is leveled by automatically removing that option for all students. Another skill suggested by Millman and colleagues is to look for the correct option to be longer or shorter than the incorrect options. A common modification strategy of making all distractors similar in length nullifies the advantage created by this test-taking skill. Ultimately, scores from a test that is perfectly accessible will not contain any variance

based on differences in test-wiseness. Tests in the future may be closer to this ideal and are also more likely to be delivered via computer; implications of this latter shift are discussed next.

Computer-Based Test-Wiseness

Since the inception of digital scoring in the 1930s, CBTs have become increasingly ubiquitous in the arena of student assessment. In 2009, Tucker reported over half of U.S. states used computers to deliver some portion of their federally mandated tests. Despite this upward trend, the majority of current CBTs remain mere transpositions of paper-and-pencil tests (PPTs). Most CBTs consist of item sets that look and function the same as those used in their PPT forebears, with the overlay of an interface that permits test-takers to use a mouse or keyboard to select responses. In these cases, the advantages of the computer over the traditional test booklet primarily benefit the test administrators. This is not to say that computers cannot or will not be used to improve testing in other ways, such as to reduce test length and increase score accuracy; indeed, as test developers utilize innovations in computer technology and psychometrics, the accuracy and efficiency of delivery and scoring likely will increase. Presently, however, in the arena of large-scale and/or commercial testing programs, little has changed in terms of item and test presentation, and most of the traditions of PPT have been carried forward in their CBT descendants. Thus, on the surface it appears little difference exists between the influence of test-taking skill use on PPTs compared to CBTs.

To the contrary, while the threat to inferential validity from the kinds of test-taking skills discussed elsewhere in this chapter may remain unchanged with the migration to CBTs, other aspects of CBTs may in fact transmute the problem to another theater. In [Chapter 15](#), Russell discusses the many potential sources of error introduced by computer tests due to access issues that may not exist in other types of tests, issues that

likely are particularly problematic for students with special needs. Simultaneously, he notes that certain innovations in CBTs offer opportunities for addressing access concerns that have long been problematic for users of PPTs. He describes recent CBT interfaces that successfully reduce the demand for scribes, readers, and other human testing accommodation delivery agents, interfaces that are preferred by many test-takers over their traditional alternatives. Specifically, he refers to *test alterations* made permissible by computers. These alterations include (a) altered presentations (e.g., changing font size, reducing the amount of content presented on a page), (b) altered interactions (e.g., pacing, content masking, scaffolding), (c) altered response modes (speech-to-text interfaces, touch screens), and (d) altered representations (tactile representations of visuals, translations into different languages).

These access tools are in no way included with the intention of permitting test-takers to increase their scores via strategy use. However, to the degree test access tools (e.g., those recommended by Russell) find their way into CBTs, the issue of test-wiseness must be translated to a new playing field. The tools that Russell recommends, particularly those that may be integrated into test interfaces and made universally available to the entirety of the test-taker population, potentially offer a new set of “opportunities” for test-wiseness to cause construct-irrelevant variation in test scores. Indeed, test-takers who learn to use these tools appropriately and efficiently may have an advantage over those who fail to learn them or those who use them improperly or inefficiently.

Commensurate with the putative accessibility problems introduced by the migration of assessment to CBTs, the potential exists for measurement error to increase as a result of efforts to promote solutions to these problems. Indeed, collinearity between access tool use and test-taker demonstration of construct-relevant skills may obscure the meanings of test scores and threaten the validity of score inferences. It is critical that efforts to eliminate access barriers for one group of test-takers do not inadvertently create advantages for others.

Suggestions for Developers of Computer-Based Tests

Developers of CBTs should ensure, to the extent possible, test-delivery systems – including integrated access tools – are optimally accessible for all test-takers. To this end, developers should consider applying the following recommendations from Schneiderman’s (1997) work on effective interface design: (a) strive for consistency, (b) cater to universal usability, (c) offer informative feedback, (d) design dialogs to yield closure, (e) prevent errors, (f) permit easy reversal of actions, (g) support internal locus of control, and (h) reduce short-term memory load. The integration of these principles into the design of test interfaces, with particular attention to features designed to reduce specific access barriers, likely will result in accessible test events for more students (see Chapter 9, this volume).

Even after integrating these principles into the test-delivery system, however, there likely will remain some test-takers for whom a user interface that integrates best practices of usability and user-friendliness will create cognitive overload based on the *unfamiliarity* of the interface alone. Test developers, therefore, are advised to build interface training modules into each assessment system to reduce the potential for threats to validity based on variation across the population in *interface expertise*. These modules, which can either be delivered prior to the test event as part of the general curriculum or included in the test event itself, should serve two purposes. Specifically, they should (a) equip all test-takers with the expertise to use the interface at the highest level possible and (b) assess the expertise of each test-taker in the use of the interface.

There are potential problems that arise with the need to incorporate a learning activity into a test. Among these is the fact that many of the test-takers for whom such an activity is necessary tend to perform poorly on tests (and often are accustomed to school failure). For these students, test anxiety is an important concern. The addition of an instructional task to a test event that already provokes negative emotions, in a potentially high-pressure environment, is a decision that should be

approached delicately and with ample attention to the emotional and academic needs of this target student population.

A Lesson from Video Games

In the arena of academic learning, some educators have argued video games are the harbingers of effective computer-based learning tools. Curiously, the notion that video game technologies could offer solutions to learning challenges is plausible, for several reasons. First, due to the ubiquity of video games (e.g., computer games, Sony Playstation, and Nintendo) there is a high probability most students will have some measure of experience with them. Second, the increasing complexity of gameplay mechanics in recent years has compelled developers to integrate mechanics that iteratively train the player to use the features of the game.

Much research has been conducted on the development and use of computer games for instruction (e.g., Gee, 2005; Becker, 2006; Blumberg, 2008). The similarities between instruction and assessment, however, suggest many of the strategies utilized in effective learning games can and should be applied to CBTs. One example is *chain gameplay*, the simple, repeatable process of subtly increasing difficulty that is found in many video games (Sivak, 2009). Sivak explained chain gameplay as follows:

A gameplay chain is any set of interlocking mechanics that must be done together in order to achieve a goal. . . [an idea] somewhat similar to Ian Bogost's theory of Unit Operations. . . Each basic mechanic is used as a building block and seamlessly connects to the next mechanic creating a complex gameplay structure that is far more interesting than the sum of the individual links. . . [helping] to initiate a state of flow for the player. Flow. . . is the feeling of total immersion in an activity (p. 284).

In the context of CBTs that utilize integrated access tool interfaces, chain mechanics may be used as part of training modules to ensure test-takers, like gameplayers, are equipped with the necessary expertise to utilize the tools with a modicum of cognition.

In Curry's (2009) chapter on the design features that comprised the classic 1985 Nintendo video game *Super Mario Bros.* and that have contributed to its enduring popularity among video game critics, the author detailed a number of interface aspects that may prove useful for the development of assessment training modules. Specifically, Curry described four key ingredients that led to *Mario's* success as a game. First, the game was *instantly accessible*. Curry noted that "when someone sits down with a new game, be it a board game, a word game in the newspaper, or videogame, his first question is always, 'ok, what am I trying to do here?'" (p. 14). He wrote that *Super Mario Bros.* immediately made the user aware of the objective of the game simply by eliminating the ability of the user to interact with any elements that did not contribute to the objective (which, in *Mario's* case, was to run to the right.) In the context of assessment training, the user should be presented with a single button, or tool, with limited choices, such that the user invariably accomplishes the first goal and experiences firsthand the primary objective of the test (e.g., to finish the test.) Second, the game was *easy to control*. It had clear rules that, while limiting the user's freedom, offered enough variety so as to avoid monotony and gave the user a sense of empowerment to make his or her way to the objective. Third, the game was *challenging at an appropriate threshold*. It utilized a mechanic similar to Sivak's gameplay chain insofar as the interface equipped the user with initial skills, followed by more emergent abilities to master. Fourth, the game was *overflowing with rewards*. Each advancement to a new level or the equipment with a new tool or skill was accompanied by visceral audio and video that served as rewards.

While these design features may seem a far cry from current CBTs, they should not be disregarded as trivial priorities. Indeed, as the application of universal design principles to assessment systems results in the inclusion of access tools to accommodate a wide variety of student needs, there is increased likelihood that computer interfaces will become more complex. It is critical that computer interfaces with vast libraries of "helpful tools" do not result in disparities between

students who have expertise in the use of these interfaces and those who do not. If some students are able to increase their scores relative to other students on the basis of their strategic use of access tools that are inaccessible to other students, their resulting score inferences will reflect an indistinguishable combination of the level of the target construct they possess and their expertise with the test system.

Practical Implications

The theory and findings on test-wiseness presented have numerous implications for developers and users of tests. The following are five lessons about test-taking skills and accessibility:

Design Assessments to Minimize Test-Taking Skills

Educators who employ sound principles of assessment design tests that minimize the influence of test-taking skills. In other words, rather than try to make test-takers less vulnerable to poor test design, we recommend improving assessments to reduce the construct-irrelevant variance introduced by differences in test-taking skills among respondents (see Rogers & Harley, 1999; Scruggs & Mastropieri, 1988, on how to design tests to reduce the influence of test-wiseness on test scores). This recommendation is extended to producers and users of CBTs, who need to train test-takers to the point that they are competent with the test's interface, so that differences in this familiarity are not reflected in the test scores. As CBTs become more sophisticated and commonplace in large-scale assessment, lessons can also be taken from successful video game design.

Evaluate the Influence of Test-Taking Strategies on a Case-By-Case Basis

Although teachers are good judges of relative student performance (i.e., knowing which students will do better or worse on tests), they are

less effective when predicting test performance for students with disabilities than for peers without disabilities, and they are less effective when predicting student performance on standardized versus teacher-made tests (Hurwitz, Elliott, & Braden, 2007). Given that the effectiveness of test-wiseness training interacts with students' knowledge, test-taking skills, and test design, we believe it will be extremely difficult to anticipate the influence of test-wiseness training on the scores of students with and without disabilities. Therefore, we advocate single-case designs to determine whether a specific training program will work with a given student for a particular type of test. We further recommend that teachers vary at least two instructional interventions – one aimed at test-wiseness and the other aimed at the content knowledge being tested – to evaluate the relative “lost opportunity costs” of devoting instruction to test-wiseness versus instruction of academic content. If there is a consistent difference favoring the test-wiseness training over subject matter instruction, then test-wiseness training should be recommended; otherwise, instruction in academic content would offer fewer ethical issues and produce at least equal effects.

Spend a Lot of Time Teaching Content and a Little Time Teaching Test-Taking Skills

Teaching test-taking skills should not be controversial if the training only takes a little class time, because it can help students avoid making mistakes that might affect their scores but not accurately reflect their achievement. Research indicates that direct instruction of content, along with drilling and practice testing, results in greater effect sizes compared to instruction in test-taking skills (Bangert-Drowns et al., 1983). These findings are reflected by the focus of the countless test preparation sites on the world wide web (e.g., www.usatestprep.com, www.studyzone.org, www.studyisland.com), which focus on practice tests and methods of drilling content, along with minimal reference

to test-taking skills. While spending 15–20 min once per year teaching test-taking skills is not likely to cause concern, including isolated instruction on test-taking skills repeatedly in one's lesson plan at the expense of grade-level content is not an appropriate strategy, and may widen the achievement gap between students without disabilities and students with disabilities.

Consider Interaction with Other Methods of Increasing Access to Tests

The applicability of test-taking skills to a test will vary in great degree based on how accessible its items are. Modifications to a test may enhance it so that the test-taking skills that one would emphasize are no longer relevant. Also, the assessment accommodations that a student uses should be considered along with the test-taking skills, because some of the skills may become counterproductive when combined with accommodations.

Teach for the Assessment

We recommend following Crocker's (2005) suggestion of teaching for the assessment, which includes an emphasis on instruction linked to content standards beyond the subset that is assessed, with faith in the assessment to accurately measure what students know and can do. This policy includes some instruction in test-taking skills, but ultimately prepares students for success in a number of different settings, rather than just the specific testing situation.

Conclusions

Every test has a format, including a delivery system, directions, method of responding, and other features, that if working optimally should provide respondents access to show what they know or can do related to the construct of interest. Test developers should make any and all modifications to a test that will reduce construct-irrelevant variance – connected to comfort with the test's

format – from being included in variance in the final test score. Beyond that commitment, training students in test-taking skills that will bring all respondents above a threshold necessary to remove the construct-irrelevant variance is also justifiable, so long as taking the time to do so does not remove students' opportunity to learn on the construct of interest. The positive impact of teaching test-taking skills appears to be small on a group basis, but it can be meaningful for individuals who might make mistakes on tests that are not related to their achievement on the intended construct. It is for this reason that limited training in test-taking skills can ethically be included within a larger plan of teaching for an assessment and increasing access for all students.

References

- Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1983). Effect of coaching programs on achievement test performance. *Review of Educational Research*, *53*, 571–585.
- Becker, K. (2006). Video game pedagogy: Good games= good pedagogy. *Journal of Design Research*, *5*, 1–47.
- Blumberg, F. C., Rosenthal, S. F., & Randall, J. D. (2008). Impasse-driven learning in the context of video games. *Computers in Human Behavior*, *24*, 1530–1541.
- Crocker, L. (2005). Teaching for the test: How and why test preparation is appropriate. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 159–174). Mahwah, NJ: Lawrence Erlbaum Associates.
- Curry, P. (2009). Everything I know about game design I learned from super Mario Bros. In D. Davidson (Ed.), *Well played 1.0: Video games, value and meaning* (pp. 13–36). Halifax: ETC Press. Retrieved May 8, 2010, from <http://etc.cmu.edu/etcpress/>
- Dunn, A. E. (1981). *An investigation of the effects of teaching test-taking skills to secondary learning disabled students in the Montgomery County (Maryland) public schools learning centers*. Unpublished doctoral dissertation, George Washington University, Washington, DC.
- Ehrmann, C. N. (2001). *The effect of test preparation on students' Terranova scores*. Unpublished master's thesis, University of Wisconsin – Madison, Madison, WI.
- Elliott, S. N., Kratochwill, T. K., & Schulte, A. G. (1999). The assessment accommodations checklist: Facilitating decisions and documentation in the

- assessment of students with disabilities. *Teaching Exceptional Children*, *Nov/Dec*, 10–14.
- Elliott, S. N., & Marquart, A. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children*, *70*(3), 349–367.
- Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-Learning*, *2*, 5–16.
- Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgments of students' test performance. *School Psychology Quarterly*, *22*(2), 115–144.
- Individuals with Disabilities Education Act Amendments, 20 U.S.C. §1400 *et seq.* (1997).
- Kettler, R. J., & Elliott, S. N. (2010). Assessment accommodations for children with special needs. In B. McGaw, E. Baker, & P. Peterson (Eds.), *International encyclopedia of education* (pp. 530–536) (3rd Ed.). Oxford, UK: Elsevier.
- Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, *84*(4), 529–551. doi:10.1080/01619560903240996
- Kettler, R. J., Rodriguez, M. R., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (in press). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*.
- Mehrens W. A. (1989). Preparing students to take standardized achievement tests. *Practical Assessment, Research & Evaluation*, *1*(11). Retrieved May 28, 2010, from <http://PAREonline.net/getvn.asp?v=1&n=11>
- Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless or fraudulent? *Educational Measurement: Issues and Practices*, *8*(1), 14–22.
- Millman, J., Bishop, H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*(3), 707–726.
- Nintendo of America, Inc. (1985). *Super Mario Bros.* (Video game).
- No Child Left Behind Act, 20 U.S.C. §16301 *et seq.* (2001).
- Rogers, W., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to test-wiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*(2), 234–247.
- Scruggs, T., & Mastropieri, M. (1988). Are learning disabled students 'test-wise'? A review of recent research. *Learning Disabilities Focus*, *3*(2), 87–97.
- Shneiderman, B. (1997). *Designing the user interface: Strategies for effective human-computer interaction*. Boston: Addison-Wesley Longman.
- Sivak, S. (2009). Each link is a chain in a journey: An analysis of the legend of zelda: Ocarina of time. In D. Davidson (Ed.), *Well played 1.0: Video games, value and meaning*. (pp. 283–314). ETC Press. Retrieved May 8, 2010, from <http://etc.cmu.edu/etcpress/>
- U.S. Department of Education. (2007a, April). *Modified Academic Achievement Standards: Non-Regulatory Guidance*. Washington, DC: Author.
- U.S. Department of Education. (2007b, revised July). *Standards and assessments peer review Guidance*. Washington, DC: Author.

Part III

Test Design and Innovative Practices

Accessibility Theory: Guiding the Science and Practice of Test Item Design with the Test-Taker in Mind

9

Peter A. Beddow, Alexander Kurz,
and Jennifer R. Frey

Test accessibility is defined as the extent to which a test and its constituent item set permit the test-taker to demonstrate his or her knowledge of the target construct (Beddow, Elliott, & Kettler, 2009). The principles of accessibility theory (Beddow, in press) suggest the measurement of achievement involves a multiplicity of interactions between test-taker characteristics and features of the test itself. Beddow argued achievement test results are valid to the degree the test event controls these interactions and yields scores from which inferences reflect the amount of the target construct possessed by the test-taker. Test score inferences typically are based on the assumption that the test event was optimally accessible; therefore, the validity of an achievement test result depends both on the precision of the test score and the accuracy of subsequent inferences about the test-taker's knowledge of the tested content after accounting for the influence of any access barriers. In essence, the accessibility of a test event is proportional to the validity of test results.

This chapter contains three primary sections. In the first section, we describe the current theoretical state of assessment accessibility. Specifically, we (a) describe the recent focus on universal design (UD; Mace, 1991; Rose &

Meyer, 2002) in education and briefly survey the broad effort to apply universal design principles to educational assessment and (b) introduce accessibility theory (Beddow, in press) and describe how it advances the current research and theory, focusing on the influence of Chandler and Sweller's (1991) cognitive load theory (CLT) and its relevance to the development of accessible tests. In the second section of this chapter, we describe a comprehensive decision-making instrument for evaluating and quantifying the accessibility of test items and applying modifications to improve their accessibility, demonstrating an iterative strategy for modifying test items and defining the various theoretical principles that undergird the process. We conclude with an examination of the relevance of accessibility theory across the educational environment.

Universal Design: The End or the Beginning?

When test-taker characteristics interact with the test such that access barriers are presented, issues of fairness are raised and ultimately the validity of decisions is impacted. Thus, it is critical student characteristics are considered throughout the test design process and test developers strive to integrate test features that promote accessibility during all stages of development. Principles of universal design (UD; Mace, 1997) provide a broad framework to conceptualize these considerations.

P.A. Beddow (✉)
Department of Special Education, Peabody College of
Vanderbilt University, Nashville, TN 37067, USA
e-mail: peterbeddow@gmail.com

In its initial conception, UD was applied within an architectural design context and was conceptualized as designing all buildings and products to be usable by as many people as possible to reduce the need for adaptations or specialized designs for specific populations (Mace, Hardie, & Place, 1996; Mace, 1997). Within the UD framework, products, buildings, and environments should be designed in such a way that they are useful and accessible to people with different abilities and interests (Mace, 1997).

Federal legislation (e.g., The Fair Housing Amendments Act of 1988; The Americans with Disabilities Act of 1990; The Telecommunications Act of 1996; The Assistive Technology Act of 2004) has required programs and public facilities to be accessible to all individuals. The principles of UD state that designs should (a) be useful, appealing, and equitable to people with different abilities; (b) accommodate a wide range of individual preferences and abilities; (c) be easy to understand; (d) communicate necessary information through multiple modes (i.e., visual, verbal, tactile); (e) promote efficient and easy use; and (f) provide appropriate size and space (Mace, 1997). In 2004, with the reauthorization of the Individuals with Disabilities Education Act, the principles of UD were extended to the development and administration of educational assessments.

Based on the principles of UD, assessments should be developed and designed to allow a wide range of students, with varying abilities and disabilities, to participate in the same assessment. Universally designed assessments should eliminate access barriers to the test and allow for more valid inferences about a student's performance. Thompson, Johnston, and Thurlow (2002) proposed seven elements of universally designed assessments: inclusive assessment population, clearly defined test constructs, non-biased test items, open to accommodations, clear directions/procedures, maximally readable and understandable, and maximally legible. By incorporating these elements into the development and design of an assessment, they hypothesized the test should be accessible to as many students as possible. Universally designed tests should

reduce the need for accommodations, but they do not eliminate the need for testing accommodations, as universally designed tests are not necessarily accessible to *all* students (Thompson, Johnstone, Anderson, & Miller, 2005).

Accessibility Theory: The Test-Taker, Not the Universe

The expectation that tests be universally designed is insufficient insofar as the broad charge by UD proponents to integrate universal access tools into assessment instruments lacks, to a large extent, a theoretical foundation with clear relevance to measurement. Figure 9.1 consists of Beddow's (in press) model of test accessibility, which the author has referred to as *accessibility theory*. The left side of the model represents the test event – that is, the sum of interactions between the test-taker and the test. The first column in the test event consists of the skills and abilities possessed – or at least required – by the test-taker during the test event. Each has a parallel in one or more features in a test or test item (the second column.) The right side of the model illustrates the presumed causal pathway among these interactions and the measurement of the target construct, the resulting test score, inferences about the amount of the target construct possessed by the test-taker, and subsequent decisions based on these test score inferences.

Accessibility theory defines five domains of test event interactions: physical, perceptive, receptive, emotive, and cognitive (see also Ketterlin-Geller, 2008, who similarly proposed four interaction categories, including cognitive, sensory, physical, and language). Depending on the design features of a particular test, each of these interactions may require the test-taker to demonstrate an access skill in addition to – or prior to – demonstrating knowledge of the target construct of the test or test item. Tests that demand test-takers possess certain access skills (or even have certain characteristics) to demonstrate the target construct, irrespective of the portion of the test-taker population that possesses

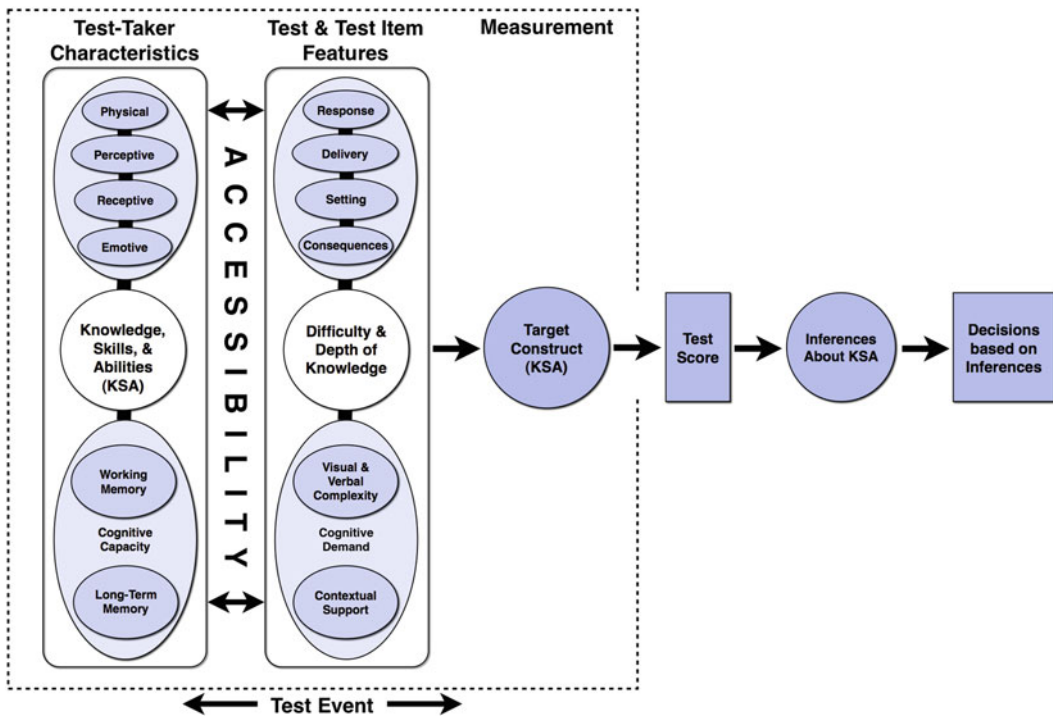


Fig. 9.1 Test accessibility theory. From Beddow (in press). (Reprinted with permission)

these skills or characteristics, inadvertently measure them. Thus, the design objective should not be to create tests that are designed to be *universally accessible*; rather, the goal is to ensure tests are accessible to the specific test-takers for whom the test is intended to be used. This is why it is critical for test developers to have familiarity with the target population of the test.

For instance, Russell (Chapter 15) addresses the accessibility issue of computer-based tests for test-takers who are deaf or have impaired hearing. The access barrier for these test-takers is caused by the perceptive interaction between the audio features of the test and their limited ability to hear the elements of the test that are presented through audio. In addition to the demand for the test-taker to perceive sound, the audio features of the test also generate a receptive interaction during the test event whereby the test-taker must comprehend information via audio. Russell describes a test-delivery system that presents these test-takers with avatars that present the same information in sign language. To the degree this

accessibility tool reduces the influence of hearing ability on the subsequent test scores of deaf test-takers, the interactive effect of this access barrier is reduced and that particular feature of the test is equally accessible for those test-takers compared to test-takers whose hearing is not impaired.

These skills, then, represent *ancillary requisite constructs* (ARCs); that is, in addition to measuring the target construct, the test also measures constructs that may obstruct the stated objective of the test. This is true for the entirety of the test-taker population. Of course, if these ARCs do not present access barriers for any test-taker, their effects are nil and the validity of test score inferences remains uncompromised. Thus, the goal of accessible test design is to control all interactions (i.e., to ensure the total effect of ARCs does not influence the test score of any test-taker) such that, barring any threats to validity apart from accessibility (e.g., low reliability of test scores), the test measures only the target construct.

Cognitive Load Theory: From Teaching to Testing

Of the interaction domains defined in accessibility theory, arguably the one that warrants the most attention in the design of accessible tests and test items is the cognitive domain. In the framework of accessibility theory, we discuss cognitive interactions in terms of limited test-taker working memory (i.e., the immediate availability of useful cognitive resources; Baddeley, 2003; Miller, 1956) and the cognitive demands of the test. CLT (Chandler & Sweller, 1991) provides an intuitive model for understanding these interactions and adjusting test features to control them. Sweller (2010), like most cognitive scientists, disaggregates memory (i.e., cognitive functioning) into long-term memory and working memory. This conceptualization will facilitate our understanding of the various cognitive interactions that may occur during the test event.

CLT primarily has been used to develop instructional tasks to facilitate efficient learning (e.g., Chandler & Sweller, 1991; Clark, Nguyen, & Sweller, 2006; Mayer & Moreno, 2003; Plass, Moreno, & Brunken, 2010). Theorists discuss cognitive load in terms of three categories: intrinsic (i.e., essential cognition for engaging and/or completing the task), germane (i.e., cognition that facilitates the transfer of information into long-term memory), and extraneous (i.e., requisite cognitive demand that is irrelevant to the task.) In spite of its lack of application to educational measurement, CLT offers an array of evidence-based strategies that can be applied to testing (e.g., Kettler, Elliott, & Beddow, 2009; Elliott et al., 2010).

Proponents of CLT recommend designers of instructional materials aim to eliminate extraneous load while maximizing intrinsic load and deliberately managing germane load to enhance knowledge acquisition. This ensures the learner is permitted to allocate his or her cognitive resources to the primary objectives of the task (Sweller, 2010). In testing, the inclusion of extraneous and/or construct irrelevant demands must be addressed at both the test and item levels to

ensure the test yields scores that represent, to the greatest extent possible, a measure of the target construct that is free from the influence of ancillary interactions due to access barriers. To this end, CLT offers a useful lens through which to evaluate and modify tests and their respective items to increase test-taker access to the target construct.

Clark et al. (2006) described a set of 29 principles for facilitating efficient learning according to CLT. Many of these can be directly applied to the design of accessible paper-pencil tests or computer-based tests (Elliott, Kurz, Beddow, & Frey, 2009). Table 9.1 consists of a distillation of several principles of CLT that may offer strategies for managing cognitive load in assessment instruments. These include guidance on the efficient design of visuals, including graphs, charts, tables, and pictures; text economy to reduce information overload; page organization and layout; highlighting and bolding of essential data; avoiding redundant multimodal presentation of material; textual or visual support for high-complexity material; and the use of audio for learners with low prior knowledge. Additionally, Mayer and Moreno's cognitive theory of multimedia learning (2003) integrates many cognitive load effects and applies them to the design of computer-based instructional materials, providing a useful perspective for test designers.

Developing Accessible Test Items: Identify, Quantify, and Modify

A set of tools has emerged from the framework of accessibility theory for examining and rating test items with a focus on increasing access, drawing from two decades of research on CLT. Specifically, two instruments were developed out of the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) and Consortium for Modified Alternate Assessment Development and Implementation (CMAADI) projects aimed at improving accessible tests for students with special needs: the *Test Accessibility and Modification Inventory (TAMI)*; Beddow,

Table 9.1 Application of cognitive load theory guidelines to assessment

Test element	Cognitive load theory concept	Application to testing
Visuals	<p>Eliminate unnecessary visuals, including those included to promote interest:</p> <ul style="list-style-type: none"> • Visual elements can distract attention from essential task demands. <p>Use diagrams to promote some types of learning (e.g., spatial relationships):</p> <ul style="list-style-type: none"> • Using an apt visual can offload cognitive demand to permit the learner to utilize resources for other essential task demands. 	<ul style="list-style-type: none"> • All elements in a test item (i.e., stimulus, stem, answer choices, and visuals) should be viewable on one page. • Avoid text when a simple visual will suffice (i.e., visuals should replace text rather than duplicate it). • Use visuals when spatial reasoning is necessary (unless the target construct requires the test-taker generating a visual from text). • When necessary, use visuals to facilitate understanding. • Use minimal amount of text to facilitate understanding of complex visuals (e.g., labels).
Page layout	<p>Avoid <i>split attention</i> and <i>redundancy</i> effects:</p> <ul style="list-style-type: none"> • Integrating knowledge from two sources using the same modality (e.g., two visuals) is cognitively demanding. • Including redundant information increases cognitive load. <p>Reduce need for <i>representational holding</i>:</p> <ul style="list-style-type: none"> • Maintaining information in working memory for use in another physical location, such as on another page or screen, can limit the resources available for other task demands. 	<ul style="list-style-type: none"> • Use one integrated visual rather than two similar visuals. • Text should not be added to self-explanatory visuals. • Information should not be presented redundantly in both text and visuals. • Text and related visuals should not be separated on a page, on different pages, or screens. • Integrate explanatory text close to related visuals on pages and screens. • Integrate requisite reference sheets or other material into the item so the test-taker is not required to hold material from other pages in working memory to respond.
Item text	<p>Pare content down to essentials:</p> <ul style="list-style-type: none"> • Complex text demands high working memory capacity; • Simplifying text and adding signals help prevent cognitive overload. <p>Attend to the complexity of text to manage cognitive load for readers with various levels of expertise:</p> <ul style="list-style-type: none"> • Write highly coherent texts for low knowledge readers, requiring minimal inferences; • Avoid redundant information for high knowledge readers. 	<ul style="list-style-type: none"> • Eliminate redundant but related technical content. • Eliminate unnecessary language to reduce reading load. • Vocabulary and sentence structure should be as simple as possible, if the target construct permits. • Use bold font and/or <u>underlining</u> to highlight words that are essential for responding. • Whenever possible, permit test-takers to access definitions of unfamiliar words. • Item stems should be written as simply as possible. • Only use headers and titles to cue test-takers to the content and types of items. • Use caution in breaking up text for low knowledge readers (e.g., by a visual).
Audio	<p>Avoid overloading working memory with the use of audio to support learning:</p> <ul style="list-style-type: none"> • The <i>phonological loop</i> and the <i>visual-spatial loop</i> are two theoretical components of the working memory system. 	<ul style="list-style-type: none"> • Do not add audio when visuals are self-explanatory. • Don't describe visuals with words presented in both text and audio narration. • Sequence on-screen text after audio to minimize redundancy. • Avoid audio narration of lengthy text passages when no visual is present. • When a visual requires further explanation, integrate text with audio to avoid the split-attention effect. • If possible, use audio to provide explanatory words for complex visuals rather than add written definitions.

Table 9.1 (continued)

Test element	Cognitive load theory concept	Application to testing
Delivery system	<p>Teach system components before teaching the full process:</p> <ul style="list-style-type: none"> • Ensure learners master the steps of a procedure before they are required to perform it as a whole; <p>Give learners control over pacing:</p> <ul style="list-style-type: none"> • When pacing must be instructional controlled, cognitive load must be managed through the design of instruction and materials. <p>Use <i>completion examples</i> to promote learner processing of examples:</p> <ul style="list-style-type: none"> • Use hybrid practice problems and worked examples. Essentially, the first step or steps is/are done for the learner. Afterward, the learner completes the steps independently. 	<ul style="list-style-type: none"> • Train test-takers in the test-delivery system prior to the test date. • Teach supporting knowledge, separate from teaching procedure steps. • Test navigation should be trained to mastery before the test-taker is required to use the test-deliver system. • Computer-based tests should permit the test-taker to navigate to any item during the test session, save progress, and take breaks when needed. An on-screen progress indicator and/or clock is recommended. • Test training should result in test-taker mastery of the test-delivery system. Item examples should be included. • If several types of items are included in a test, test training should include examples of each type.

Kettler, & Elliott, 2008) and the *TAMI Accessibility Rating Matrix (ARM)* (Beddow et al., 2009). Both instruments were designed with the dual purpose of documenting and evaluating features of tests and test items, and each generates information that can be used to guide strategic modifications to reduce the influence of ancillary interactions on test results. In sum, the process – which we argue is best conducted in teams – consists of three steps: the item writing/accessibility team *identifies* features of the items that may present access barriers for some test-takers, *evaluates* the item on the basis of these features, and *modifies* the item according to suggestions made by the team.

The *TAMI ARM* (Beddow et al., 2009) represents the most recent advancement in guiding the systematic improvement of test item accessibility. *ARM* ratings can be used to identify features of individual test items that may hinder access

for some test-takers, quantify the accessibility of item elements and overall items, and offer evidence-based recommendations for increasing access for more test-takers. The *ARM* is comprised of two analytic rubrics. The first rubric consists of a matrix for rating the five basic elements of a typical multiple-choice item: the item passage and/or stimulus, the visual, the item stem, the answer choices, and the page or item layout. Raters use a 4-point Likert-type scale to rate the accessibility of each item element (see Fig. 9.2). It should be noted that when evaluating constructed response items, such as short answer or essay questions, the answer-choices element is not rated. The second rubric aggregates the element matrix to yield an overall accessibility rating for the item. Because individual item elements can disproportionately influence the overall accessibility of a test item, the overall rating is not simply an average of the matrix

Level	Description	Heuristic
4	Maximally Accessible for Nearly All Test-Takers	Optimal accessibility for between 95–99% of the population
3	Maximally Accessible for Most Test-Takers	Optimal accessibility for between 90–95% of the population
2	Maximally Accessible for Some Test-Takers	Optimal accessibility for between 85–90% of the population
1	Inaccessible for Many Test-Takers	Optimal accessibility for fewer than 85% of the population

Fig. 9.2 Accessibility levels (Beddow, Elliott, & Kettler, 2010)

Table 9.2 Key characteristics of optimally accessible multiple-choice items

Item element	Optimally accessible if:	Key references
Item passage/stimulus	It contains only essential words. The text is minimal in length and written as plainly as possible. The vocabulary and sentence structure are grade-appropriate. The directions or pre-reading text are clear.	Clark, Nguyen, and Sweller (2006); Mayer, Bove, Bryman, Mars, and Tapangco (1995); Mayer and Moreno (2003)
Item stem	The text is minimal in length, written as plainly as possible. It reflects the intended content standards and/or objectives. The target construct is evident. It is positively worded and uses the active voice.	Haladyna, Downing, and Rodriguez (2002)
Visuals	Are necessary for responding to the item. They clearly depict the intended images and are as simple as possible. They contain only text that is necessary for responding. They are unlikely to distract test-takers.	Mayer and Moreno (2003); Mousavi, Low, and Sweller (1995); Torcasio and Sweller (2010)
Answer choices	Are minimal in length and written as plainly as possible. The key and distractors are balanced with regard to length, order, and content. All distractors are plausible. Only one answer is correct.	Halyadyna, Downing, and Rodriguez (2002); Mayer and Moreno (2003); Rodriguez (2005)
Page/item layout	The entire item and all necessary information for responding are presented on one page or screen. All visuals are integrated with the other item elements rather than being placed off to the side. The layout is well organized and presented in a manner that facilitates responding. There is sufficient white space to facilitate comprehension of necessary item elements. The text and item elements are large and readable.	Sweller and Chandler (1994); Chandler and Sweller (1996); Moreno and Mayer (1999)

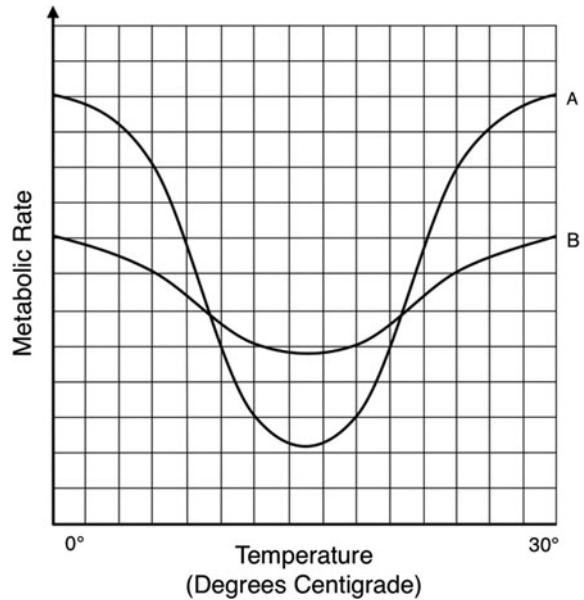
ratings. Table 9.2 contains the key characteristics of optimally accessible multiple-choice items.

It is essential *TAMI* accessibility raters have (a) experience with test item writing; (b) knowledge of the domain content at the grade level(s) of the target test; and (c) familiarity with the range of abilities and needs of the target population of the test. Validity studies on the use of the *ARM* indicate, with minimal time (i.e., 4–6 h), assessment professionals with no prior experience can be trained to rate the accessibility of

items with a high level of agreement with expert raters (Beddow, Elliott, & Kettler, 2010).

To facilitate an understanding of this “pathway to accessibility,” we will model the process of identification, evaluation, and modification with a high school science item (see Fig. 9.3). The item was written, to measure a particular performance objective found in the Grade 11 Content Standards from an unnamed state, as follows: “Use knowledge of biological concepts to draw conclusions from graphed data.”

Fig. 9.3 Grade 11 science item, original form



Although the item is hypothetical, its content, structural features, and target construct all are derived from an actual end-of-course state test item. The item represents several features of common test items that may influence their accessibility for some test-takers. Figures 9.4 and 9.5 (see pp. 175–176) present the same item in two subsequent phases of modification. Each reflects an increasing degree of accessibility resulting from this iterative process.

The item in Fig. 9.3 (“original” form) consists, primarily, of a passage, stimulus, visual, stem, and answer choices. The item begins with a passage that describes a phenomenon whereby low levels of salinity near the equator and at the poles are caused by rainfall and low temperatures, respectively, while the water at the “mid-latitudes” (i.e., between the equator and the poles) contains a higher concentration of salt. The passage is followed by a stimulus that precedes a graph of the metabolic rates of two organisms (referred to as Organism A and Organism B) as a function of water temperature. The item stem directs the test-taker to select the answer choice that contains the best conclusion based on the data presented in the graph. Each of the answer choices refers to the degree to which one of the two organisms is better adapted to life in different locations or salinity levels than the other.

Before undertaking the effort to identify item features that may present access barriers for some test-takers, the rater must first assume the role of the test-taker and complete the item independently. It is by experiencing the item firsthand (without peeking at the item key!) that the rater is able to ascertain the various aspects of the item that demand the use of cognitive resources. Greater the degree the rater attends to the variety of individual differences, that may impact test-takers while completing the item (i.e., using his or her knowledge of, and experience with, the target population of the test), the more information he or she may be able to use to identify features of the item that may influence accessibility.

The following represents the recorded thoughts of a *TAMI*-trained test-taker, describing his attempt to complete the original item as presented in Fig. 9.3:

When I began to read from the top of the question, it immediately became clear that the text was abstruse and very complex. While the vocabulary did not appear to exceed grade-level language expectations for upper-grade high school students, the block of text at the top of the item was difficult to comprehend. The graph was similarly challenging. The axis labels used font sizes that were large enough to read, but it was difficult to keep track of all of the parts of the graph to make sense of the data.

Reaching the item stem and answer choices, the key verb adapted – found in each answer choice – seemed to spring out of nowhere. Indeed, no text preceding the answer choices included the words adapt, adaptation, or any variation thereof. I found words such as metabolic rate, evaporation, precipitation, and I eventually inferred the connection between metabolism and adaptation. Once I made this connection, I referred back to the graph to attempt to draw the necessary conclusion, namely that the metabolic rate of Organism A (and thus its apparent adaptability) is highest on the extremes of the X-axis. I then re-read the passage to learn that those extremes represent the poles and the equator. Re-reading the answer choices – several times over – I did not find a statement about Organism A’s adaptability at the poles and equator that matched my conclusion. I then examined the data for Organism B and found that its metabolic rate is highest at the mid-latitudes. At this point, I searched the answer choices again for a statement that confirmed my latest conclusion, but again, my search was fruitless. Frustrated, I finally decided to read each of the answer choices individually and refer to the graph to determine whether it was potentially true. Thankfully, response option A – “Organism A is less well adapted for life at the mid-latitudes than Organism B” – was indeed correct. I realized that I had skimmed over the word less and wrongly assumed all of the choices concerned which organism was better adapted than the other (except choice D, which equated the two organisms.) Even then I had to double-check to ensure the mid-latitudes were indeed represented by the unlabeled space between the extremes of the abscissa and that the metabolic rate of Organism A was indeed lower there than that of Organism B. Satisfied (and relieved) I moved on to the next question.

It should be noted that there are many ways a test-taker may have opted to respond to this item, and the person above did not explore every permutation in his think-aloud. It is for this reason that cognitive interview studies with representative participants from the target test-taker population (e.g., Roach, Beddow, Kurz, Kettler, & Elliott, 2010; Johnstone, Bottsford-Miller, & Thompson, 2006) can offer useful information for understanding the various solution processes attempted by students who are likely to take the test. The greater the degree to which test developers who pursue to improve the accessibility of tests can assimilate these perspectives into the development process, the more likely their item generation, evaluation, and modification

procedures will result in items that are free from access barriers for the intended test-takers.

Now that we have completed the item, we may begin the process of identifying potential access barriers. Following the *ARM* (Beddow et al., 2010), we shall begin with the topmost item element, the passage or item stimulus.

Item Stimulus

The stimulus begins as follows: “The temperature of the Pacific Ocean ranges from near freezing (32° Fahrenheit, 0° Centigrade) at the poles to 86° Fahrenheit (30° Centigrade) in close proximity to the equator.”

Cognitive load theory offers a helpful perspective for understanding the mental demands of this first sentence. According to Sweller (2010), the dominant source of complexity in learning tasks is from *element interactivity*. In this sentence, the test-taker simultaneously must hold two informational elements in working memory to integrate them into a unit of understanding that permits him or her to proceed to the next sentence. Specifically, the relation between temperature and location must be clear before the test-taker can move onto the next sentence, which begins like this: “The salinity of the water is highest at the mid-latitudes.” If the test-taker has not integrated the temperature-location knowledge from the first sentence such that it is available for use in working memory, the second sentence may be difficult to follow, as the word *salinity* is presented for the first time without an explicit statement of how temperature and location are related to salinity.

Several additional sources of potential element interactivity can be found in the first paragraph of the item stimulus alone – namely, the test-taker must understand the relation between salinity and location to make the later connection between salinity and temperature in answer choice C, and the relation between Fahrenheit and Centigrade temperature systems likely must be understood in order to mentally dismiss one temperature system or the other to avoid confusion. In fact, it may be possible to eliminate both measurement systems

altogether by describing the water using the terms *warm* and *cold* instead of numeric degrees.

Clearly, the item stimulus contains numerous features that likely pose access barriers for at least some test-takers. At the lowest accessibility level for the item stimulus, the ARM item analysis matrix contains this descriptor: “the majority of text is likely to be difficult to understand for some test-takers.” Based on the factors above, the highest ARM rating the item stimulus can receive is 1 or *inaccessible for many test-takers*.

Item Stem

The next step is to identify features of the item stem that may contain access barriers for some test-takers. At first glance, the item stem appears to be worded plainly. However, two features of the stem could be changed to conform to professional guidance from CLT and item writing guidelines. Specifically, the word *best* uses all capital letters as opposed to using bold or underline for highlighting the importance of the word (Clark et al., 2006). Second, the stem is written partially in the passive voice (“can be drawn”), which may cause confusion for some test-takers (Haladyna et al., 2002). While the accessibility of the stem is not optimal, these concerns are relatively minor; based on the item analysis matrix, the item stem receives a rating of 3 or *maximally accessible for most test-takers*.

Visuals

The next item element to examine is the visual. The first question to ask is, “Is the visual *necessary for responding* to the item?” In terms of item evaluation, visuals include any pictures, charts, tables, or graphs that appear on the page with the other item elements. Evidence indicates the inclusion of *nonessential* visuals may hinder reading comprehension for some learners (Torcasio & Sweller, 2010) and negatively influence student test performance (Kettler et al., in press). In its current form, the item in Fig. 9.3 requires the test-taker to use the graph to respond; therefore,

the graph must be examined in terms of its clarity, simplicity, and its location with respect to the other item elements to identify features that may cause access problems for some test-takers.

The graph’s Y-axis label is *Metabolic Rate*, which has no reference term in any other item element and its relation to adaptation must, therefore, be deduced. The axis has an arrow symbol at the top to indicate low versus high metabolic rate. This tiny graphical element may be missed by some test-takers. Additionally, to the degree the test-taker has prior knowledge that metabolism may be used as an indicator of fitness (and that *metabolic rate* is the measure of metabolism), and if he or she understands that fitness or survival rate may be used as a proxy for adaptability, the label does not present an access barrier. This raises the question of the target construct of the item. Is the demand for these connections with prior knowledge an ARC, or is it a component of the target construct of the item? The performance objective (our nearest definition of the intended measurement construct) is, as stated in the content standards, “Use knowledge of biological concepts to interpret graphed data.” The item stimulus includes ample discussion of the relation between temperature and salinity, the assimilation of which likely demands extensive intrinsic load. The focus should be to reduce extraneous cognitive load wherever possible. Arguably, replacing the word *metabolic* with *survival* eliminates the demand for using the transitive property (if a equals b and b equals c then a equals c , or *if metabolic rate equals survival rate and survival rate equals adaptability, then metabolic rate equals adaptability*), thus reducing cognitive load. With this modification, the test-taker is still required to understand that survival rate is a proxy for adaptability – a biological concept – so the change does not dilute the item’s measure of target construct of the item. It should be noted that this issue supports the need for item modification to be a collaborative process consisting of interactive discussion among assessment experts, content area specialists, and educators with familiarity with the target population of a test. It could be said that this is an example of where science meets art in the

arena of test item writing (Rodriguez, 1997). The X-axis label is *Temperature*, with the clarifier *Degrees Centigrade*. Removing one temperature system from the stimulus – or even eliminating both systems in favor of the *warm* and *cold* terms mentioned earlier – would permit the elimination of this clarifier and reduce the reading load of the visual.

Taken together, the features of this item visual are likely to cause access barriers for many test-takers. Therefore, the visual receives a rating of 1 or *inaccessible for many test-takers*. As we will observe shortly, the visual is the central feature of the item and should receive abundant attention if our goal is to improve the accessibility of this item.

Answer Choices

The next item element is the answer choices. Haladyna et al. (2002) listed a number of guidelines for writing effective answer choices, including recommending all answer choices should be plausible and balanced with regard to length and content. Further, based on a meta-analysis of 80 years of research, Rodriguez (2005) argued three options are optimal for multiple-choice tests. Finally, Rodriguez has argued that the use of the term *distractor* is inappropriate given the purpose of incorrect options in a multiple-choice item; namely, the incorrect options should consist of common errors. The author has argued the term *attractors* would be more appropriate since the objective for item writers should be to discriminate test-takers who know the material from those who still tend to make these types of errors. Implausible distractors, on the other hand, tend to be the least-selected and do not discriminate as well.

Based on this guidance, the answer choices in Fig. 9.3 are problematic for a number of reasons. First, the choices are unbalanced with regard to content. Choices A and D begin with Organism A; choices B and C begin with Organism B. Choices A and D use the term “well adapted”; choices B and C use “better adapted.” Choices A, B, and D refer to geographic locations

(mid-latitudes, poles, and equator, respectively); choice C refers to water salinity. Choices A and C use the comparative term *than*; choice B uses *compared to*. All of these factors increase the cognitive demand of the element, particularly insofar as they require multiple element interactivity (salinity with location, temperature with survival, Organism A with Organism B). Second, choice D is highly implausible. Except for two specific (unlabeled) locations on the graph, the data for each organism are clearly disparate from the other across the range of temperatures. This is a “throw-away” response option (and is a clear candidate for elimination according to Rodriguez, 2005). Much work can be done to make the answer choices accessible for more test-takers, and according to the ARM (Beddow et al., 2010) matrix, the answer choices element receives a rating of 1 or *inaccessible for many test-takers*.

Page/Item Layout

The final element to examine for accessibility concerns is the layout of the page or item. Again, at first glance everything about this element appears to be in order and nothing “clangs” of access barriers. The item is embedded in the center of the item between the item stimulus and stem, as recommended by cognitive load theorists (e.g., Clark et al., 2006), to reduce the need for representational holding. As it is currently presented, the item allows the test-taker to find all of the information needed to solve the item within the item space. Indeed, if the item visual were presented on a facing page – or worse – on the separate page that is out of immediate view, the element interactivity concerns noted earlier would be compounded by the fact of the test-taker being required to “carry” this knowledge in his or her working memory across the gap between pages or parts of the page. However, the block of text at the top of the page may be intimidating to test-takers with poor reading skills (again, we must remember that the target domain of the item is science, not reading) and there is little space between lines of text. The graph appears

cluttered by text and lines and one rater quipped about a similar visual that “my eyes went go googly trying to figure this out.” Finally, the item number beside the item stimulus is small and, depending on the relative location of this item to the one before it, may be missed by some test-takers. Based on these concerns, the page and item layout for this item receives a rating of 2 or *maximally accessible for some test-takers*.

Overall Accessibility Rating

The *ARM* contains an Overall Analysis rubric that is designed to be completed after rating the accessibility of each of the individual item elements. Although the rater is advised to use the element ratings to inform his or her determination of the overall rating, it should be reiterated that the overall accessibility of the item may be disproportionately influenced by one or more item feature over the others, so averaging across the individual item element ratings is not recommended. In this case of the item in Fig. 9.3, while the item stem may be maximally accessible for most test-takers, the extraneous cognitive load in the stimulus and visual both decrease the accessibility of the item as a whole. The opposite is possible, of course: An item can consist of an optimally accessible passage and visual, but the stem contains so much extraneous cognitive load that the test-taker is unable to understand how to select the correct answer choice even if he or she possesses a sufficient amount of the target construct to demonstrate his or her knowledge on a more accessible item. Based on the rubric, the item in Fig. 9.3 receives a rating of 1 or *inaccessible for many test-takers*.

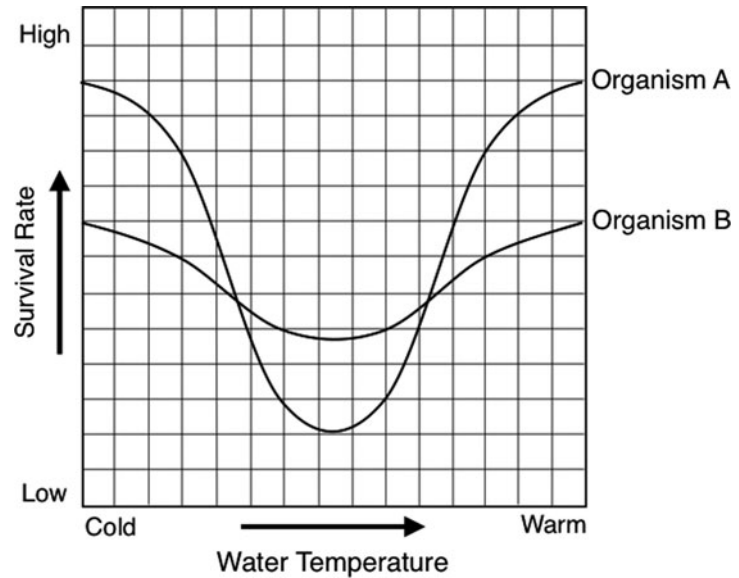
Figure 9.4 consists of the same science item following modification based on the identification and evaluation procedures described above. Using the *ARM* Record Form (Beddow et al., 2009), the raters who evaluated the original item suggested several modifications that may improve the accessibility of the item. For the stimulus, the raters suggested simplifying or shortening the text, reorganizing information, and changing the text formatting to reduce cognitive

demand. The resulting stimulus, which consists of an introductory sentence consisting of a directions statement, is as follows: “Use these facts about the Pacific Ocean to answer the question.” This directive is followed by two bulleted statements containing bold text for key terms or information elements (i.e., those necessary for responding; e.g., *salinity*, *lowest*, and *very cold*). These statements precede a simplified stimulus for the visual.

These modifications are likely to improve access for more test-takers for a number of reasons. First, the multiple element interactivity – the primary cause of cognitive overload – is reduced. The temperature systems from the original items have been removed in favor of qualitatively descriptive text labels for water temperature, as have the extraneous information about the locations where the high- and low-salinity water is found. On the latter, objections could be raised: Is the information about the high-salinity water being found at the poles and equator in fact extraneous, or is it part of the target construct of the item (i.e., the “biological knowledge”)? Unless the item is intended to measure the test-taker’s ability to use knowledge of the geographical locations on the earth where salinity of water is highest or lowest to interpret graphed data, including this information arguably adds extraneous complexity to an item that already contains access barriers for many test-takers. If, on the other hand, a secondary purpose of the item was to *teach* test-takers this knowledge, then its inclusion would be necessary. In this example, however, the item was designed purely for measurement purposes and its inclusion does not facilitate measurement. Indeed, the material may occupy cognitive resources that test-takers need to respond to the actual target construct, possibly causing some to respond incorrectly to the item because of cognitive overload.

Even after eliminating this element interaction, however, the item still requires the test-taker to retain several element interactions in working memory. The subject of the stimulus is salinity; the subject of the visual is water temperature; and the subject of the answer choices is also salinity. Thus, the test-taker must be able to use reflexive

Fig. 9.4 Grade 11 science item, modified form A



logic to select the correct response option, which, it could be argued, is another ancillary requisite construct. In its modified form, the stimulus receives a rating of *maximally accessible for most test-takers* – an ARM rating of 3 – with the recommendation that it be simplified further to improve its accessibility.

Bold font was used to facilitate identification of the stem among the other item elements. In its modified form, the stem receives a rating of 4 or *maximally accessible for nearly all test-takers*. The item stem was reworded in the active voice, to read as follows: “Based on the graph, what could you conclude about the two organisms?” It should be noted that some test writers recommend avoiding the word *you* in the questions, arguing that it makes the item sound less professional or could lead test-takers to believe the item demands an opinion rather than a correct response. Use of the second person notwithstanding, the stem receives a rating of 4 or *maximally accessible for nearly all test-takers*.

Several changes made to the visual during modification reduced the cognitive demand of the item. The word *Metabolic* on the Y-axis label was changed to *Survival* to facilitate test-takers’ comprehension of the connection between the graphed data and the answer choices, since it is generally expected that the biological

concept of adaptation is delivered in the elementary grades (e.g., National Center for Education Statistics, 2011), while metabolism as an indicator of organism survival is taught much later. The arrow embedded in the Y-axis was separated and increased in size, along with the addition of the words *low* and *high* to simplify the characterization of the range of survival rates. Similarly, an arrow was added to the X-axis to facilitate the test-taker’s interpretation of the water temperature progression. Finally, the two data series were labeled with the specific organism titles instead of using simply *A* and *B*. The modified visual receives an ARM rating of 3 or *maximally accessible for most test-takers*, with the recommendation that it be simplified even further.

The answer choices were revised significantly. To achieve balance, all three choices are positively worded (i.e., better adapted, equally well adapted). All three refer to *water with low salinity*. The implausible option D was eliminated. The modified answer choices element receives a rating of 4 or *maximally accessible for nearly all test-takers* – the highest ARM rating.

Finally, the page and item layout are moderately improved. The bulleted text at the top of the item creates more white space than the original item. The item number is increased in size to make it easier for the test-taker to see where one

item ends and the next item begins. Additionally, there is an implication that the test-taker is now permitted to respond on the actual page instead of recording his or her answers on a separate answer sheet, thus reducing the need for representational holding and eliminating the possibility of the dreaded classic “bubble sheet misalignment boo-boo,” which needs no further explanation. The modified page and item layout receives an ARM rating of 4 or *maximally accessible for nearly all test-takers*, with the recommendation that white space be increased to improve the accessibility of the item.

Overall, the item receives an ARM rating of 3 or *maximally accessible for most test-takers*. It should be noted that when the rater is able to generate ways to revise an item further to improve the accessibility of the item, the authors of the *TAMI* argue the item should not receive the highest accessibility rating. They argue item development, particularly when there is a focus on accessibility, is an iterative process and items should not be used for measurement for accountability or decision making until they are deemed to be optimally accessible for the target test-taker population. Although this version of the item was rated highly using the *ARM*, the rater suggested changes to the item stimulus, visual, and item

layout. Figure 9.5 contains the same item following a second set of modifications to improve its accessibility for more test-takers. In its final form, you will note the item has changed considerably.

Specifically, the word count of the item stimulus has been shortened by 43%. The first sentence of the item in Fig. 9.4 (situating the item in the context of the Pacific Ocean) was deemed extraneous, as was the inclusion of the relation between salinity and water temperature. By eliminating the demand for reflexive logic (i.e., the cognitive shift from the concept of salinity to temperature and then back to salinity) the item now requires the test-taker to hold one potentially novel informational element in working memory to interpret the graph; namely, the definition of salinity. In fact, the graphed data are interpretable even without this knowledge. (It should be noted, however, that were this primer removed, some test-takers may feel anxious if the word *salinity* is unfamiliar to them.)

Clearly, the graph is easier to interpret in its current form than in either of the two previous versions. The elimination of the temperature connection greatly simplified the data series, and the elimination of the gray boxes makes the item appear much simpler. In fact, it is likely some readers will object to these changes, since

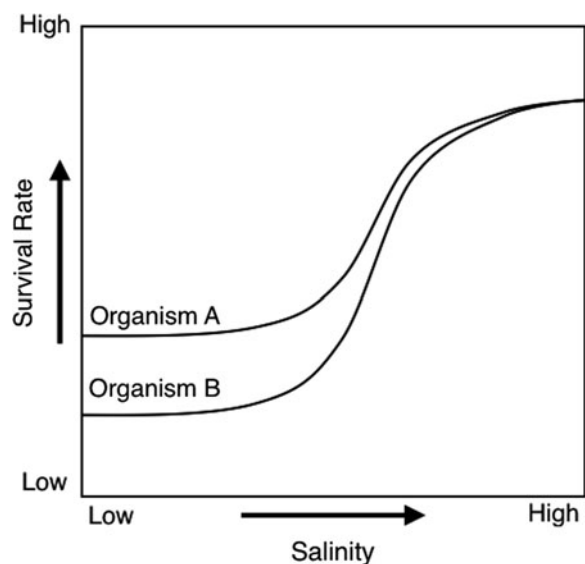


Fig. 9.5 Grade 11 science item, modified form B

the item in its current form demands only that the test-taker know (a) that survival rate is a proxy for adaptability – a concept which, as we noted earlier, typically is covered in elementary school – and (b) how to read a fairly simple two-way graph consisting of two data series. Consider, however, the target construct as stated earlier: “Use knowledge of biological concepts to interpret graphed data.” In its final form, this item demands test-takers use their knowledge of the fact that organisms who are able to adapt to their environments are more likely to survive in that environment to interpret a simple graph comparing the survival rates of two organisms in different environmental conditions. This is the target construct of the item – our stated purpose from the beginning!

Technical Evidence to Verify Intended Effects: The Accessibility Proof Paradox

There are several item-level indices that may be used as indicators of the effects of modification procedures to enhance accessibility, including word count, readability, depth of knowledge ratings (DOK; e.g., Webb, 2002), item difficulty (p), item discrimination indices such as item-total correlations (r), and reliability indices such as Cronbach’s alpha (α). For many of these indicators, expected changes are mostly straightforward. Given the relation between cognitive demand and reading load, for example, word count is likely to decrease for most items following enhancement, with the occasional exception of items for which contextual support or clarifying text is added (e.g., definitions for construct-unrelated but essential terms, labels for graphs). Similarly, readability indices may indicate reduced reading load (though for some assessment instruments it may be important to ensure readability does not fall below grade level expectations). It should be noted that typical readability indices utilize weighted formulas that include various factors, including sentence length and average syllables. When two or more

readability indices are compared, results often are highly disparate (Wright, 2009). When using readability as a proxy for the grade level of text, results should be interpreted with caution. Generally, readability is considered acceptable when the grade level of a text falls anywhere within the range of the intended grade level (e.g., 7.0–7.9 would be acceptable readability values for a grade 7 reading item).

By contrast, determining the desired effects of item changes, as estimated by other indices, is somewhat counterintuitive and may even present logical problems. The first of these challenges is the question of item difficulty. It would seem increases in item difficulty are always desirable, since they demonstrate a greater proportion of test-takers responded correctly to the item (which must be a good thing!). However, there is a critical set of conditional limitations that determines whether item difficulty should increase if access barriers are removed from items. For example, item difficulty is expected to increase following modification *only* if the accessibility of the original item were less than optimal for a portion of the tested population, *and* if some of the same test-takers were able to demonstrate their knowledge of the tested content had the item been optimally accessible. Thus, *increases* in item difficulty are not always the best indicators of intended effects.

Ideally, item discrimination will remain the same or increase if modifications resulted in the elimination of access barriers for a portion of the target population. It should be noted, however, that in this case, the *ideal* is an optimally accessible test! If other test items contain access barriers for some test-takers, then the *total* component of the item-total correlation that comprises the item discrimination index likely will reflect these access skills, thus resulting in *decreased* item discrimination even if the accessibility of the individual item has improved. Of course, these are challenging concerns for researchers interested in examining the effects of item modifications or enhancements because decreases in item discrimination may indicate either (a) the modification changed the target construct of the item, while the

tested construct represented by the balance of the items remained pure, or (b) the modification isolated the target construct of the item, while the tested construct represented by the balance of the items remained adulterated by ARCs.

The solution to both the item difficulty and item discrimination paradoxes harkens back to the essence of test validation and requires test developers to generate a validity argument based on all available data about the current test. In most cases, the reliability of an accessible test, often indicated by α , should be similar to that of validated tests of similar constructs. To wit, if the average of the item-total correlations for the universe of test items is high, the deduction is that the items map to the same construct (one that hopefully is uncontaminated by ARCs, a deduction that must be tested). One solution may be to examine point-biserial correlations with validated measures of the same construct (e.g., correlating items from the current mathematics test with total scores of a math test with established validity evidence). Additionally, test developers should examine the test results of students who have demonstrated access needs against the scores of students who are able to demonstrate their knowledge of the target construct, notwithstanding the presence of access barriers for other test-takers. If the modified version of the test or test items contains fewer access barriers than the original version, evidence should exist of a differential (greater) boost for the test-takers for whom accessibility has been improved (Kettler et al., in press). Evidence of discriminant validity (i.e., evidence that a measure correlates weakly with tests of theoretically orthogonal constructs) may include examining correlations with measures of test-taker working memory (e.g., Baddeley, 2003) or – in the case of a test of a construct other than reading – reading fluency. In essence, we recommend test developers generate *a priori* criteria for making a determination about whether access strategies have been successful. Additionally, we contend it is unwise to rely on one index alone for making such a determination; rather, an argument about the effectiveness of accessibility strategies must be based on the triangulation of multiple item- and test-level indices.

Conclusion

Universal Design: The End or the Beginning?

Universal design is a useful starting point for understanding the need to develop tests that are accessible for the entirety of the intended population. However, the principles of UD contain little operational guidance for developing accessible assessments. Indeed, tests are not designed to be *universally* accessible; rather, they are intended to be accessible for a group of test-takers with a common set of individual characteristics. Test developers should aim to design tests such that variation in this population does not influence test scores and subsequent inferences about the target construct.

Accessibility Theory: The Test-Taker, Not the Universe

Accessibility theory extends guidance from universal design and universal design for assessment (Thompson et al., 2002) and operationalizes test accessibility as the sum of interactions between features of the test and characteristics of each test-taker. To enhance accessibility, therefore, we contend the test developer must be grounded both in effective test design strategies as well as have familiarity with the range of abilities and needs of the population for which the test is designed. Accessibility theory integrates principles of working memory, cognitive load theory, and research on test and item development to provide specific guidance for ensuring tests and test items do not contain features that prevent test-takers from demonstrating their knowledge of the target construct.

Cognitive Load Theory: From Teaching to Testing

Cognitive load theory arguably contains the deepest wellspring of conceptual and operational guidance for the development of highly

accessible tests. By attending to cognitive load during the design phase, test item writers can manage cognitive resource demands, eliminating interference from access skills and isolating the target construct of the test.

Developing Accessible Test Items: Identify, Quantify, and Modify

While accessibility theory certainly offers guidance that can be used to inform the development of accessible items from their inception, we contend the process of item writing with a focus on accessibility is iterative. We recommend test developers examine existing test items to *identify* features that may present access barriers for some test-takers, *quantify* the accessibility of the items, and *modify* items with poor accessibility, removing access barriers to improve their accessibility for more test-takers.

Technical Evidence to Verify Intended Effects: The Accessibility Proof Paradox

After systematically evaluating the accessibility of a test and ensuring, to the extent possible, that it is free from access barriers, evidence should be collected to verify the effectiveness of the process. Notwithstanding the complexities of establishing a test accessibility argument, it is critical test developers generate explicit expectations and utilize multiple indices at both the item and test levels (e.g., readability indices, depth of knowledge, item difficulty, item discrimination, and relations with other measures such as working memory and reading fluency) to test their veracity.

The Access Pathway: Accessibility Across the Educational Environment

The goal of the discussed item modifications via the *TAMI Accessibility Rating Matrix* (Beddow et al., 2009) was to increase student access to the item's target construct. The systematic

application of such accessibility reviews (and subsequent modifications) across a test's constituent item set is thus meant to strengthen the validity of test score inferences about a student's knowledge and skills. That is, the test score from an optimally accessible test permits a more accurate inference about what a student knows and can do because inaccessible tests inextricably measure an unintended nuisance dimension, namely the student's ability to manage (extraneous) cognitive load caused by construct-irrelevant item features. Porter (2006) defined the content related to tested knowledge and skills as the *assessed curriculum*. As such, test accessibility is concerned with increased student access to the assessed curriculum at the time of the test event.

Besides the assessed curriculum, researchers (Anderson, 2002; Kurz & Elliott, 2011; Porter, 2006) have identified additional curricula of the educational environment such as the *intended curriculum* (e.g., content expressed in state standards) and the *enacted curriculum* (e.g., content covered by teacher instruction). Kurz and Elliott argued that an important goal of schooling is to provide students with the opportunity to learn the intended curriculum. They further noted three major barriers to access for students with disabilities (see Fig. 1.1 in Chapter 1), including insufficient opportunity to learn the intended curriculum, inappropriate testing accommodations (or lack thereof), and construct-irrelevant variance on assessments. The design and/or modification of more accessible tests using tools such as the *ARM* (Beddow et al., 2009) clearly support the promotion of greater student access to assessed curriculum. Unfortunately, we have to recognize that even the most accessible test is unable to overcome access barriers related to the assessed curriculum preceding the test event such as a lack of instruction covering the tested content.

The aim to enhance accessibility across the educational environment should be focused on students' access to the content of its various curricula, including the intended, enacted, and assessed curriculum. The teacher's enacted curriculum is supposed to cover the knowledge and skills put forth in the intended curriculum, while

the assessed curriculum is designed to sample across the various domains of the intended curriculum. Access to the assessed curriculum thus can be compromised via a teacher's enacted curriculum that fails to cover the content prescribed by the standards. Kurz discussed this aspect of accessibility in greater detail under the concept of opportunity-to-learn (OTL) in [Chapter 6](#). Many students, especially students with disabilities, may require additional instructional adaptations to facilitate student learning of the teacher's enacted curriculum. The chapter by Ketterlin-Geller and Jamgochian explicated how accessible instruction can be supported through instructional accommodations and modifications (see [Chapter 7](#)), and Phillips highlighted how accessibility for students with disabilities is integrated into a legal framework that mandates their physical and intellectual access to curriculum, instruction, and assessment (see [Chapter 3](#)). In this chapter, we elucidated on the final stretch of the access pathway, the test event. While Tindal ([Chapter 10](#)) and Kettler ([Chapter 13](#)) discussed appropriate testing accommodations as an avenue to provide students with disabilities access to tested content at the time of the test event, item modifications, such as the ones discussed in this chapter, are less concerned with the test's administration including its presentation, timing, mode of response, or environment and more concerned with the accessibility of the test itself.

As evidenced by the many chapters in this book, there are a number of access points to the intended curriculum that can be viewed as occurring along an access pathway, which (a) begins with physical access to buildings and classrooms; (b) continues with instruction that provides students with the opportunity to learn what is expected and measured; (c) supports student learning with needed instructional accommodations and modifications; and (d) culminates in assessments that grant students optimal access to demonstrating their achievement of the intended curriculum. The final access points for students occur during the test event, which may present access barriers for many test-takers. To overcome these access barriers it is necessary to consider the interactions between test-taker

characteristics and features of the test and its administration. Currently, two primary methods are used to reduce the influence of access barriers: testing accommodations and test or item modifications. Just as the decision to assign a particular testing accommodation should consider the interaction between a test-taker's disability-related characteristics and certain aspects of the test administration, which may limit his or her access to demonstrate achievement of the target construct (e.g., a scribe for a student with a physical disability), the design and/or modification of items, with a focus on accessibility, must consider the interaction between specific test-taker characteristics and certain features of the test itself (e.g., large print for a student with a visual impairment). An important distinction between the two access strategies, however, is that test-taker characteristics are considered on an *individual* basis in the case of testing accommodations, whereas item modifications consider the characteristics of a *group* of test-takers. Of course, the purpose of enhancing accessibility across the access pathway is to afford students with greatest opportunity to learn and to enable them to demonstrate their learning. The ultimate goal, however, is much grander, and infinitely elusive: namely, to provide educational services that are optimized to meet the specific needs of each child in our schools.

References

- Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory into Practice, 41*(4), 255–260.
- Baddeley, A. (2003). Working memory: Looking back and looking forward. *Neuroscience, 4*, 829–839.
- Beddow, P. A. (2010). Beyond universal design: Accessibility theory to advance testing for all students. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques* (1st ed., pp. 383–407). New York: Information Age Publishing.
- Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009). *TAMI accessibility rating matrix (ARM)*. Nashville, TN: Vanderbilt University.
- Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2010). *Test accessibility and modification inventory (TAMI) technical supplement*. Nashville, TN: Vanderbilt University.

- Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test accessibility and modification inventory (TAMI)*. Nashville, TN: Vanderbilt University.
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293–332.
- Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*, 151–170.
- Clark, R. C., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-Based guidelines to manage cognitive load*. San Francisco: Jossey-Bass.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., et al. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children, 76*, 475–495.
- Elliott, S. N., Kurz, A., Beddow, P., & Frey, J. (2009, February). *Cognitive load theory: Instruction-Based research with applications for designing tests*. Paper presented at the national association of school psychologists' annual convention, Boston.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–333.
- Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical report 44). National Center on Educational Outcomes, University of Minnesota, 25.
- Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education, 84*, 529–551.
- Kettler, R. J., Rodriguez, M. R., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (in press). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*.
- Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*, 3–16.
- Kurz, A., & Elliott, S. N. (2011). Overcoming barriers to access for students with disabilities: Testing accommodations and beyond. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques*. Charlotte, NC: Information Age Publishing.
- Mace, R. L. (1991). *Definitions: Accessible, adaptable, and universal design (Fact Sheet)*. Raleigh, NC: Center for Universal Design, NCSU.
- Mace, R. (1997). *The principles of universal design* (2nd Ed.). Raleigh, NC: Center for Universal Design, College of Design. Retrieved May 20, 2010, from http://www.design.ncsu.edu/cud/pubs_p/docs/poster.pdf
- Mace, R. L., Hardie, G. J., & Place, J. P. (1996). *Accessible environments: Toward universal design*. Retrieved May 20, 2010, from http://www.design.ncsu.edu/cud/pubs_p/docs/ACC%20Environments.pdf
- Mayer, R. E., Bove, W., Bryman, A., Mars, R., & Tapangco, L. (1995). When less is more: Meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of Educational Psychology, 88*, 54–73.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*, 43–52.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.
- Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*, 348–368.
- Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*, 319–334.
- National Center for Education Statistics. (2011). *The Nation's Report Card: Science 2009 (NCES 2011–451)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- Plass, J. L., Moreno, R., & Brunken, R. (Eds.). (2010). *Cognitive load theory*. New York: Cambridge University Press.
- Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–159). Mahwah, NJ: Lawrence Erlbaum.
- Roach, A. T., Beddow, P. A., Kurz, A., Kettler, R. J., & Elliott, S. N. (2010). Incorporating student input in developing alternate assessments based on modified academic achievement standards. *Exceptional Children, 77*, 61–80.
- Rodriguez, M. C. (1997, August). *The art & science of item-writing: A meta-analysis of multiple-choice item format effects*. Paper presented at the annual meeting of the American Education Research Association, Chicago, IL.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*, 3–13.
- Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Moreno & R. Brunken (Eds.), *Cognitive load theory* (pp. 29–47). New York: Cambridge University Press.

- Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognitive Instruction, 12*, 185–233.
- Thompson, S. J., Johnstone, C. J., Anderson, M. E. & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S. J., Johnston, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Torcasio, S., & Sweller, J. (2010). The use of illustrations when learning to read: A cognitive load theory approach. *Applied Cognitive Psychology, 24*(5), 659–672.
- Webb, N. L. (2002, April). *An analysis of the alignment between mathematics standards and assessments for three states*. Paper presented at the American educational research association annual meeting, New Orleans, LA.
- Wright, N. (2009). *Towards a better readability measure – The Bog index*. Retrieved June 5, 2010, from <http://www.clearest.co.uk/files/TowardsABetterReadabilityMeasure.pdf>

Validity Evidence for Making Decisions About Accommodated and Modified Large-Scale Tests

10

Gerald Tindal and Daniel Anderson

The *Standards on Psychological and Educational Tests* are typically referenced as the starting point to any argument about general education testing. Likewise, they should be the starting point for testing students with disabilities, whether the decision focuses on the use of an accommodation or a recommendation to participate in an alternate assessment based on modified academic achievement standards (AA-MAS). Using five different kinds of evidence, the standards provide a frame of reference focused on making inferences and decisions. “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests . . . It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself” (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999, p. 9).

The process is iterative with different kinds of evidence being collected at different points in a testing program or in reference to varying inferences and decisions. For example, in standards-based testing, content-related evidence is considered essential. Usually, this type of evidence comes from an alignment study with state standards. When the focus is on the constructs being tested, evidence on internal structures tends

to dominate. In large-scale testing programs, criterion-related evidence may be less valued because of the emphasis on achieving proficiency on grade-level content standards, but nevertheless may be important when considering independent information from other measures of student academic performance. In assessing students with limited English or being served in special education programs, the focus may turn to response processes, or the degree to which relevant target skills are being measured rather than superfluous access skills. Finally, in any standards-based, large-scale testing program, the consequence of assigning student proficiency arises from documenting adequate yearly progress (AYP). In the end, validity is about inferences and the support behind them.

Generally, these sources of evidence are used to counter threats to construct misrepresentation or construct under-representation. In the former, irrelevant constructs are being conflated with those the test was designed to measure. For example, many mathematics tests require considerable reading skills and therefore misrepresent mathematics achievement. In the latter, constructs are under-represented by tapping primarily prerequisite or partial skills that do not fully reflect the construct for which the test was designed. A common example of this might be a simple mathematics story problem that reflects only computation skills, not the decision-making skills needed to weigh relevant from irrelevant information for solving mathematical problems. These sources of evidence, along with documentation of

G. Tindal (✉)
University of Oregon, Eugene, OR 97403, USA
e-mail: gerald.tindal@mac.com

reliability (or reproducibility), however, are really only part of the argument; indeed, they represent only the beginning of the argument.

Although proper test design (see Downing & Haladyna, 2006) is used to assemble a validity argument and organize the collection of supporting evidence, further procedural and statistical documentation is needed to document the process (test administration and scoring) and the outcome (decisions and consequences). In standards-based testing, participation is a critical issue by subgroup and performance levels. Standardization is another issue, often spilling into the use of accommodations or changes in the test so that specific subgroups have equitable access. To fully understand the outcomes, evaluations of testing programs need to include concurrent information about populations of students and implementation by teachers.

A problem exists, however, in that student groups are often more fluid than would be desired and administration-scoring more flexible than normally assumed. For example, in the mid-1980s, the findings from the Institute for Research on Learning Disabilities (IRLD – University of Minnesota) raised serious doubts about the identification of students with learning disabilities (Ysseldyke et al., 1982). Therefore, reporting test outcomes by disability groups may be problematic. Another significant problem is that the Standards (AERA, APA, NCME, 1999) confuse critical terms in test administration for students with disabilities, treating modifications and accommodations as the same within a validity argument: “A variety of test modification strategies have been implemented in various settings to accommodate the needs of test takers with disabilities” (p. 103). As a consequence, confusion may exist in defining the (substantive) nature of changes: When is a change sufficiently substantial that it reflects a different construct and should be reserved for a different kind of decision?

In both instances, validity evidence can no longer be assembled from test design and development. Rather other sources of evidence need to be collected to document the effects of testing specific groups and making specific changes.

From the perspective of the testing program (not just the test), these sources of variance are embedded in practice: in the training and implementation process and in the procedures used in giving the test to students.

We turn to the evidence from quasi-experimental and experimental studies conducted with varying levels of control that seek to affirm causal inference or, at the very least, clarity of interpretation. At this point, the validation process turns to research studies in which tests and measures are used in the field under varying conditions. In the course of the study, students are assigned testing conditions, teachers are trained in testing administrations, and data are collected on the outcomes (using some type of dependent variable). This research base (experimental and quasi-experimental) also needs to be considered in addition to the evidence collected as part of proper test design and construction. Here the term *validity* refers to “the approximate truth of an inference” (Shadish, Cook, & Campbell, 2002, p. 34).

In considering the evidence from these research studies, different criteria are used to judge their adequacy (or more properly, the adequacy of the causal description or explanation). Generally, four types of validity evidence can be documented, with each type focusing on a threat that counters or weakens causal inferences (Shadish et al., 2002). The first type involves the *internal validity* and focuses on the control of the study and how false conclusions can be made between cause and effect: “The validity of inferences about whether observed co-variation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured” (p. 38). For example, many studies involve intact groups of students without random assignment and administer tests over long periods of time. As a consequence, confounding variables may be present (from participant backgrounds, histories, settings, testing, etc.) that make any causal inferences suspect. The second type of validity is *statistical conclusion validity* or “the validity of inferences about the correlation (co-variation) between treatment

and outcome, including effect sizes” (p. 38). For example, accommodation studies have varying sample sizes receiving a treatment of unknown integrity, use specific outcome measures, and conclusions based on various significance tests. The central question is whether the study can reflect co-variation by having sufficient power, in the presence of a known treatment, and using appropriate analytical techniques on an adequate outcome measure. The third type of validity is *external validity* or “the validity of inferences about whether the causal relationships holds over variation in persons, settings, treatment variables, and measurement variables” (p. 38). Can the findings from the study be generalized to others who are similar to them and given similar circumstances (treatments and outcomes)? For example, in studies of students with disabilities, the manner in which they have been labeled, recruited, treated, or measured may limit generalizations. Finally, the fourth type of validity is *construct validity*, “defined as the degree to which inferences are warranted from the observed persons, settings, and cause and effect operations included in a study to the constructs that these inferences might represent” (p. 38). For example, with a specific accommodation administered to a population in a study with relevant quasi-experimental controls and findings (conclusions) generated, can the study provide an appropriate explanation?

In the next section of this chapter, we focus our review from these two validation perspectives. The first focuses on the response processes from the Standards for Educational and Psychological Tests (1999). Have researchers adequately reflected upon the test and its demands in conjunction with the students and their characteristics? We use two areas from accommodations research (extended time and read aloud) to illustrate the research basis. The second part of the review approaches the validation process from the criteria used to judge the adequacy of quasi-experimental and experimental research (Shadish et al., 2002). We focus primarily on construct validity as it addresses the explanation in the most comprehensive manner. Again, we consider only studies within two

areas of accommodations (extended time and read aloud).

Current Practice in Testing Approaches and Implications for Response Processes

With increased attention to large-scale testing and the disaggregation of outcomes by disability and English-language proficiency required by the No Child Left Behind Act (NCLB), concurrent emphasis has been placed upon participation and accommodation, which are a necessary component of all large-scale testing programs (No Child Left Behind Act, 2002). Thompson, Johnstone, Thurlow, and Altman (2005) report that all states now document accommodations use on statewide testing. According to a 2007 report by the National Center on Educational Outcomes (NCEO), 51% of students with Individual Educational Programs (IEPs) receive accommodations on statewide tests (Thurlow, Altman, Cuthbert, & Moen, 2007). When accommodations are viewed as inadequate to successfully document student proficiency, some states have developed AA-MAS. These assessments are designed to be aligned with grade-level content standards but reduce the complexity of the items: “The expectations of content mastery are modified, not the grade-level content standards themselves” (U. S. Department of Education, April, 2007, p. 9). They are designed for a small group of students with a disability that would prevent their attainment of grade-level proficiency in the same time frame as their peers.

In this section, we focus on response processes using two lines of test accommodations research to explicate the relation between task demands and student characteristics in understanding response processes for: (a) extended time, and (b) read aloud. We use the Standards (AERA, APA, NCME, 1999) to define this construct (response processes) as “the fit between the construct and the detailed nature of the performance or response actually engaged in by examinees . . . validation may include empirical studies

of how observers or judges record and evaluate data along with analyses of the appropriateness of these processes to the intended interpretation of construct definition” (pp. 12–13). In our analysis, we believe it is important to go beyond the focus on effectiveness of accommodations because the procedures are quite disparate and the findings considerably inconsistent, with little attention given to the manner in which accommodations are aligned with task demands and student characteristics.

Early research on accommodations was quite limited (National Center on Educational Outcomes, undated) with only few published peer-reviewed studies (National Center on Educational Outcomes, undated). In 1999, 134 studies were found and described in a summary by Tindal and Fuchs (1999); this research base was further updated by the National Center on Educational Outcomes (2001). However, both reports provided only a description of the studies, including their methodologies but without results on the outcomes of various accommodations. Finally, in the past decade, research on the effects from test accommodations has expanded with two meta-analyses completed (Chiu & Pearson, 1999; Sireci, 2004).

Nevertheless, little research has focused on the decision-making process itself, either in the descriptive summaries of research or in the documentation of outcomes. While there have been helpful manuals and procedural documents that explicate the process for selecting accommodations (Thompson, Morse, Sharpe, & Hall, 2005), very few empirical studies have been conducted on these processes. As a consequence, we know little about how teachers *actually* recommend or select specific accommodations. Generally, what little research exists is not particularly supportive of the process (see Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000 and Hollenbeck, Tindal, & Almond, 1998).

As we view the decision-making process for either recommending specific accommodations or assigning students to participate in the AA-MAS, the focus needs to be on (a) the task presented (i.e., the specific domain the measure is assessing) and (b) the student’s individual

background, knowledge, and skills. Together, these characteristics should guide the decision to either (a) provide an accommodation to the standard test or (b) recommend the student take an AA-MAS. This decision is likely to reflect the interaction between the two; hopefully, however, it also is guided by the available research. As we noted earlier, an additional consideration should be the research basis and its quality.

In our view, however, the definition and analysis of task demands need to be specific enough to be useful. For instance, an investigator may state that they have used a “math” measure, without explicitly stating or analyzing the different sub-domains within the measure. Given that math tasks may require qualitatively different skills (i.e., algebra versus geometry), the effect of the decision may be confounded. For example, it may be inferred that the accommodation provides a “differential boost” (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000) to students on items assessing geometric relationships but not on items assessing algebraic relationships. A similar problem exists in the way student characteristics are considered or the manner in which students have been labeled. Students with a range of academic difficulties may all be grouped under the label “learning disabled” (Morris et al., 1998). Yet the effect of the decision may vary substantially between students with the same label but different skills as well as between students with different labels. Unfortunately, students receive labels with little detail in specific behaviors associated with the label (and not all labels are about disabilities). For instance, if a read aloud accommodation is being investigated, how do students classified as “in-need” of the accommodation qualify? Is it because they have received some sort of label by the school? Or, is it more specific, such as the student demonstrating low fluency? If so, how low is “low”?

Between each of these variables (task demands and student characteristics) are potential interactions that may influence the outcome. Future research may examine these factors explicitly; however, paying greater attention to the task demands and student characteristics may alleviate considerable confusion. In the

remainder of this chapter, we present a review of literature but limit it to experimental and quasi-experimental research in the K-12 settings. Additionally, because considerably more research has been conducted on accommodations, we focus primarily on this area but argue that the same issues apply to decisions about selecting the AA-MAS for students. To keep the chapter within a reasonable length, we further limit our focus to two areas of accommodations research that have been most heavily investigated: extended time and read aloud. These two areas are among the most prevalent accommodations used in practice and are, therefore, among the most researched (Thurlow, 2007).

Although evidence of accommodation effectiveness is quite disparate, ranging from highly effective (Huesman & Frisbie, 2000; Johnson, 2000) to completely ineffective (Cohen, Gregg, & Deng, 2005; Meloy, Deville, & Frisbee, 2002), we are less concerned with the outcome than the relation between task demands and student characteristics. For example, some research shows how accommodations provide a differential boost for students with disabilities (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998) while others show the effect as homogenous across groups (Elliott & Marquart, 2004; Meloy et al., 2002; Munger & Loyd, 1991). Still others show a differential boost for students without a disability (Elbaum, 2007; Elbaum, Arguelles, Campbell, & Saleh, 2004; Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Tindal et al., 1998). Examining the task demands and the student characteristics may help explain these results. Unfortunately, as we

document with the research, this information is often unavailable for both variables.

In Tables 10.1 and 10.2, we present a summary judgment of adequacy for a sample of research literature on *extended time* and *read aloud accommodations*. The judgments come from an analysis of the study descriptions of task demands and student characteristics. Task demands are rated on a 0–2 scale. Studies describing the task as a generic “math” or “reading” test received a score of 0; studies describing, but not analyzing, sub-domains received a score of 1; studies both describing and analyzing sub-domains received a score of 2. Student characteristics are rated on a 0–3 scale. Studies that grouped students into generic, school assigned “LD/non-LD” categories received a score of 0; studies providing any additional distinctions received a score of 1; studies providing a specific label (i.e., dyslexic) received a score of 2; and studies using specific measures to categorize students (i.e., all students scoring below the *n*th percentile), received a score of 3. These criterion ratings are then totaled to provide an overall criterion rating on a 0–5 scale.

We primarily include only recent studies just prior to and following passage of NCLB, although we reference the Munger and Loyd (1991) study because of its strong research design. We also focus on experimental and quasi-experimental studies rather than field-based descriptive studies. The Cohen et al. (2005) study was non-experimental, but is included because the sample was large enough to be limited to only students receiving one accommodation (extended time). In this particular study, differential item

Table 10.1 Descriptions for task demands and student characteristics for extended time

Article	Task description rating (0–2)	Accommodation grouping rating (0–3)	Total criteria rating (0–5)
Elliott and Marquart (2004)	0	1	1
Munger and Loyd (1991)	1	0	1
Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000)	2	0	2
Cohen et al. (2005)	2	0	2
Fletcher et al. (2006)	2	3	5
Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al. (2000)	2	0	2
Huesman and Frisbie (2000)	2	0	2

Table 10.2 Descriptions for task demands and student characteristics for read aloud

Article	Task description rating (1–3)	Accommodation grouping rating (1–4)	Total criteria rating (2–7)
Crawford and Tindal (2004) – reading	3	2	5
Elbaum (2007) – reading	2	2	4
Fletcher et al. (2006) – reading	3	4	7
Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al. (2000) – reading	3	1	4
Hale et al. (2005) – reading	3	3	6
Elbaum et al. (2004) – math	3	1	4
Helwig et al. (1999) – math	2	4	6
Helwig et al. (2002) – math	2	2	4
Johnson (2000) – math	1	2	3
Tindal et al. (1998) – math	1	1	2
Meloy et al. (2002) – all subjects	3	2	5

functioning (DIF) analyses were then conducted to examine the difficulty of items in the accommodated condition versus a random sample of students in the un-accommodated condition.

Extended Time Research

Typically, extended time involves providing a reasonable amount of extra time (often up to one-half the standard amount of time) so that students may finish the test. This is an important accommodation because standards-based tests are designed to measure what students can do, not just what they did do. If students' performance is low because an arbitrary time limit is reached, no one can know what the student can do. Furthermore, it helps distinguish missing data (items left blank) from incorrect data (items with the wrong option selected). Otherwise, these two types of data are conflated (and missing items become de facto incorrect items).

Task Demands

Many studies provide exemplary descriptions of the task demands (e.g., Cohen et al., 2005; Fletcher et al., 2006; Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al., 2000; Fuchs, Fuchs,

Eaton, Hamlett, & Karns, 2000; Huesman & Frisbie, 2000). Rather than simply stating the task as “math” or “reading,” the authors of these studies describe domains such as “plane geometry” (Cohen et al., 2005), “reading comprehension” (Fletcher et al., 2006; Huesman & Frisbie, 2000), or “concepts and applications” (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000). Additionally, each of these studies analyzes the effect of the accommodation on each sub-domain, rather than with the test as a whole. Other studies (e.g., Munger & Loyd, 1991) provide a description of the sub-domains, but provide no analysis on them. Therefore, it is not possible to document any potential interaction between the accommodation and the task sub-domain. Still other studies provide detailed accounts of the test development process, while the analysis is based on a generic label (Elliott & Marquart, 2004).

Student Characteristics

While the majority of studies describe the task demands, student characteristics are often not described beyond a generic label (Cohen et al., 2005; Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al., 2000; Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Huesman & Frisbie, 2000; Munger & Loyd, 1991). Elliott and Marquart (2004)

provide a more detailed account of student characteristics than is typical, classifying students into three groups: students identified as LD, non-LD, and educationally at-risk in math. The at-risk label came from teacher ratings on an evaluation scale. This is a helpful step in going beyond the standard labels, although, as the authors themselves highlight, students grouped under the generic LD label were heterogeneous, comprised of “students with mild learning disabilities, emotional disabilities, behavioral disabilities, mild physical disabilities, speech and language disabilities, and mild cognitive disabilities” (Elliott & Marquart, 2004, p. 354). It is reasonable to expect that students within each of these disability categories respond to accommodations differently. Fletcher et al. (2006) used a detailed descriptive label (dyslexic) and was the only study found that classified students as in need of the accommodation based on documented performance.

Read Aloud Research

This accommodation involves having tests read to the student (a) to provide access to the test which would otherwise be inaccessible, and (b) so that irrelevant access skills may be distinguished from relevant target skills. However, considerable variation exists in (a) which parts of the test can be read and (b) the manner in which the reading is done. For example, test directions often are read; on occasion, the prompts (but not the options) are read; finally, the subject matter often is considered in the use of read aloud accommodations, with mathematics more typically targeted than reading tests. The delivery itself also varies whether the accommodation is read by a trained person, or by a controlled reader through either a compact disc or a computer.

Task Demands

Of the 12 studies reviewed on the read aloud accommodation, six describe and analyze the task demands in detail (Crawford & Tindal,

2004; Elbaum et al., 2004; Fletcher et al., 2006; Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al., 2000; Hale, Skinner, Winn, Allin, & Molloy, 2005; Meloy et al., 2002); four describe the task demands in detail but conduct no analyses on the sub-domains (Elbaum, 2007; Helwig, Rozek-Tedesco, & Tindal, 2002; Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999); and two do not describe task demands in sufficient detail (Johnson, 2000; Tindal et al., 1998).

Reading comprehension was the most common task demand specifically described. However, the way this task was accommodated differed markedly among studies – ranging from the student reading the test aloud on their own (Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al., 2000), to the test being administered with a video read aloud (Crawford & Tindal, 2004) or by a designated proctor (Elbaum, 2007; Hale et al., 2005). Still others read aloud only the proper nouns, stems, and answer options to the student (Fletcher et al., 2006). Both the Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al. (2000) and the Hale et al. (2005) studies further described the item demands, stating the number and type of each item. These descriptions were not carried on to item-level analyses in either study; however, unlike differences between domains, the subtleties of item type differences within a domain are not likely to confound the observed effectiveness of a particular accommodation. Meloy et al. (2002) read aloud four sub-tests of the Iowa Test of Basic Skills (ITBS), covering Reading Comprehension, Science, Math Problem-Solving and Data interpretation, and Usage and Expression; separate analyses were conducted for each test.

In mathematics, the task demands were often described in detail (Elbaum, 2007; Helwig et al., 2002, 1999; Johnson, 2000; Tindal et al., 1998), including tasks such as “number sense and operations, geometry, data analysis and probability, algebraic thinking, and measurement” (Elbaum, 2007, p. 221) or “number concepts, mathematic relationships, geometry, estimation, statistics, and measurement” (Helwig et al., 1999, p. 116). However, sub-domain analyses were not conducted in any of the studies, which resulted in

Table 10.3 Description of research elements for extended time studies

Author (year)	Participants <i>n</i> (gen/sped)	Content area	Design	Treatment	Measure
Elliott and Marquart (2004)	97 8th graders from 4 middle schools in 4 Iowa districts; 23 students with disabilities identified, 23 students "at risk" and 51 students at or above grade level	Math	Crossed	Extended/not extended	Two equivalent short forms of Terra Nova 18 Accommodations Survey Academic competence Evaluation Scale (ACES)
Elliott et al. (1999)	Rhode Island 4th graders with total <i>n</i> : 11,273–11,429 Specd: 1264–1306	Math Writing Health	Comparative/ Nested	Examined whether or not students received an accommodation and if so, what type and its effect on scores by student groups	Rhode Island Performance Assessment
Munger and Loyd (1991)	222 5th graders from 18 elementary schools in 6 school districts in Virginia. 6 were physically handicapped, 94 LD, 112 no handicapping condition	Math & Language Arts	Crossed	Timed/not timed	ITBS: Language usage & expression (<i>n</i> = 109) ITBS: Math concepts (<i>n</i> = 113)
Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000)	373 students 190 LD: 129 G4, 63 G5 181 non-LD: Grade 4	Math	Crossed	Administered CBM under standard conditions then under various accommodations	<ul style="list-style-type: none"> • Two alternate forms of computation CBM • Four alternate forms of concepts and application CBM • Five alternate forms of problem solving CBM • All at 3rd grade level
Cohen et al. (2005)	Data obtained from 2003 administration of Florida Comprehensive Assessment Test. 211,601 9th graders evaluated Random sample of 1,250 LD & 1,250 without LD evaluated	Math	Comparative	Only the students receiving extended time as their only accommodation were included in analysis	Florida Comprehensive Achievement Test
Huesman and Frisbie (2000)	Students from two districts: 6th grade 129 LD; 409 non-LD	Reading	Crossed	61 LD had both extended and standard time, 68 had only extended; non-LD student had both conditions	Form K – Level 12, of reading comprehension ITBS

Table 10.4 Description of research elements for read aloud studies

Author (year)	Participants <i>n</i> (gen/sped)	Content area	Design	Treatment	Measure
Crawford and Tindal (2004)	338 students with 89 Title I and 76 special education) in 4th and 5th grades	Reading (five passages with 5–8 questions in each form)	Crossed design (within subjects analysis)	Videotape of items read aloud versus standard test administration	Form A and B for each Reading Test
Elbaum (2007)	643 students in grades 6–10 (388 with LD)	Mathematics	Classrooms assigned to one of four condition (treatment and form) allowing for 1 between – 1 within ANOVA	Teacher read aloud with timed pacing for students to respond versus standard	Two author constructed multiple choice math tests (30 items per form) using practice items from state tests and controlled for linguistic complexity
Fletcher et al. (2006)	3rd grade students (91 with Dyslexia and 91 average)	Reading	Students randomly assigned to either standard or accommodated condition and tested in small groups	Extended session to two blocks and teachers read proper nouns as well as stems and possible responses versus standard test administration	Texas Reading Assessment of Knowledge and Skills (TAKS)
Fuchs et al. (2000)	4th and 5th grade students with LD (181) and without LD (184)	Reading	Crossed (using within-subjects comparisons)	Student received extended time, large print, and read the test aloud versus standard test administration	Iowa Test of Basic Skills: Reading Comprehension
Hale et al. (2005)	4 male secondary students diagnosed as emotionally disturbed	Reading	Crossed (time series design over 9 sessions)	Listening and Listening while reading versus a silent reading control condition	Timed Reading passages (number and rate in comprehension)
Helwig et al. (1999)	247 elementary-age students grouped as high and low on reading fluency and math computation	Mathematics	Crossed design (using within subject comparisons)	Video taped read aloud of a state math test with paced presentation versus a standard administration	21-question computational test and a statewide mathematics test
Johnson (2000)	115 4th grade students	Mathematics	Crossed design: Students assigned according to education placement: (a) general education control (<i>n</i> = 39), (b) general education who were poor or good readers (<i>n</i> = 38), and (c) special education for reading (<i>n</i> = 38) versus a standard administration	Proctors read items verbatim to half of students in group B and to students in group C	Washington statewide achievement test (WASL)

Table 10.4 (continued)

Author (year)	Participants <i>n</i> (gen/sped)	Content area	Design	Treatment	Measure
Kettler et al. (2005)	118 4th graders (49 with disabilities) and 78 8th graders (39 with disabilities)	Reading and Mathematics	Crossed: Students with disabilities assigned to accommodations based on their IEP and then a student without disabilities paired with them: Counter-balanced in administration with and without the accommodation	Project assistants administered the accommodation	Terra Nova Multiple Assessment Battery: Two reading and two math subtests
Meloy et al. (2002)	A total of 260 students: <ul style="list-style-type: none"> • 98 6th graders • 84 7th graders • 78 8th graders 62 students with reading disabilities and 198 without	All subject areas: Reading, Writing, Mathematic, and Science	Nested: Students randomly assigned to take all tests in either an accommodated or a standard administration	The read aloud condition was scripted (with slightly more time)	Iowa Test of Basic Skills: Science, Usage and Expression, Math Problem-Solving and Data Interpretation, and Reading Comprehension
Tindal et al. (1998)	114 4th grade students with and without disabilities	Mathematics	Nested: Students randomly assigned to either a read aloud condition or a standard test administration (allowing use of 1 within and 1 between ANOVA)	The test was administered in both conditions by trained graduate students	Oregon Statewide Test

lower ratings in Tables 10.3 and 10.4. The overall lower ratings for math over reading suggests either (a) more stringent standards used to evaluate math articles, or (b) a lack of research evaluating math sub-domains. Unlike reading comprehension, the accommodated read-aloud condition in math was quite consistent: Conditions included the test being read by a trained teacher/proctor (Elbaum, 2007; Johnson, 2000; Tindal et al., 1998), or a video read aloud (Helwig et al., 2002, 1999).

Student Characteristics

Of the studies reviewed, only Fletcher et al. (2006) and Hale et al. (2005) used a measure percentile cutoff approach to classify students. Fletcher et al. “required that the participating students clearly demonstrate word recognition difficulties” (p. 139). To ensure this, letter-word identification and word attack subtests of the Woodcock-Johnson III were administered. If the student scored above the 26th percentile or the student’s teacher had another assessment showing them functioning above the 25th percentile, the student was no longer evaluated. This approach actually documented the achievement level of students who were deemed “in need” of the accommodation. Students in the Helwig et al. (1999) study took a test of oral reading fluency along with a basic math skills test. Data analyses were then conducted on four student groups: low reader/high math, medium reader/high math, high reader/high math, and low reader/low math. However, too few students were available in the medium reader/low math or high reader/low math for meaningful analyses to be conducted.

Fletcher et al. (2006) also used a *specific* school-assigned disability label to describe students (dyslexic), as did Hale et al. (2005) with their focus on emotionally disturbed. The most common approach to classifying students was to use a *generic*, school-assigned label (i.e., LD). Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al. (2000) have noted that “the appropriate unit of analysis in accommodation decisions is the individual, not the LD label” (p. 69) and that the heterogeneity within the LD group “makes

conceptual analysis of meaningful test accommodations impossible” (p. 68). Despite this, the authors did not go beyond the school-identified labels of LD and non-LD during data analysis, similar to others (Elbaum et al., 2004; Tindal et al., 1998).

Some studies classified students by generic labels, but provided careful descriptions of how the school deployed this label (i.e., discrepancy model, RTI, etc.) or provided other achievement information, such as state test results (e. g., Elbaum, 2007; Hale et al., 2005; Meloy et al., 2002). Other researchers stratified students into more than the typical two or three groups (Special Education, General Education, and Title I; Crawford & Tindal, 2004), while still other studies used a specific label (learning disabled in reading, LD-R; Helwig et al., 2002; Johnson, 2000; Meloy et al., 2002). The LD-R label was deemed slightly more specific than a general LD label; however, students identified as LD-R are still a largely heterogeneous group. Johnson (2000) never explicitly used the LD-R label, but stated that the sample, “consisted of students receiving special education services for reading disabilities, as defined by Washington’s Administrative Code” (pp. 262–263). Meloy et al. (2002) used the school’s definition of LD-R, and further qualified the sample by stating that “students with additional diagnoses or service labels such as behaviorally or emotionally disturbed were not included in the study” (p. 250).

Research Designs and Quality of Research

As noted earlier, the experimental and quasi-experimental research can be analyzed by the degree to which causal inferences are plausible from known treatments (accommodations) to documented outcomes (performance on a state test or other measure of achievement).

In this analysis, threats to *internal validity* focus on the *co-variation* of other causes associated with an effect and can arise from a number of sources: ambiguous temporal sequences, subject selection biases, history and maturation,

regression, attrition, testing, or instrumentation effects. As can be seen in the sources of threats to internal validity, the focus is on causal inference in the context of study designs. Each of the threats presents a counter-explanation of why an effect was achieved irrespective of the treatment, which otherwise should provide the best explanation. Internal validity is concerned with clarity of causal reasoning and the errors that can arise from explanations attributed from a specific cause that is related to a specific effect.

Threats to *statistical conclusion validity* reflect the *strength* of covariance between cause and effect and come from a number of sources: violated assumptions of the statistical tests, low statistical power, repeated analyses of data, restricted range of performance, marked heterogeneity among populations, unreliable treatments, extraneous factors in the settings, unreliability of measures, and inaccurate effect sizes. In these threats, the focus is on the degree to which co-variation can be documented; each threat weakens the ability of experiment to ascertain the presence of an effect, given a presumed cause. Statistical conclusion validity is concerned with errors resulting from statistical analyses used to document co-variation.

With *external validity*, the focus is on *generalizations* across students, settings, treatments, and measures. A number of threats can result in errors resulting from an interaction of causal relations among these four dimensions of a study, limiting inferences to only those students, settings, treatments, or measures that were used in a study. External validity concerns the causal inferences (size or direction) interacting with students, settings, treatments, or measures.

Finally, and perhaps most importantly, there are threats to *construct validity* that concern the relation between the operations used in a study and the *constructs* used to describe them. A number of threats are present including inadequate explication of constructs, confounding constructs, bias from mono-method or mono-operations, constructs interacting with each other, measurement reactivity, study participation and feedback, reactivity to experimental situations, expectancies from experimenters,

novelty and disruption, compensatory equalization or rivalry, demoralization, and treatment diffusion. Construct validity addresses explanatory descriptions of the study and the generalizations that can be made.

Extended Time

Of the six studies we reviewed on extended time, the *internal validity* of the studies reflected a clear direction of causal inference. Researchers conducted studies that were clear on the temporal sequence (of the treatment) and provided clean comparisons of subjects across treatment and control conditions. However, most studies also used convenience samples, so subject selection biases may have been present. All studies were conducted over short time periods, so history and maturation were clearly not viable threats as well as regression or attrition. Because of the crossed conditions, testing effects may have been present in studies that did not counter-balance introduction of the conditions. Finally, with crossed designs, testing and instrumentation effects may have been present. In general, researchers on accommodations have produced studies that avoid many threats to internal validity, thus providing assurance of causation when (significant) effects were found.

Most of the research on extended time accommodations reflected studies with appropriate *statistical conclusion validity*. Researchers generally used appropriate statistical tests and appropriately mined the data files. All studies had sufficient sample size to conduct various statistical tests. With some populations (e.g. students with disabilities), restricted ranges of performance were present, perhaps lessening the ability to show the strength of co-variance between treatments and outcomes; nevertheless, the research on accommodations tended to include a wide range of (general education) students reflecting considerable heterogeneity of students. Few treatments were specifically defined or monitored, although they contained few extraneous factors. Finally, reliability of the measures was seldom documented. The research on accommodations

in general included few controls for threats to statistical conclusion validity, with considerable improvements needed before inferences can be made about the presence (or strength) of co-variation between accommodations and outcomes.

External validity is the area with the most threats. Only Elliott, Bielinski, Thurlow, DeVito, and Hedlund (1999) and Cohen et al. (2005) had sufficient numbers of students to generalize to large-scale test populations. At the other extreme was a study by Hale et al. (2005) which included a small sample size ($n = 4$). All of the studies in accommodations reflected very idiosyncratic contexts, limiting generalizations to other students, settings, treatments, and measures. No studies have been replicated.

Threats to *construct validity* (including consideration of the treatment and outcome measure) were present across all studies. With extended time, treatment integrity was rarely an issue as it was either defined or explicitly monitored. However, extended time in the field is rarely operationalized as a unitary construct but often confounded with setting and grouping to allow schools an efficient manner for implementation. Clearly, then, the research on extended time has more construct clarity than is executed in practice. Most of the researchers used crossed designs, in which students served as their own controls (e.g., they received both the experimental accommodation treatment and the control standard test administration). Because crossed designs were primarily used, a number of threats were minimized in the research, including study participation, reactivity to experimental conditions, novelty or disruption, and compensatory artifacts. Experimenter expectancies were never addressed in the research.

Read Aloud Accommodations

We reviewed 10 studies and found similar variation among the researchers as we found with extended time. This variation affects threats to validity in different ways with the general effect

of providing the field only a nascent accumulation of findings to guide practice.

The studies on read aloud reflect a general consistency in the manner that threats to *internal validity* have been controlled. Nearly all studies provided a treatment with a clear temporal sequence (some kind of treatment was uniquely introduced). Most of the studies were conducted within a short fixed period of time (typically, a class period). Generally, students were conveniently sampled by using entire classrooms, an artifact that confounds students nested in teachers. Some studies have actually used random assignment to treatments, a feature that precludes subject selection bias. For example, Tindal et al. (1998) used a nested design but their random assignment of students controlled for threats to internal validity. Likewise, Fletcher et al. (2006) randomly assigned students to accommodated and control groups to take a state test. Other factors also were considered in selecting students. For Kettler et al. (2005) the IEP was considered and then they matched the student (with a disability) to a student without a disability. Performance on a curriculum-based measure administered under an accommodated condition was used to assign students to accommodations (Fuchs, Fuchs, Eaton, Hamlett, Binkley, et al., 2000). Hale et al. (2005) used a time series design and therefore may have established the best control for various threats to internal validity (history, maturation, regression, attrition, testing, and instrumentation). Elbaum (2007), as well as most other researchers, counter-balanced forms and order to control threats from testing and instrumentation. In general, because the time period for treatment was so constrained, few studies had many threats to internal validity from these latter variables. Testing and instrumentation effects were generally ignored (neither controlled nor described). Published tests were either a state test (or approximation of a state test; Helwig et al., 2002) or a researcher-developed test (Hale et al., 2005).

Studies on read aloud accommodations addressed *statistical conclusion validity* in a number of ways. The most significant issue

was obtaining a sufficient sample size to document co-variation between the treatment and the outcome. Because the focus has been on understanding the effects for students with disabilities, over-sampling has been used often to attain sufficient numbers for conducting parametric analyses (Elbaum, 2007; Fletcher et al., 2006). In some research, sample sizes have been determined ahead of time by considering power and effect sizes needed (Elbaum et al., 2004). Nevertheless, as in the extended time studies, student populations have been somewhat restricted in performance (special education) as well as heterogeneous (general education); limited studies have been done where various skill levels have been documented of students participating in the study. Probably, the most significant issue in read aloud accommodations has been the unreliability of treatment implementation and extraneous factors in the settings that appeared along with the accommodation. The problem with these latter two threats is that any effect may be partially associated with or caused by other factors in the procedures or setting. Few studies have addressed reliability of measures even though considerable attention has been given to their construction.

Threats related to *external validity* are prominent in research on read aloud accommodations. However, it is difficult to ascertain the degree of influence as no replications have ever appeared in this body of research. Each study represents a unique group of students, treated uniquely in both the settings and the manner of delivering an accommodation, and measured using study-specific outcomes. Ironically, the external validity of this accommodation is stronger with studies using more generic treatments with a range of students sampled by classrooms taking measures typical of large-scale testing programs. For example, the scripted directions from Elbaum (2007) may be easier to use in practice than the videotaped administration used by Crawford and Tindal (2004).

Construct validity may be the most troubling aspect of research on read aloud accommodations. Great variation existed among study operations and controls that may or may not have been

present and the constructs used to explain them. Furthermore, considerable variation existed in both what was read and how it was read. For example, Elbaum (2007) had teachers read with time paced; Johnson (2000) used proctors, and Crawford and Tindal (2004) used videotaped read alouds. Finally, read aloud accommodations were rarely implemented as a uniform or singular accommodation. Meloy et al. (2002) not only used a script, but they also implemented extra time. Likewise, Fuchs, et al. (2000) also assigned extended time and large print along with the read aloud. And almost endemic to this research, little connection has been documented between the use of this accommodation and the use of assisted reading in the classroom (where students have materials read aloud to them). Therefore, the construct validity of read aloud accommodations has not been clearly established and the very nature of its implementation has often been associated with other accommodations (e.g., a quiet and separate setting or individual test administration). In the end, several threats associated with construct validity have been present, particularly confounding constructs and constructs interacting with each other. On the other hand, most studies have used crossed designs in which students served as their own control and therefore limited possible threats associated with measurement reactivity, study participation and feedback, reactivity to experimental operations, novelty and disruption, and rivalry and demoralization. The only potential problem associated with this design has been treatment diffusion, an unfortunate side effect from the same teachers implementing both treatments (the accommodation and the comparator condition).

Conclusions on Validity Evidence

Large-scale assessment programs reflect a complex process that requires appropriate test design and development along with the procedural and statistical documentation that provides various test-focused validity evidence for making inferences on content proficiency, internal structures and invariance across subgroups, relations with

other measures, or the relation between access and target skills (response processes). In addition, however, validation evidence needs to arise from testing programs in practice, using quasi-experimental and experimental research conducted in the field after measures have been introduced. This is particularly true when specific populations (e.g. students with disabilities) become the focus of test design and administration.

Although we primarily addressed accommodations to standards-based grade-level tests, our findings have a number of implications for students being assigned to alternate assessments judged against modified achievement standards. Using response processes as an important form of validity evidence, we reported on a considerable amount of research on two accommodations: extended time and read aloud. By considering both the demands of the test and the characteristics of the students, we found a number of studies with sufficient specificity to provide the field a solid research base. We also reported on the study designs and the degree to which adequate threats to validity were controlled. Again, a number of studies have been completed on these two accommodations to provide recommendations to the field.

However, a number of shortcomings also were apparent, particularly in the overall manner in which tests have been studied, to warrant several recommendations for improving practice. We have organized these conclusions around five topics: (a) definitions of constructs, (b) operationalization of test design and implementation, (c) training of teachers, (d) organization of systems, and (e) validation of outcomes in an iterative fashion.

Definitions of Constructs

Measurement practices need to improve considerably in defining constructs, particularly when testing students with disabilities. Although the Standards (AERA, APA, NCME, 1999) provide a definitive source for clarifying validity evidence,

they remain out of synch with the field of accommodations with language precariously perched between definitions of test changes and populations for whom the tests were designed. They do not directly address issues of universal design; nor do they address policy (e.g., federal regulations) in which specific populations of students are targeted. This problem becomes compounded when considering threats to construct misrepresentation, particularly in reference to response processes (i.e., analysis of task demands and consideration of student characteristics).

Operationalization of Test Design

The principles of test design have been well articulated for decades and procedures well established for how items should be developed and displayed, whether they use selected or constructed responses. The problem is that testing (its design and implementation) needs to also attend to administration. Any variation is considered a threat to valid inferences, particularly construct misrepresentation (Haladyna & Downing, 2004). Standardization becomes key in test administration even though it is far more encompassing and includes test planning and development (test specifications, item development, test design, assembly, and production):

The administration of tests is the most public and visible aspect of testing. There are major validity issues associated with test administration because much of the standardization of testing conditions relates to the quality of test administration. . . Standardization is a common method of experimental control for all tests. Every test (and each question or stimulus within each test) can be considered a mini experiment (Van der Linden & Hambleton, 1997). The test administration conditions – standard time limits, procedures to ensure no irregularities, environmental conditions conducive to test taking, and so on – all seek to control extraneous variables in the “experiment” and make conditions uniform and identical for all examinees. Without adequate control of all relevant variables affecting test performance, it would be difficult to interpret examinee test scores uniformly and meaningfully. This is the essence of the validity issue for test administration conditions. (Downing, 2006, p. 15)

Unfortunately, the research on accommodations has not led to consistent results, primarily as a result of inadequate attention to response demands interacting with student characteristics or clear (or consistent) test administration procedures in the context of experimental control.

Training of Teachers

Probably the weakest link in a validity argument on appropriate accommodations is the teacher (or more properly, their role in making decisions and the information that is used to make those decisions). In general, very little attention has been given to teachers in the process of measuring students. In both our analysis of response demands with student characteristics along with the design/execution of research studies, insufficient attention has been devoted to training teachers on fundamental constructs that comprise the entire measurement process. Whether the constructs are being measured as part of a large-scale testing program or manipulated as part of a specific accommodation study, teachers have not been the focus of attention.

Organization of Systems (Research and Practice)

The difficulty in developing a fully functioning large-scale testing program is that often research and practice are in conflict with each other. For example, to attain adequate control in establishing sufficient validity evidence, the research needs to be done very carefully with all steps in the design and implementation monitored. Yet typically, this attention to detail cannot be attained when considering systems implementation with thousands of students. Most of the research has been conducted in specific ways to maximize experimental control; however, this approach is antithetical to informing the field of practice.

Validation of Outcomes

Too often, the research is a single study even though the validation process is iterative and studies should be done in relation to each other. Typically, findings from studies do not build from previous research and application to practice is often unclear. So much variation exists in the test demands and student characteristics that it is not possible to build an adequate validity argument in the design and implementation of tests. This variation is then carried forward to the experimental and quasi-experimental research on accommodations where even more variation exists. No two studies approach the same problem in a consistent and systematic manner. The students who are included in the research vary; the treatments (accommodations) varied in every manner possible, with little control over either their definitions or potential confounds; though often the sample sizes are sufficient to avoid Type II errors, certainly the generalization to large-scale testing programs is insufficient; in the end, the very construct of an accommodation remains elusive. Little consistency exists on extended time and read aloud, the two most researched areas; it is unlikely to be present in other areas of accommodation.

References

- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards of educational and psychological testing*. Washington, DC: American Psychological Association.
- Chiu, C. W. T., & Pearson, P. D. (1999). *Synthesizing the effects of test accommodations for special education and limited English proficiency students*. Paper presented at the Annual Large-scale Assessment Conference of the Council of Chief State School Officers, Snowbird, UT.
- Cohen, A., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice, 20*, 225–233.
- Crawford, L., & Tindal, G. (2004). Effects of a read-aloud modification on a standardized reading test. *Exceptionality, 12*, 89–106.

- Downing, S. (2006). Twelve steps for effective test development. In S. Downing & T. Haladyna, M. (Eds.), *Handbook of test development* (pp. 3–25). Mahwah, NJ: Lawrence Erlbaum Associates.
- Downing, S., & Haladyna, T., M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Elbaum, B. (2007). Effects of an oral testing accommodation on mathematics performance of secondary students with and without learning disabilities. *Journal of Special Education, 40*, 218–229.
- Elbaum, B., Arguelles, M., Campbell, Y., & Saleh, M. (2004). Effects of a student-reads-aloud accommodation on the performance of students with and without learning disabilities on a test of reading comprehension. *Exceptionality, 12*, 71–87.
- Elliott, J., Bielinski, J., Thurlow, M., DeVito, P., & Hedlund, E. (1999). *Accommodations and the performance of all students on Rhode Island's Performance Assessment*. Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.
- Elliott, S., & Marquart, A. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children, 70*, 349–367.
- Fletcher, J., Francis, D., Boudousquie, A., Copeland, K., Young, V., Kalinowski, S., et al. (2006). Effects of accommodations on high-stakes testing for students with reading disabilities. *Exceptional Children, 72*, 136–150.
- Fuchs, L., Fuchs, D., Eaton, S., Hamlett, C., Binkley, M., & Crouch, R. (2000). Using objective data sources to supplement teacher judgment about reading test accommodations. *Exceptional Children, 67*, 67–82.
- Fuchs, L., Fuchs, D., Eaton, S., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29*, 65–85.
- Haladyna, T., & Downing, S. (2004). Construct-irrelevant variance in high stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.
- Hale, A., Skinner, B., Winn, R., Allin, J., & Molloy, C. (2005). An investigation of listening and listening-while-reading accommodations on reading comprehension levels and rates in students with emotional disorders. *Wiley InterScience, 42*, 39–51.
- Helwig, R., Rozek-Tedesco, M., & Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *The Journal of Special Education, 36*, 39–47.
- Helwig, R., Rozek-Tedesco, M., Tindal, G., Heath, B., & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research, 93*, 113–125.
- Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *The Journal of Special Education, 32*, 175–183. doi: 10.1177/002246699803200304
- Huesman, R., & Frisbie, D. (2000). *The validity of ITBS reading comprehension test scores for learning disabled and non learning disabled students under extended-time conditions*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- Johnson, E. (2000). The effects of accommodations in performance assessments. *Remedial and Special Education, 21*, 261–267.
- Kettler, R., Neibling, B., Mroch, A., Feldman, E., Newell, M., Elliott, S., et al. (2005). Effects of testing accommodations on math and reading scores: An experimental analysis of the performance of students with and without disabilities. *Assessment for Effective Intervention, 31*(1), 37–48.
- Meloy, L., Deville, C., & Frisbee, D. (2002). The effect of a read aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education, 23*, 248–255.
- Morris, R., Stuebing, K., Fletcher, J., Shaywitz, S., Lyon, G., Shankweiler, D., et al. (1998). Subtypes of reading disability: Variability around a phonological core. *Journal of Educational Psychology, 90*, 347–373.
- Munger, G., & Loyd, B. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research, 85*, 53–57.
- National Center on Educational Outcomes. (2001). *A summary of research on the effects of test accommodations: 1999 through 2001*. Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.
- National Center on Educational Outcomes. (u.d.). *Testing accommodations for students with disabilities: A review of the literature*. Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.
- No Child Left Behind Act. (2002). *No Child Left Behind Act*.
- Sireci, S. G. (2004). *Validity issues in accommodating NAEP reading tests*. Washington, DC: National Assessment Governing Board.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston: Houghton Mifflin.
- Thompson, S., Johnstone, C., Thurlow, M., & Altman, J. (2005). *2005 State special education outcomes: Steps forward in a decade of change*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S., Morse, A., Sharpe, M., & Hall, S. (2005). *Accommodations Manual, How to select, administer, and evaluate use of accommodations for instruction and assessment of Students with Disabilities*. Retrieved February 28, 2011 from http://www.osepideasthatwork.org/toolkit/accommodations_manual.asp <http://www.cehd.umn.edu/NCEO/Presentations/AERA07Thurlow.pdf>

- Thurlow, M. (2007). *Research impact on state accommodation policies for students with disabilities*. Paper presented at the American Educational Research Association. Retrieved February 28, 2011, from <http://www.cehd.umn.edu/NCEO/Presentations/AERA07Thurlow.pdf>
- Thurlow, M., Altman, J., Cuthbert, M., & Moen, R. (2007). *State performance plans: 2004–2005: State assessment data*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64*, 439–450.
- Tindal, G. & Fuchs, L. (1999). *A summary of research on test changes: An empirical basis for defining accommodations*. Lexington, KY: Mid-South Regional Resource Center.
- U. S. Department of Education. (2007, April). *Modified academic achievement standards – Non-regulatory guidance*. Washington, DC: U. S. Department of Education.
- Van der Linden, W., & Hambleton, R. (1997). Item response theory: Brief history, common models, and extensions. In W. Van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). New York: Springer.
- Ysseldyke, J., Thurlow, M., Graden, J., Wesson, C., Deno, S., & Algozzine, B. (1982). *Generalizations from five years of research on assessment and decision making*. Minneapolis, MN: Institute for Research on Learning Disabilities.

Michael C. Rodriguez

Four Thousand Years of Limited Accessibility

Evidence about the format and function of assessments dates back to at least 2200 BC during the time of the performance assessments of Chinese civil service exams. Records of oral and written exams can be traced back to the 1200s at the University of Bologna and 1400s at Louvain University. During the 1500s, Jesuits (e.g., St. Ignatius of Loyola) adopted written tests for student placement and evaluation. By the 1800s, written exams were commonly replacing oral exams because of questions of fairness. However, these questions were more concerned about fairness in scoring and subjectivity, rather than accessibility or equity in access. Were questions about validity and access ever asked? There is no evidence of assessment-related research during these periods of performance assessments, even though they were all “high stakes.”

Assessment in K-12 education settings became prominent during the mid-1840s as Boston public schools adopted short-answer items in district-wide tests. Once the multiple-choice (MC) item format was formally adopted in the Army Alpha and Army Beta tests in 1917, group-administered intelligence tests modeled after the Binet-Simon Scale of school

readiness (1908, as cited in DuBois, 1970), test design was becoming a science. Thorndike (1904) was the first to introduce principles of test construction in his *Introduction to the Theory of Mental and Social Measurements*. Item-writing research (to which we have access in psychology and education journals) began in the 1920s, at a time when the College Board began using the new MC item types in college entrance exams. Item writing and test design quickly advanced with the landmark chapter on item writing by Ebel in the 1951 first edition of *Educational Measurement*. This period of item-writing advances and modern measurement theory was intensified in the late 1960s, through the introduction of the National Assessment of Educational Progress (NAEP) (see Jones & Olkin, 2004, for a history of NAEP), the assessment we have come to know as the nation’s report card.

The history of testing is instructive, as comprehensively reviewed by DuBois (1970), providing clues to important stages in thinking and practice. One interesting note is that formal assessments were originally in the form of performance tasks, including archery, horsemanship, and accounting tasks (as in the ancient Chinese civil service exams; see DuBois for a more descriptive review). When the MC item format was introduced, it was favored over the subjective oral and written exams. The MC item offered an advancement that was greatly appreciated until the 1960s, when the fairness of MC items was questioned because of poor testing practices that led to

M.C. Rodriguez (✉)
University of Minnesota, Minneapolis, MN 55455, USA
e-mail: mcrdz@umn.edu

inequitable placement of black and poor children in special education programs (Williams, 1970). There was a renewed interest in the promise of performance assessments that led to their adoption in classrooms and some large-scale testing programs in the 1970s through the 1990s – whereby performance assessments were viewed as an innovation and panacea for addressing the achievement gap; they were, for a brief time, the new assessment model. Because of costs, poor quality and psychometric performance, and little evidence of improved equity and fairness, performance assessments have once again fallen out of favor for large-scale, high-stakes testing.

Access and Assessment as Policy Tools

Through a series of legislative and federal administrative rules and regulations (see Zigmond & Kloo, 2009, for a review), there has been intense attention to issues related to inclusion, fairness, equity, and access to instruction and assessments that support multiple educational and accountability purposes. Several research centers and federally funded research projects have devoted substantial resources to addressing the most recent regulations to accommodate students with modified and alternate academic achievement standards in NCLB (No Child Left Behind) assessments. These researchers have contributed to a sparse volume of research on item writing. Although such research has been conducted for nearly 90 years at this point, the science of item writing is relatively undeveloped compared to the science of scoring, scaling, and equating. However, psychometric attention to alternate assessments is renewing interest in item-writing research, as a means of improving accessibility at the item level (Rodriguez, 2009).

This chapter will review existing guidance on item writing, the available body of item-writing research, recent research on item writing and test development with alternate assessments, and, briefly, innovations in item types and their relevance to accessibility. In the broad context of item writing and test development, ensuring test accessibility is a simple exercise

of employing best practice. Good item writing, founded on what limited evidence exists, supported through the application of current item-writing guidance, is implicitly designed to ensure accessibility. Good item writing has always been about accessibility. Other promising areas, including language learning, cognitive load theory, cognition and learning models, universal design, and instructional design, support this goal of test accessibility (many of which are reviewed in this book). Since the majority of test items are in the MC format, this will be the focus of this chapter. However, similar issues related to constructed-response (CR) items and performance tasks will receive comment when appropriate.

MC Formats

There are many formats of MC items, including innovative types, enabled through computer-based testing. Similarly there are several ways to score each of these formats (more on scoring later). The most common formats for MC items include the following (based on formats presented by Haladyna, Downing, & Rodriguez, 2002):

Conventional MC

Dropping *poor* items from a test on the basis of item analysis data alone may result in

- A. lower content-related validity.
- B. lower score reliability.
- C. a more heterogeneous test.

Multiple True-False

You are a psychometrician and providing advice to a novice test author. The test author asks several questions about improving test score reliability. Are these true or false concepts that, done in isolation, may improve reliability?

1. Adding more test items of similar quality improves test score reliability.
2. Increasing the sample size will improve test score reliability.
3. Obtaining a sample with more test score variability improves reliability.
4. Balancing positively and negatively worded stems improves test score reliability.

Alternate-Choice

Coefficient alpha is a more appropriate index of reliability for

- A. criterion-referenced test scores.
- B. norm-referenced test scores.

True-False

If a distribution of raw scores is negatively skewed, transforming the raw scores to z-scores results in a normal distribution.

Matching

Match each term on the right with the description on the left.

- | | |
|-------------------------------|------------------------|
| 1. score consistency | A. systematic error |
| 2. test-wiseness | B. random error |
| 3. score accuracy | C. item difficulty |
| 4. proportion correct | D. item discrimination |
| 5. point-biserial correlation | E. reliability |

F. validity

Context-Dependent Item Set

A college class of 200 students was given a final exam. The item analysis based on students' responses to five of the multiple-choice items is reported in the table. Response frequencies to correct options are in brackets []. For example, the correct option to item #1 is A, correctly answered by 125 of the 200 students.

Write the *item number* on the blank following questions #19 to #24.

Item number	Number of students selecting each choice				Corrected item-total correlation
	Choice A	Choice B	Choice C	Choice D	
1	[125]	25	20	30	.35
2	0	15	[185]	0	.15
3	[90]	25	15	70	-.40
4	40	30	[100]	30	.14
5	35	85	5	[75]	.30

1. Which item has the best discrimination? _____
2. Which item is the most likely candidate for revision into a true-false item? _____
3. Which item is the most difficult? _____
4. Which item is the most likely to have two correct answers? _____

Note: Each item above is a 5-option MC item.

Complex Multiple-Choice (Type K)

Which are criterion-referenced interpretations of test scores?

1. John's score is three standard deviations above the class mean.
 2. Mary answered 80% of the items correctly.
 3. Eighty% of the class scored above a T-score of 45.
 4. The average math score for Arlington High School is equal to the district average.
 5. Antonio is proficient in 5th grade reading.
- A. 1 and 3.
B. 1, 3, and 4.
C. 2 and 5.
D. 2, 4, and 5.
E. All 5.

CR Formats

The many forms of CR items are less systematically organized across the literature. For the most part, CR items are distinguished from MC items because they require the examinee to generate or construct a response. Many alternate forms exist, many of which are innovative items enabled through computer testing (described below). More traditional forms of CR items were classified by Osterlind and Merz (1994), described in the next section, in an attempt to create a taxonomy of CR items. Osterlind and Merz, as well as Haladyna (1997), described over 20 formats for CR items, including more common formats, such as essays, grid-in responses, research papers, short-answer items, and oral reports – as well as fill-in-the-blank and cloze procedures, which are not recommended for use in tests (Haladyna). They also included, what most measurement specialists would include within the range of performance assessments, considerably different forms of tasks like portfolios, performances, exhibitions, and experiments. These tasks require much more extensive scoring rubrics and substantially more time, as well as require planning and preparation time so they are typically not adaptable to on-demand testing.

There is no theoretical framework or taxonomic scheme that is all-encompassing across

the many forms of MC and CR items and performance assessment tasks to make meaningful distinctions in all cases. Items are not consistently differentiated by features of the items themselves, in response modes allowed for various item formats, or in the scoring processes for various item formats. In large-scale achievement tests, one typically finds grid-in items, short-answer items, and essay formats. In many cases, responses to such item formats can be objectively scored, particularly through the use of automated scoring (see, for example, Attali & Burstein, 2006). In alternate assessments (mostly for AA-AAS), alternate forms of responding to these more constrained CR items are allowable, including verbal responses, drawings, and word lists or construct maps.

Item-Writing Guidelines and Taxonomies

In the arena of accessibility, there is a renewed interest in accommodations, item writing, test modification, and assessment design. A number of resources are available, including recent special issues of journals, for example the *Peabody Journal of Education* (Volume 84, 2009) and *Applied Measurement in Education* (Volume 23, Number 2, 2010).

Every educational measurement textbook (of which there are dozens) contains one or more chapters on item writing. A number of chapters are designed to be somewhat exhaustive and instructive regarding item writing, including Chapters 12, 13, and 14 in the *Handbook of Test Development* (Downing, 2006; Welch, 2006; and Sireci & Zenisky, 2006; respectively) and more generally Chapters 9 and 16 in *Educational Measurement* (Ferrara & DeMauro, 2006; and Schmeiser & Welch, 2006). There are also a handful of books devoted to aspects of item writing – for example, *Writing Test Items to Evaluate Higher Order Thinking* (Haladyna, 1997) and *Developing and Validating Multiple-Choice Test Items* (Haladyna, 2004).

MC Item-Writing Guidelines

The first comprehensive taxonomy of MC item-writing guidelines was developed by Haladyna and Downing (1989a) with a companion piece summarizing the available empirical evidence (1989b). This taxonomy was revised to include additional empirical evidence and a meta-analytic review of some of that evidence (Haladyna et al., 2002). These reviews were largely informed from a review of item-writing advice from over two dozen textbook authors, with support from existing empirical literature. Most of these guidelines were based on logical reasoning and good writing practices; very few were based on empirical evidence. Guidance on item writing covers four elements of MC items: content, formatting and style, writing the stem, and writing the options.

Content concerns are paramount and the subject matter expert is the key to writing a successful item. Items should be carefully constructed to tap important relevant content and cognitive skills. These guidelines are largely based on logical argument and experience of item writers and examinee reactions. Aside from some general research on clarity and appropriate vocabulary use, there are no specific tests of these item-writing guidelines. The guidelines regarding content concerns include:

1. Every item should reflect specific content and a single specific mental behavior, as called for in test specifications (two-way grid, test blueprint).
2. Base each item on important content to learn; avoid trivial content.
3. Use novel material to test higher-level learning. Paraphrase textbook language or language used during instruction when used in a test item to avoid testing for simply recall.
4. Keep the content of each item independent from content of other items on the test.
5. Avoid over-specific and over-general content when writing MC items.
6. Avoid opinion-based items.

7. Avoid trick items.
8. Keep vocabulary simple for the group of students being tested.

Formatting and style concerns are largely based on general good writing practice. There is some empirical evidence to support general use of most item formats (Haladyna et al., 2002), whereas others, like the complex MC format, appear to introduce greater difficulty that may be unrelated to the construct being measured in most settings. There is evidence to support the recommendation regarding format of the MC item (guideline 9). Additional evidence is available regarding guideline 13. These guidelines include:

9. Use the question, completion, and best answer versions of the conventional MC, the alternate choice, true-false (TF), multiple true-false (MTF), matching, and the context-dependent item and item set formats, but AVOID the complex MC (Type-K) format.
10. Format the item vertically instead of horizontally.
11. Edit and proof items.
12. Use correct grammar, punctuation, capitalization, and spelling.
13. Minimize the amount of reading in each item.

Writing the stem is also an area with limited empirical evidence. With the exception of negatively worded stems, which the empirical results suggest should be rarely done, these guidelines are extensions of style concerns, specifically applied to the stem of the item. Although the work of Abedi and others provides evidence to support these guidelines, their work was not intentionally designed to test the validity of specific item-writing guidelines. The guidelines regarding writing the stem include:

14. Ensure that the directions in the stem are very clear.
15. Include the central idea in the stem instead of the choices.
16. Avoid window dressing (excessive verbiage).

17. Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface.

Writing the choices is the area with the largest volume of empirical evidence. Issues related to MC options were studied in the first published empirical study of item writing (Ruch & Stoddard, 1925). Since then, there have been dozens of published studies in this area. Nevertheless, of the 14 guidelines in the following, only 5 have been studied empirically (18, 24–27).

18. Develop as many effective choices as you can, but research suggests three is adequate.
19. Make sure that only one of these choices is the right answer.
20. Vary the location of the right answer according to the number of choices.
21. Place choices in logical or numerical order.
22. Keep choices independent; choices should not be overlapping.
23. Keep choices homogeneous in content and grammatical structure.
24. Keep the length of choices about equal.
25. *None-of-the-above* should be used carefully.
26. Avoid *All-of-the-above*.
27. Phrase choices positively; avoid negatives, such as NOT.
28. Avoid giving clues to the right answer, such as:
 - a. Specific determiners including always, never, completely, and absolutely.
 - b. Clang associations, choices identical to or resembling words in the stem.
 - c. Grammatical inconsistencies that cue the test-taker to the correct choice.
 - d. Conspicuous correct choice.
 - e. Pairs or triplets of options that clue the test-taker to the correct choice.
 - f. Blatantly absurd, ridiculous options.
29. Make all distractors plausible.
30. Use typical errors of students to write your distractors.

31. Use humor if it is compatible with the teacher and the learning environment.

In the context of accountability, most of these guidelines provide for clarity, efficiency in word usage, and largely employ good writing practices that should provide for the greatest accessibility to the widest audience of examinees. For those guidelines that have an evidence base, the direction of the guidance promotes the easier form of an item that supports the intended inferences (validity) compared to an alternative. This research supports the goal of accessibility and is in line with principles of universal design. This evidence is briefly reviewed below.

Constructed-Response Item-Writing Guidelines

Guidelines for writing CR items are less developed and less uniform across sources, with even less empirical evidence. An early attempt to formalize a framework for CR items was offered by Bennett, Ward, Rock, and LaHart (1990). In this work, they differentiated item types based on the degree to which openness is allowed in the response. They tested this framework by assessing the ability of judges (27 test developers) to reproduce expected item classifications (46 items) according to the framework and the ability of judges (16 test developers) to assess scoring objectivity. Their framework provided a scheme for categorizing items, in order of (0) multiple-choice, (1) selection/identification, (2) reordering/rearrangement, (3) substitution/correction, (4) completion, (5) construction, (6) presentation/performance. Judges relatively succeeded on their agreement of item classifications; however, there were disagreements on every item, even among the MC items. Oddly, there was greater agreement in scoring among items with more closed scoring keys, rather than those with open scoring keys. The authors surmised that open scoring keys provided for greater inclusion to a wide range of responses rather than more structured items which provided specific correct options.

A taxonomy for CR items was proposed by Osterlind and Merz (1994), who reviewed the work of cognitive psychologists, including Hannah and Michaels (1977), Snow (1980), Sternberg (1982), and others, as a framework for the multidimensionality of cognition. This taxonomy contained three dimensions – including (a) the type of reasoning competency employed, including (i) factual recall, (ii) interpretive reasoning, (iii) analytical reasoning, (iv) predictive reasoning; (b) the nature of cognitive continuum employed, including (i) convergent thinking and (ii) divergent thinking; and (c) the kind of response yielded, including (i) open-product and (ii) closed-product formats – producing 16 combinations. The first two dimensions address cognitive processes, whereas the third dimension addresses the kinds of responses possible. Closed-product formats are those that allow few possible response choices, possibly scored with a relatively simple scoring key; open-product formats permit many more choices, requiring scoring with more elaborate rubrics potentially allowing unanticipated innovative responses.

Most introductory educational measurement textbook authors provide guidance for CR item writing. Hogan and Murphy (2007) reviewed advice about preparing and scoring CR items from authors of 25 textbooks and chapters on educational and psychological measurement from 1960 to 2007. The 25 sources resulted in 124 points on preparing CR items and 121 points on scoring. They reference a handful of empirical research on these guidelines, but suggest that most are not based on empirical evidence. They suggest that it is important to assess the degree to which the recommendations affect the psychometric quality of tests and assessments. Among those most relevant for accessibility, the most frequently cited recommendation for preparing CR items was attention to testing time and length of the CR test (cited 20 times), followed by avoiding the use of optional items, defining questions or tasks clearly, relating content to instructional objectives, assuring items test complex processes, avoiding the use of CR items to test recall, and considering vocabulary, grammatical, and syntax appropriateness for the level of the test.

Major testing companies have developed CR item-writing guides to guide the work of their item writers. Notably, the work of Educational Testing Service (ETS) in their large-scale programs including CR items (e.g., National Assessment of Education Progress and Advanced Placement exams) has resulted in a large body of research on the quality of CR items, scoring, and scores. The ETS *Guidelines for Constructed-Response and Other Performance Assessments* (Baldwin, Fowles, & Livingston, 2005) provides a starting point. These guidelines are general, forming a basis from which more specific guidelines can be developed for specific testing programs. Most importantly, they argue that three elements must be in place prior to planning the design of CR items. Each of these elements speaks directly to equity, fairness, and accessibility (and applies more generally to all forms of assessment):

1. Make sure the people whose decisions will shape the assessment represent the demographic, ethnic, and cultural diversity of the people whose knowledge and skills will be assessed.
2. Make relevant information about the assessment available during the early stages so that those who need to know and those who wish to know can comment on the information.
3. Provide those who will take the assessment with information that explains why the assessment is being administered, what the assessment will be like, and what aspects of their responses will be considered in the scoring.

In his work on CR items and their ability to tap intended cognitive functions in the NAEP assessments, Gitomer (2007) defines CR items as “well-defined tasks that assess subject matter achievement and that ask examinees to demonstrate understanding through the generation of representations that are not prespecified and that are scored via judgments of quality” (p. 2). With respect to effective CR items and more generally to a wide range of performance assessment tasks, Gitomer argues that task demands are only clear in the context of the rubric, and that the meaning of the rubric is only clear in the context of the associated scoring process. Similarly, students

must understand what is being asked of them in the task and understand the response requirements, and the scoring system must consistently interpret the student’s response.

Both of these elements address accessibility. Gitomer argues that these requirements are typically satisfied for MC items, assuming the question is clear and students understand their task (i.e., good item writing). However, for CR tasks, these requirements are typically not satisfied, and he presents a framework for CR item design that is intended to secure these requirements in a coherent way. The intent is to ensure that students understand what is being asked of them and that scorers know how to interpret student responses appropriately. He presented seven guidelines for effective CR task development:

1. Justify the use of CR task formats; the tasks should require student thinking that cannot easily be obtained from MC or other fixed-choice formats.
2. Inferences should be explicit and construct-relevant; inferences about task demands and responses should not require implicit assumptions, such that they are a direct function of knowledge or skill on the construct of interest.
3. Distinctions should be clear across score points; clear distinctions regarding the features and quality of evidence to satisfy every score point are defined.
4. Justifications should be clear within score points; when there are different ways to accomplish a given score point, they should be cognitively equivalent.
5. Avoid over-specification of anticipated responses; combinations of response features may satisfy the required support for intended inferences even if every element required (through over-specification) to achieve a given score is not included in the response.
6. Empirically verify and modify task through pilot studies; empirical evaluation of student performance on the task and the associated rubric sufficiency needs to be considered. (Herein lies the examination of accessibility to students with varying levels of cognitive abilities.)

7. Construct measurement should be consistent across tasks; the expectations for the same cognitive aspects should be consistent across tasks within the assessment.

Choosing the Item Format

Rodriguez (2002) provided guidance regarding the choice of item format. He listed advantages and disadvantages of MC and CR items. Rodriguez argued for the primacy of purpose of the assessment as the guiding force regarding all test design stages, but particularly regarding choice of item format. For the purpose of rank-ordering academic abilities, the two formats appear to behave more alike than not. Although carefully constructed MC items can assess a wide range of cognitive processes, CR items more directly (and perhaps more easily) measure complex processes, providing for opportunities for a wider range of responses and novel solutions, particularly when the target objective requires a written response or extended explanation, reasoning, or justification. However, most CR items in operational tests behave far too much like MC items without options, making responses from the two formats nearly perfectly correlated (Rodriguez, 2003). In fact, if CR items are written like MC item stems, not really addressing different cognitive tasks, they will measure the same aspects of the construct as MC items, but with much less precision. Issues of accessibility apply equally to both formats.

Evidence

The evidence base for existing item-writing guidelines is sparse, as described earlier. Some of that evidence is briefly reviewed here. Evidence regarding item modifications to achieve the goals of accessibility is growing and is presented in the following section.

In their first comprehensive review of the empirical evidence on item writing, Haladyna and Downing (1989b) reported the frequency of support for each of their 43 item-writing guide-

lines, for which only seven had empirical evidence. These included

- avoid “none-of-the-above,”
- avoid “all-of-the-above,”
- use as many functional distractors as feasible,
- avoid complex item formats (Type-K),
- avoid negative phrasing,
- use the question or completion format,
- keep option lengths similar.

Except for the last three guidelines for which there was strong support, the others had mixed support (for and against). The issue of number of distractors was the most frequently investigated, so it will be treated separately in this review. The others were studied no more than 10 times in the 63 years of literature covered.

The largely narrative review of the item-writing guidelines was enhanced with an empirical meta-analysis of existing research by Rodriguez (1997). This analysis examined the overall effect of violating several item-writing guidelines based on published research. These results are summarized below. In a revision of the Haladyna and Downing (1989a) taxonomy of item-writing guidelines, Haladyna et al. (2002) reviewed 19 empirical studies published since 1990. They mostly found support for the existing taxonomy of item-writing guidelines and reduced the number of guidelines to 31. They also concluded that the plausibility of distractors is an area of item writing that is long overdue for empirical study.

Item-Writing Guidelines with Empirical Evidence

Not all of the seven guidelines have been investigated equally. Mueller (1975) was the only author who had examined the effects of inclusive alternatives (both none-of-the-above or NOTA and all-of-the-above or AOTA) and reported independent outcomes for AOTA. One limitation of Mueller’s design was that items were *not* stem equivalent across formats. In order to generalize these findings, we must assume that item difficulties were randomly distributed across all formats

Table 11.1 Summary of average effects of rule violations

Rule violation	Difficulty index	Discrimination index	Reliability coefficient	Validity coefficient
Using NOTA	-.035 ^a (.005) <i>n</i> = 57	-.027 ^b (.035) <i>n</i> = 47	-.001 ^b (.039) <i>n</i> = 21	.073 (.051) <i>n</i> = 11
Using negative stems	-.03 ^a (.010) <i>n</i> = 18		-.166 (.082) <i>n</i> = 4	
Using an open, completion-type stem	.016 ^b (.009) <i>n</i> = 17	-.003 ^b (.076) <i>n</i> = 6	.031 ^b (.069) <i>n</i> = 10	.042 ^b (.123) <i>n</i> = 4
Making the correct option longer	.057 ^{ab} (.014) <i>n</i> = 17			-.259 ^b (.163) <i>n</i> = 4
Using type-K format	-.122 ^{ab} (.011) <i>n</i> = 13	-.145 ^{ab} (.063) <i>n</i> = 10	-.007 ^b (.083) <i>n</i> = 4	

^aAverage effect is significantly different than zero

^bEffects are homogenous or consistent across studies based on the meta-analytic *Q*-test statistic

Note: Standard errors are in parentheses. The number of study effects is *n*

Source: Rodriguez (1997)

used. Mueller reported that items with AOTA were the least difficult among the formats examined (including NOTA and Type-K items) and where AOTA was the correct response, the item was very easy. The use of AOTA provides a clue to respondents with partial information, in that knowing two options are correct suggests that all are correct.

For five of the empirically studied guidelines, the results of the Rodriguez (1997) meta-analysis are summarized in Table 11.1. The empirical evidence of item-format effects was inconclusive. When rule violations resulted in statistically significant effects, they tended to be small. For example, using NOTA reduced the difficulty index overall (making items more difficult), but had no significant effect on item discrimination, score reliability, or validity. Using negatively worded stems made items more difficult and reduced score reliability, but with only four effects regarding reliability, this effect was non-significant. Using an open completion-type stem had no effect on any metric. Making the correct option longer (a common error made by classroom teachers) made items easier and reduced validity (although with only four effects, this was not significant). Finally, using the complex Type-K format made items much more difficult, reduced item discrimination, but had no overall effect on score reliability.

Three Options Are Optimal

A comprehensive synthesis of empirical evidence regarding the optimal number of options for MC items included 80 years of research (Rodriguez, 2005). Researchers included tests of K-12 achievement, postsecondary exams, and professional certification exams in areas including language arts, social sciences, science, math, and other specialty topics, where all but one of the researchers recommended three options for MC items (or four options if that was the minimum studied). Rodriguez found that moving from five or four options to three options did not negatively affect item or test score statistics. Negative effects were found if the number of options was reduced to two (as in alternate-choice MC items). The results were even more compelling when examining the effect of distractor deletion method used in the studies. When distractors were randomly deleted to create forms with fewer options, a significant decrease in score reliability was observed. When ineffective distractors were removed to go from five or four options to three options, there was no effect on score reliability.

In addition to the empirical evidence promoting the use of three-option items, several additional arguments support this approach; in particular, the reduction of cognitive load and

greater accessibility to the item by all students. Rodriguez (2005) argued that less time is needed to prepare three-option items and more items can be administered in the same time frame compared to four- or five-option items. Rodriguez reviewed several studies that demonstrated how many large-scale assessment programs exist where five-option items function like three-option items, since few items have more than two effective distractors. The overwhelming concern regarding the probability of random guessing likely prevents test designers from employing three-option items. However, since most items function like three-option items and the chance of getting a high score by random guessing on three-option items is remote, the evidence is compelling for its adoption. For example, on a 40-item test of three-option items (with .33 chance of randomly guessing correctly), the probability of getting 60% correct (score of 24/40) is .00033.

The most important consideration regarding the distractors in a MC item is to make them plausible, where the distractors are based on common misconceptions or errors of students, providing diagnostic information. Distractors are effective to the degree that they attract the right students, including students with misinformation, misconceptions, or errors in thinking and reasoning. We can then examine the frequency each distractor is selected to provide instructionally relevant information.

Research on Item Modifications for Accessibility

With federal requirements for complete inclusion in state NCLB assessments, attention to alternate assessments has increased substantially. Although states define the eligibility for participation in alternate assessments differently, there are two primary forms in which these assessments exist, including alternate assessments for alternate academic achievement standards (AA-AAS), generally for students with the most severe cognitive impairments, and alternate assessments for modified academic achievement standards (AA-MAS), generally for students with moderate

to severe cognitive impairments (persistent academic difficulties).

AA-AAS

The technical quality and effectiveness of AA-AAS have been examined in several ways. Initially, issues related to alignment between assessments and curriculum were examined for students with the most significant impairments (see, for example, Ysseldyke & Olsen, 1997). Others examined influences on student performance on these assessments, addressing challenges to validity of results, including technical quality, staff training, and access to general education curriculum (see, for example, Browder, Fallin, Davis, & Karvonen, 2003). Others have also examined assessment content given state standards (see, for example, Johnson & Arnold, 2004).

Recently, Towles-Reeves, Kleinert, and Muhomba (2009) reviewed 40 empirical studies on AA-AAS, updating a similar review completed earlier by Browder, Spooner, et al. (2003). They made several recommendations, such as (a) to include both content experts and stakeholders in studies to validate performance indicators (e.g., alignment studies), and (b) to design AA-AAS to produce data to inform instructional decisions. They argued that studies of technical quality of AA-AAS are few and limited and recommended more of this work as the first point on their suggested research agenda.

Finally, through their work developing technical documentation for states, Marion and Pellegrino (2006, 2009) offered an outline for AA-AAS technical documentation, particularly focused on the unique characteristics of AA-AAS in a validity-based framework. They created a four-volume model of technical documentation, including (a) a “nuts and bolts” volume that is similar to the typical technical manual for standard tests; (b) a validity evaluation combining the guidance from the *Testing Standards* (AERA, APA, & NCME, 1999), Kane (2006), Messick (1989), and others; (c) a stakeholder summary (also drawn from the first two volumes); and (d)

a transition document that provides procedures to ensure effective transition through changing contractors or state personnel. Although this effort was supported by their work on two federally funded collaborations with states on their AA-AAS, they suggested that this framework could be used for purposes of documenting technical evidence for AA-MAS. The first two volumes specify the kinds of technical information relative to demonstrating the quality of the AA-AAS, which is directly a result of accessibility for students with the greatest academic challenges.

AA-MAS

There has been substantially more attention to AA-MAS, perhaps due to the larger intended test-taker population. A special issue of *Peabody Journal of Education* (Volume 85) on AA-MAS recently provided several updates on the current state of affairs. Each of the articles, in some way, commented on technical adequacy of AA-MAS as currently employed. Much of the focus has investigated identification of appropriate participants, levels of participation, impact of accommodations, access to the general education curriculum, and overall performance. Kettler, Elliott, and Beddow (2009) provided a review of their work on a theory-guided and empirically based tool for guiding test modification to enhance accessibility. The Test Accessibility and Modification Inventory (TAMI, available at <http://peabody.vanderbilt.edu/tami.xml>) is guided by the principles of universal design, test accessibility, cognitive load theory, test fairness, test accommodations, and item-writing research (see Chapter 9, this volume).

The TAMI provides a rating system to evaluate accessibility, based on elements of (a) the reading passage or other item stimuli, (b) the item stem, (c) visuals, (d) answer choices, (e) page and item layout, (f) fairness, and (g) several aspects of computer-based tests, if applicable. The rating system is guided by a series of rubrics covering overall accessibility ratings on each of the above elements and an opportunity to recommend item modifications to improve

accessibility; several standard modifications are provided as a guide. The purposes of modifications are to ensure the use of strong item-writing guidelines to improve the item, to remove sources of construct-irrelevant variance, and to maximize accessibility for all students. They also recommend using cognitive labs to guide item and test modifications.

The TAMI has been employed in several states. Kettler et al. (in press) and Elliott, et al. (2010) have reported on results from multistate consortium research projects involving modification of state tests using the TAMI. These studies have found that modification preserves score reliability and improves performance of students (increases scores), and the performance of students eligible for AA-MAS increased more than for others. Rodriguez, Elliott, Kettler, and Beddow (2009) presented a closer analysis of the effect of modification on the functioning of distractors. The one common modification was a reduction of the number of options to three, typically done by removing the least functioning or least plausible distractors. In mathematics and reading items, the retained distractors became more discriminating. An important element in this line of research is that item modification typically involves multiple modifications. Because a package of modifications is employed, it is difficult to isolate the effect of any one modification. Given the large number of possible modifications and that each modification is tailored to the accessibility needs of the item, it makes little sense to uniformly apply any one modification across items (see Chapter 13, this volume for a review of research on packages of modifications). Moreover, items should be developed initially with maximizing accessibility as a goal, where TAMI could be used as an effective guide for item writers.

Innovations and Technological Enhancements

Computers and other assistive devices have been used as accommodations, providing a variety of means for individuals to respond to test questions. Computer-based testing also has provided

an open venue for a variety of new item formats, where innovations continue to be developed. Some have argued that these innovations have led to greater accessibility, and others have argued that innovative item types have improved the degree to which items tap the target construct. Sireci and Zenisky (2006) identified 13 item formats that are enabled by computer-based testing that also have potential for enhancing construct representation of the target of measurement and reducing construct-irrelevant variance. Although there is some evidence regarding the promise of these formats, we have not seen evidence regarding potential enhancement of accessibility. A couple of examples are provided here.

Extended MC items are typically found in items with reading passages, where each sentence in the passage provides an optional response to specific questions. The exam can contain questions about a reading passage, for example regarding the main idea of a paragraph, and the response is selected by highlighting the appropriate sentence in the reading passage. This format presents a large number of options (contrary to the recommendation of three-option items), but the options are the sentences within the reading passage itself – rather than rephrased ideas or out-of-context statements in the typical MC format. How these features interact and impact accessibility is unknown. Among the many innovative item types examined in the literature, there is little evidence regarding their ability to enhance accessibility.

Other formats include connecting ideas with various kinds of links (dragging and connecting concepts) and other tasks like sorting and ordering information. The computer environment allows for other innovative response processes, including correcting sentences with grammatical errors or mathematical statements, completing statements or equations, and producing or completing graphical models, geometric shapes, or trends in data. Computers provide a wide range of possibilities. Unfortunately, the development of these formats has been done without the investigation of their effect on accessibility and fairness.

Computer-enabled innovations have been examined in postsecondary and professional

exam settings. Innovations in the GRE and TOEFL have led to many studies of impact. Bennett, Morley, Quardt, and Rock (1999) investigated the use of graphical modeling for measuring mathematical reasoning in the GRE, where respondents create graphical representations. The results provided highly reliable scores that were moderately related to the GRE quantitative total score and related variables. For example, examinees may plot points on a grid and then use a tool to connect the points. Although examinees agreed that these graphing items were better indicators of potential success in graduate school, they preferred traditional MC items (which is a result commonly found when comparing MC and CR items in other settings).

Bridgeman, Cline, and Levin (2008) examined the effect of the availability of a calculator on GRE quantitative questions. They found that relatively few examinees used the calculator on any given item (generally about 20% of examinees used the calculator) and the effects on item difficulty were small overall. There were also no effects for gender and ethnic group differences. Bridgeman and Cline (2000) earlier examined response times for questions on the computer-adaptive version of the GRE, examining issues related to fairness. The issue regarding differences in item response time is critical in adaptive tests because examinees receive different items, depending on their response patterns. They found no disadvantage for students who received items with long expected response times. They also found a complication in that the item response time also depended on the point in which the item was administered in the test. Gallagher, Bennett, and Cahalan (2000) examined open-ended computerized mathematics tasks for construct-irrelevant variance. They hypothesized that experience with the computer interface might advantage some. Although no evidence of construct-irrelevant variance was detected, some examinees experienced technical difficulties and expressed preference for paper forms of the test. It appeared that it took much longer to enter complex expressions as compared to simply writing them out on paper forms.

Broer, Lee, Rizavi, and Powers (2005) looked for differential item functioning (DIF) among GRE writing prompts. They found moderate to large DIF for African American and Hispanic examinees among some items. In examining these prompts more closely, they found that sentence complexity and the number of different points made in the passage increase processing load and might explain DIF results in these cases.

In the context of accommodations (see Chapter 10, this volume for further background), assistive technologies have been used effectively in the area of reading test accessibility. The Technology Assisted Reading Assessment (TARA) project conducts research and development to improve reading assessments for students with visual impairments or blindness. It works in conjunction with the National Accessible Reading Assessment Projects (NARAP), through the Office of Special Education Programs and the National Center for Special Education Research (<http://www.narap.info/>). The *Accessibility Principles for Reading Assessments* (Thurlow et al., 2009) presents five principles with multiple specific guidelines addressing the implementation of each principle:

1. Reading assessments are accessible to all students in the testing population, including students with disabilities.
2. Reading assessments are grounded in a definition of reading that is composed of clearly specified constructs, informed by scholarship, supported by empirical evidence, and attuned to accessibility concerns.
3. Reading assessments are developed with accessibility as a goal throughout rigorous and well-documented test design, development, and implementation procedures.
4. Reading assessments reduce the need for accommodations, yet are amenable to accommodations that are needed to make valid inferences about a student's proficiencies.
5. Reporting of reading assessment results is designed to be transparent to relevant audiences and to encourage valid interpretation and use of these results.

There is a great deal of overlap between the *accessibility principles for reading assessments*

with the elements of the TAMI and foundational concepts in cognitive load theory and good item writing. Guideline 1-A requires understanding of the full range of student characteristics and experiences, so that item writers produce items “that have low memory load requirements” (p. 5). Guideline 1-B requires application of universal design elements at the test development stage, including “precisely defined constructs; nonbiased items; and simple, clear, and intuitive instructions and procedures” (p. 6). Guideline 2-D suggests developing criteria to specify when visuals are used or removed, recognizing the mixed results in the research literature on the use of visuals to enhance reading assessments. Guideline 3-B addresses item development and evaluation, where the “content and format of the items or tasks may be modified, to some extent to increase accessibility for all subgroups” (p. 14). This includes the importance of conducting item analysis and think-aloud studies, examining item functioning across relevant subgroups. Guideline 3-C speaks to test assembly, examining factors including “length of a test, the way items are laid out on a page, whether the test is computer-administered” (pp. 14–15). These principles and guidelines are intended to provide direction for future test design and “a road map for improving the accessibility of current assessments” (p. 23). The principles are a compilation of existing guidance and the first three principles largely reflect the guidance in the TAMI. The TAMI provides more detailed and explicit direction for item development and modification.

Alternative Scoring Models

Measurement specialists have argued that important information in most forms of items goes untapped. In part, alternative scoring methods are associated with alternate administration methods – the administration methods themselves can enhance scoring by providing additional information about respondents during the administration process. Attali, Powers, and Hawthorn (2008) investigated the effect of immediate feedback and revision on the quality of open-ended

sentence-completion items. As expected, being able to revise answers resulted in higher scores and the reliability of revised scores and correlations with criterion measures were higher. Writing high-quality accessible items or modifying existing items to maximize accessibility is important on the front end to secure responses that are construct-relevant, but the degree to which we obtain construct-relevant information is a function of scoring.

There are several alternate scoring methods for any given item format, including MC items. Most of these alternate methods are intended to capture partial knowledge, an idea that is underutilized when assessing students with limited access to the general curriculum and persistent academic difficulties. By ignoring information available in wrong responses, we lose what little information might be available about students with the most challenging academic learning objectives. For example, the ability to recognize that some options or responses are more correct than others should be rewarded. This may take place by allowing students to select more than one possibly correct response and obtain partial credit. Another option is to allow students to rate their confidence in the correctness of their answers, where the confidence rating can be used to weight responses.

Elimination testing was introduced in 1956 by Coombs, Milholland, and Womer, who allowed students to mark as many incorrect options as they could identify, awarding one point for each option correctly identified, with a deduction if the correct option was identified as incorrect. This process yields scores in a range from -3 to $+3$ (for four-option items), providing for a range of completely correct response to partial correct to completely incorrect. A related method allows students to identify a subset of options that includes the correct answer, with partial credit given based on the number of options selected and whether the correct option is in the selected subset. Chang, Lin, and Lin (2007) found that elimination testing provides a strong technique to evaluate partial knowledge and yields a lower number of unexpected responses (guessing that results in responses inconsistent with overall ability, in an Item Response Theory framework)

than standard number correct scoring. Bradbard, Parker, and Stone (2004) also found that elimination testing provides scores of similar psychometric quality, reduces guessing, measures partial knowledge, and provides instructionally relevant information. They noted that in some courses, the presence of partial or full misinformation is critical (e.g., health sciences), and instructors will have to be more purposeful when developing distractors, so that they reflect common errors and misconceptions (so that they attract students with partial information and meaning can be inferred from their selection). Bush (2001) reported on a liberal MC test method that is parallel to the elimination procedure, but instead of asking students to select the wrong options, they may select more than one answer and are penalized for incorrect selections. Bush reported that the higher achieving students liked the method (because of the additional opportunities to select plausibly correct options), but the lower achieving students strongly disliked it (mostly regarding the negative markings for incorrect selections).

Models allowing students to assign a probability of correctness to each option or assign confidence to their correct response have also shown to be useful. Diaz, Rifqi, and Bouchon-Meunier (2007) argued that allowing students to assign a probability of correctness to one or more options in MC items allows “the student to involve some of the choices that otherwise he wouldn’t consider, enriching his answer” (p. 63), and thus providing a closer picture of the student’s knowledge, skills, and abilities.

Pulling It All Together

As item writing and test design continue to improve, through the application of a wide range of accessibility principles, good item writing, and good test design practice, the full arena of assessment must work in a coherent fashion. This book covers much of this arena, from politics to classrooms to the technology of test design to obtaining student input. While item writing is only part of the equation, it is a core component, since everything else depends on the quality of the information captured by test items.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Attali, Y., & Burstein, J. (2006). Automated scoring with e-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3), 1–30.
- Attali, Y., Powers, D., & Hawthorn, J. (2008). *Effect of immediate feedback and revision on psychometric properties of open-ended sentence-completion items* (ETS RR-08-16). Princeton, NJ: Educational Testing Service.
- Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed-response and other performance assessments*. Princeton, NJ: Educational Testing Service.
- Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (1999). *Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning* (ETS RR-99-21). Princeton, NJ: Educational Testing Service.
- Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items*. Princeton, NJ: Educational Testing Service. ED 395 032.
- Bradbard, D. A., Parker, D. F., & Stone, G. L. (2004). An alternate multiple-choice scoring procedure in a macroeconomics course. *Decision Sciences Journal of Innovative Education*, 2(1), 11–26.
- Bridgeman, B., & Cline, F. (2000). *Variations in mean response times for questions on the computer-adaptive GRE general test: Implications for fair assessment* (ETS RR-00-07). Princeton, NJ: Educational Testing Service.
- Bridgeman, B., Cline, F., & Levin, J. (2008). *Effects of calculator availability on GRE quantitative questions* (ETS RR-08-31). Princeton, NJ: Educational Testing Service.
- Broer, M., Lee, Y.-W., Rizavi, S., & Powers, D. (2005). *Ensuring the fairness of GRE writing prompts: Assessing differential difficulty* (ETS RR-05-11). Princeton, NJ: Educational Testing Service.
- Browder, D. M., Fallin, K., Davis, S., & Karvonen, M. (2003). Considerations of what may influence student outcomes on alternate assessment. *Education and Training in Developmental Disabilities*, 38(3), 255–270.
- Browder, D. M., Spooner, F., Algozzine, R., Ahlgrim-Delzell, L., Flowers, C., & Karvonen, M. (2003). What we know and need to know about alternate assessment. *Exceptional Children*, 70, 45–61.
- Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, 25(2), 157–163.
- Chang, S.-H., Lin, P.-C., & Lin, Z. C. (2007). Measures of partial knowledge and unexpected responses in multiple-choice tests. *Educational Technology & Society*, 10(4), 95–109.
- Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement*, 16(1), 13–37.
- Diaz, J., Rifqi, M., & Bouchon-Meunier, B. (2007). Evidential multiple choice questions. In P. Brusilovsky, M. Grigoriadou & K. Papanikolaou (Eds.), *Proceedings of Workshop on Personalisation in E-Learning Environments at Individual and Group Level* (pp. 61–64). 11th International Conference on User Modeling, Corfu, Greece. Retrieved September 25, 2010 from <http://hermis.di.uoa.gr/PeLEIGL/program.html>
- Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Mahwah, NJ: Lawrence Erlbaum.
- DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.
- Ebel, R. L. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 185–249). Washington, DC: American Council on Education.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., et al. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children*, 76(4), 475–495.
- Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 324–345). Westport, CT: Praeger Publishers.
- Gallagher, A., Bennet, R. E., & Cahalan, C. (2000). *Detecting construct-irrelevant variance in an open-ended, computerized mathematics task* (ETS RR-00-18). Princeton, NJ: Educational Testing Service.
- Gitomer, D. H. (2007). *Design principles for constructed response tasks: Assessing subject-matter understanding in NAEP* (ETS Unpublished Research Report). Princeton, NJ: Educational Testing Service.
- Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn & Bacon.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 1, 37–50.
- Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 1, 51–78.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15, 309–334.
- Hannah, L. S., & Michaels, J. U. (1977). *A comprehensive framework for instructional objectives*. Reading, MA: Addison-Wesley.

- Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, 20(4), 427–441.
- Johnson, E., & Arnold, N. (2004). Validating an alternate assessment. *Remedial and Special Education*, 25(5), 266–275.
- Jones, L. V., & Olkin, I. (Eds.). (2004). *The nation's report card: Evolution and perspectives*. Bloomington, IN: Phi Delta Kappa Educational Foundation.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). New York: American Council on Education, Macmillan.
- Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, 84, 529–551.
- Kettler, R. J., Rodriguez, M. R., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (in press). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and the interaction paradigm. *Applied Measurement in Education*.
- Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluation the technical quality of alternate assessment. *Educational Measurement: Issues and Practice*, 25(4), 47–57.
- Marion, S. F., & Pellegrino, J. W. (2009, April). *Validity framework for evaluation the technical quality of alternate assessments based on alternate achievement standards*. Paper presented at the annual meeting of the national council on measurement in education, San Diego, CA.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education, Macmillan.
- Mueller, D. J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. *Educational and Psychological Measurement*, 35, 135–141.
- Osterlind, S. J., & Merz, W. R. (1994). Building a taxonomy for constructed-response test items. *Educational Assessment*, 2(2), 133–147.
- Rodriguez, M. C. (1997, March). *The art & science of item writing: A meta-analysis of multiple-choice item format effects*. Paper presented at the annual meeting of the American educational research association, Chicago, IL.
- Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 213–231). Mahwah, NJ: Lawrence Erlbaum.
- Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40(2), 163–184.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3–13.
- Rodriguez, M. C. (2009). Psychometric considerations for alternate assessments based on modified academic achievement standards. *Peabody Journal of Education*, 84, 595–602.
- Rodriguez, M. C., Elliott, S. N., Kettler, R. J., & Beddow, P. A. (2009, April). *The role of item response attractors in the modification of test items*. Paper presented at the annual meeting of the national council on educational measurement, San Diego, CA.
- Ruch, G. M., & Stoddard, G. D. (1925). Comparative reliabilities of objective examinations. *Journal of Educational Psychology*, 16, 89–103.
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 324–328). Westport, CT: Praeger Publishers.
- Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Lawrence Erlbaum.
- Snow, R. E. (1980). Aptitude and achievement. In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade. New Directions for Testing and Measurement* (Vol. 5, pp. 39–59). San Francisco: Jossey-Bass.
- Sternberg, R. J. (1982). *Handbook of human intelligence*. Cambridge, MA: Cambridge University Press.
- Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Teachers College, Columbia University.
- Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., et al. (2009). *Accessibility principles for reading assessments*. Minneapolis, MN: National Accessible Reading Assessment Projects. Retrieved September 25, 2010.
- Towles-Reeves, E., Kleinert, H., & Muhomba, M. (2009). Alternate assessment: Have we learned anything new? *Exceptional Children*, 75(2), 233–252.
- Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303–327). Mahwah, NJ: Lawrence Erlbaum.
- Williams, R. I. (1970). Black pride, academic relevance & individual achievement. *The Counseling Psychologist*, 2(1), 18–22.
- Ysseldyke, J. E., & Olsen, K. R. (1997). *Putting alternate assessments into practice: What to measure and possible sources of data* (Synthesis Report No. 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Zigmond, N., & Kloo, A. (2009). The “two percent students”: Considerations and consequences of eligibility decisions. *Peabody Journal of Education*, 84, 478–495.

Jamal Abedi

Issues concerning assessments of English language learner (ELL) students are among the top national priorities as the number of these students is rapidly increasing in the nation. According to a recent report by the U.S. Government Accountability Office, about 5 million ELL students are currently enrolled in schools, representing approximately 10% of all U.S. public school students (GAO, 2006). Nationally, ELL enrollment has grown 57% since 1995, whereas the growth rate for all students has been at less than 4% (Flannery, 2009). This rapid growth demands that we consistently and accurately determine which students require English language services (Abedi & Gándara, 2006).

However, literature on the assessment of English language learners clearly and consistently shows a large performance gap between ELL and non-ELL students in all content areas (Abedi, 2006; Abedi & Gandara, 2006; Solano-Flores & Trumbull, 2003; Wolf et al., 2008). The performance gap decreases as the level of language demand of test items decreases (Abedi, Leon, & Mirocha, 2003). Thus, the unnecessary linguistic complexity of assessments may interfere with ELL students' ability to present a valid picture of what they know and are able to do.

J. Abedi (✉)
Graduate School of Education, University of California,
Davis, CA 95616, USA
e-mail: jabedi@ucdavis.edu

Language is an important component of any form of assessments. However, a distinction must be made between language that is part of the assessments, and therefore relevant to the construct being measured, and language that may be irrelevant to the focal construct. Unnecessary linguistic complexity may affect the overall accessibility of assessments, particularly for students who are non-native speakers of English. For example, English language learners may have the content knowledge, but may have difficulty understanding the complex linguistic structure of content-based assessments such as those measuring student's mathematics and science content knowledge. Therefore, the measure of a student's content knowledge may be confounded with their level of English language proficiency. To make assessments more accessible for ELL students, test items must clearly communicate the tasks that students are supposed to perform.

The Concept of Linguistic Accessibility in Content-Based Assessments in English

Researchers have shown that large performance gaps exist between ELL and non-ELL students throughout content areas. These performance gaps widen as the language demand of the test items increase. In subject areas with high levels of language demand, such as reading/language

arts, the performance gap between ELL and non-ELL students is the largest. The gap decreases in content areas such as math and science, as the language demand of the items decrease within these subjects (Abedi, 2010; Abedi, 2006; Abedi, 2008; Solano-Flores, 2008).

The main issue with language on assessments is relevancy of the language to the content being assessed. In content areas where the focal construct is not language, the concept of language accessibility could be applied. Assessments could be modified to make them linguistically accessible to all students, particularly those at the lower level of English proficiency spectrum.

Assessments are linguistically accessible if they do not contain unnecessary linguistic complexity. Some of the linguistic features that are discussed in this chapter affect understanding of test items but may not be content related and may not be relevant to the content being measured. Linguistically accessible assessments are less impacted by such features. For example, as Abedi (2010) explained, accessible reading assessments use: (1) familiar or frequently used words, (2) words that are generally shorter, (3) short sentences with limited prepositional phrases, (4) concrete rather than abstract or impersonal presentations, (5) fewer or no complex conditional or adverbial clauses, and (6) fewer or no passive voices. In general, as indicated above, these assessments avoid any linguistic complexities that are not part of the focal construct.

The main focus of this chapter is the impact that language factors have on the assessment of English language learners and on the linguistic accessibility of content-based assessments. I first discuss the research evidence that points to the impact that language has on content-based assessments. I then introduce the linguistic features that have been shown to have major impacts on ELL student performance in content-based assessments. Finally, I discuss research-based approaches to control for the impact of unnecessary linguistic complexity on the assessment of ELL students.

The Nature and Impact of Language Factors on Content-Based Assessment Outcomes

Performance outcomes of ELL students in content areas such as math and science may be confounded with their level of English proficiency. That is, ELL students may not have the language facilities to understand the assessment questions or express their content knowledge within open-ended questions written in English. The main problem behind the language factors on these assessments is the fact that these factors may differentially impact the performance of ELL and non-ELL students. For many non-ELLs,¹ the language of assessment may not be a key issue as it is for ELL students. Therefore, the comparability of assessment outcomes for ELL and non-ELL students may seriously impact the validity of assessments. In situations like these where assessment outcomes are differentially impacted by a source or sources of construct-irrelevant variance, the outcomes from the different groups cannot be combined. Aggregating outcomes from different subgroups seriously violates the comparability assumption across the subgroups and it may also violate assumptions underlying the classical measurement theory. The source of variance due to measurement error is assumed to be random within the classical theory of measurement (Thorndike, 2005), but in the assessment of ELLs the linguistic and cultural factors may be a systematic source of error or bias (see, Abedi, 2006 for a detailed discussion of the violation of the assumption underlying the classical test for ELL students). To develop accessible assessments for ELL students, it is imperative to identify sources of linguistic biases that pose threats to the validity of assessment outcomes for these students.

¹ Non-ELLs are those who are considered as proficient in English, which includes native speakers of English, non-native speakers who are identified as initially fluent in English and ELL students who were reclassified as fluent English speakers.

Language factors unrelated to the construct being measured strongly affect the academic assessment of ELL students. However, reasoned conclusions about the relevancy of language factors within an assessment are difficult to formulate. A team of experts may be able to make such judgments. This team would include content, language, and assessment experts who would base their decisions on their expert judgment and research evidence from studies on the assessment and accommodation of ELL students (see Abedi, 2006). The following sections of this chapter define linguistic complexity of the assessments, illustrate the impact of unnecessary linguistic complexity on content-based assessments, and show how a distinction can be made between linguistic features that are related and those that are unrelated to the focal construct. In the next sections, I provide an example of an original version of a math item and a revised version of the item in which the level of unnecessary linguistic complexity was reduced.

Creating More Linguistically Accessible Assessments

Research has shown that language factors significantly impact student performance, particularly performance of students with more challenging academic careers, such as English language learners and students with disabilities (Abedi, 2010). The impact language carries on performance becomes more critical when language is not the target of assessment and the linguistic complexity of assessment becomes a source of construct-irrelevant variance (Abedi, 2006). In this section, I discuss the concept of “unnecessary” linguistic complexity in assessments and introduce the linguistic features that could affect the accessibility of assessments. I then discuss research-based approaches that are used to revise test items to reduce linguistic complexity.

Unnecessary Linguistic Complexity in Assessments

Language is an important part of any assessments without which the assessment system (test questions and student responses) would be incomplete. For example, when presented with math word problems, students have to read and comprehend the math questions and share the details of how they propose to solve a particular problem. Language may also be present as an unessential part of the assessment, but may help facilitate understanding of assessment items. For example, test items presented in a rich language context would be more interesting for test takers. However, sometimes the language used for context in the assessment questions may be excessively complex and may cause misunderstanding and confusion. Therefore, it is of paramount importance to distinguish between the language that is a natural part of the assessment and essential to the assessment process and the language that is unrelated to the assessment process.

The first step in this process would be to identify linguistic features in the test items that could potentially impact the assessment outcomes and then provide strategies to control for these sources of systematic measurement error (bias). In the next section I discuss sources of linguistic complexity of assessments.

Linguistic Features That May Affect Accessibility of Assessments

Studies focusing on the assessment of, and the accommodations for, ELL students have identified linguistic features that may not be related to the focal construct and may seriously impact student performance outcomes for those students who are challenged by excessive linguistic complexity within assessments (for a more detailed discussion of these features see Abedi, 2006; Abedi, Lord, & Plummer, 1997). These features slow down the reader, increase the likelihood of

misinterpretation, and add to the reader's cognitive load, thus interfering with concurrent tasks. Linguistic features that affect performance outcomes of ELL students include the following:

Linguistic feature	Definition
<i>Word frequency/familiarity</i>	Words that are high on a general frequency list for English are likely to be familiar to readers.
<i>Word length</i>	Longer words are more likely to be morphologically complex.
<i>Sentence length</i>	Sentence length is an index for syntactic complexity and comprehension difficulty.
<i>Voice of verb phrase</i>	Passive voice constructions are more difficult to process than active constructions and can pose a challenge for non-native speakers of English.
<i>Length of nominals</i>	A reader's comprehension of long nominal compounds may be impaired or delayed by problems in interpreting them.
<i>Complex question phrases</i>	Longer question phrases occur with lower frequency than short question phrases, and low-frequency expressions are harder to read.
<i>Comparative structures</i>	Comparative constructions have been identified as potential sources of difficulty for ELLs.
<i>Prepositional phrases</i>	Languages such as English and Spanish may differ in the ways that motion concepts are encoded using verbs and prepositions.
<i>Sentence and discourse structure</i>	Two sentences may have the same number of words, but one may be more difficult due to the syntactic structure.
<i>Subordinate clauses</i>	Subordinate clauses may contribute more to complexity than coordinate clauses.
<i>Conditional clauses</i>	Separate sentences, rather than subordinate "if" clauses, may be easier to understand.
<i>Relative clauses</i>	Since relative clauses are less frequent in spoken English than in written English, some students may have had limited exposure to them.
<i>Concrete vs abstract presentations</i>	Information presented in narrative structures tends to be understood better than information presented in expository text.
<i>Negation</i>	Sentences containing negations (e.g., no, not, none, never) are harder to comprehend than affirmative sentences.

For a detailed description of the features, along with research support of these features, see Abedi et al. (1997).

Procedures for Linguistic Modification of Test Items

The identification of the linguistic sources that raise accessibility issues for ELL students led to the development of a linguistic modification methodology for ELL assessments. Using the linguistic features presented in Fig. 12.1, two rubrics were developed to judge, identify, and quantify these features (see Abedi, 2006;

Abedi et al., 1996). The first rubric was used to assess the level of linguistic complexity of each test item on each of the 14 linguistic features. This rubric provides an *analytic rating* for each of the test items (Abedi, 2006, Figure 12.1 of this Chapter). For example, based on this rubric, the "word frequency/familiarity" in the test item is judged on a 5-point Likert-Scale ranging from 1 (words that are frequently used and are quite familiar to the test takers) to 5 (words that are least frequently used and are quite unfamiliar to the test takers).

The second rubric provides an assessment of the overall complexity of the test items

Fig. 12.1 Holistic item rating rubric. Adapted from Abedi (2006)

LEVEL	QUALITY
1	<p>EXEMPLARY ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Familiar or frequently used words; word length generally shorter • Short sentences and limited prepositional phrases • Concrete item and a narrative structure • No complex conditional or adverbial clauses • No passive voice or abstract or impersonal presentations
2	<p>ADEQUATE ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Familiar or frequently used words; short to moderate word length • Moderate sentence length with a few prepositional phrases • Concrete item • No subordinate, conditional, or adverbial clauses • No passive voice or abstract or impersonal presentations
3	<p>WEAK ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Relatively unfamiliar or seldom used words • Long sentence(s) • Abstract concept(s) • Complex sentence/conditional tense/adverbial clause • A few passive voice or abstract or impersonal presentations
4	<p>ATTENTION ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Unfamiliar or seldom used words • Long or complex sentence • Abstract item • Difficult subordinate, conditional, or adverbial clause • Passive voice/ abstract or impersonal presentations
5	<p>PROBLEMATIC ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Highly unfamiliar or seldom used words • Very Long or complex sentence • Abstract item • Very difficult subordinate, conditional, or adverbial clause • Many passive voice and abstract or impersonal presentations

(Abedi, 2006). A 5-point Likert-Scale rating is used to provide a *holistic* rating of the linguistic complexity of the test items. The Likert-Scale ratings range between 1 (exemplary items with simple linguistic structure) to 5 (problematic) on items that have very complex linguistic structures. For example, items with an overall rating of 5 use words that are highly unfamiliar or rarely being used; have very long sentences or complex sentences; are abstract; may

have very difficult subordinate, conditional, or adverbial clauses; and have many passive voice and abstract or impersonal presentations.

After providing linguistic ratings, both in terms of individual linguistic features and overall linguistic complexity, items are grouped into two categories: those that are linguistically accessible and those that need serious revisions and restructuring. Items in the second group (need revisions) can then be put through a

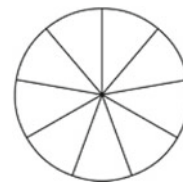
linguistic modification process. The linguistic modification process involves revising the language features that are judged to be unrelated to the construct being measured. However, the decision process of what is language related and unrelated is a challenging task and requires an expert's judgment. Therefore, a team of experts works together to first identify the complex language structure in the item that are not content related and then to propose linguistic revisions to make items more accessible for ELL students.

To illustrate the process of linguistic modification, a released test item from the National Assessment of Educational Progress (NAEP) is used. For illustration purposes only sources of grammatical complexity of the item are identified and modified. Although grammatical complexity is not a clearly defined construct (Rimmer, 2006), researchers have recommended investigating categories of words – verbs, conjunctions, nouns, etc. – to identify contributors to complex syntax used in the academic language of school (Hall, White, & Guthrie, 1986). Based on a review of literature and expert opinion, six features were examined in this question as possible indicators of grammatical complexity (Abedi et al., 2010):

- Passive verbs (PV)
- Complex verbs other than passive (CV)
- Relative clauses (RC)
- Subordinate clauses other than relative clauses (SC)
- Noun phrases (NP)
- Entities (EC)

Using these features provides a systematic means of analyzing text for grammatical complexity. I first present the original item (Fig. 12.2) with ratings on its grammatical complexity on all of the above six complexity features (Table 12.1) and then present the linguistically revised version of the item (Fig. 12.3) along with the ratings on the six features (Table 12.2).

The original test item contains a relative clause that begins with the potentially awkward combination of words “in which,” followed by the remainder of the lengthy clause “the chance of landing on blue will be twice the chance



Number of blues: _____

Number of reds: _____

Fig. 12.2 Original Item Luis wants to make a game spinner in which the chance of landing on blue will be twice the chance of landing on red. He is going to label each section either red (R) or blue (B)

of landing on red.” The directive, “Show how he could label his spinner” contains a complex verb (“could label”), but is not in the form of a question: a possible source of confusion for elementary-age students who use a question mark as a cue for what the problem requires. This NAEP-released test item also contains three noun phrases (“game spinner” – noun plus noun, and “the chance of landing on red” and “the chance of landing on blue,” both having the structure noun plus two prepositional phrases). The three entities are “Luis,” “the chance,” and “you” (understood in “You show how he could label his spinner”).

The grammatical complexity in this problem can be reduced by eliminating the relative and subordinate clauses and the complex verb form; syntax elements are often identified as sources of difficulty in reading (Larsen, Parker, & Trenholme, 1978; Haladyna & Downing, 2004). Although using character names in reading has not been specifically identified as a complex feature, presenting the problem as a simple directive to the student (You make a . . . spinner. . .) is a more straightforward presentation of what is required of the test taker. The requirement to count the red and blue sections is in

Table 12.1 Linguistic complexity rating for the original item

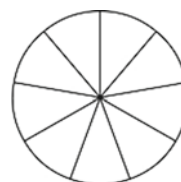
PV	CV	RC	SC	NP	EC	Total
	1	1	1	3	3	9

Fig. 12.3 Linguistically modified item

Make a red and blue game spinner and follow this rule:

The chance of landing on a blue section is two times the chance of landing on a red section.

How many blue sections and how many red sections do you make?



Number of blues: _____

Number of reds: _____

Table 12.2 Linguistic complexity rating for the revised item

PV	CV	RC	SC	NP	EC	Total
				3	3	6

question form, possibly bringing attention to what is required. The linguistically modified problem, shown below, is reduced from nine grammatically complex features to six.

Linguistic Modification: Practical Implications

The linguistic modification approach has been introduced as a language-based accommodation to make assessments more accessible for ELL students. As discussed in this chapter, this approach has been supported by research evidence as an effective and valid accommodation for ELL students. We will discuss the feasibility and logistics surrounding the implementation of this approach as an effective, valid, and relevant accommodation for ELL students. Unlike many other language-based accommodations that are implemented during the actual testing

process (such as providing a dictionary or glossary), the linguistic modification approach is implemented during the test development process. A team of experts carefully examines linguistic features of the newly developed tests and recommends changes in the linguistic structure of the items, if necessary. These recommendations are based on two important assumptions: (a) the recommended changes are not content related and do not alter the focal construct, and (b) they make the assessment more accessible to ELL students.

Research and Methodological Issues Related to the Linguistic Accessibility of Assessments

There are several methodological issues related to the linguistic modification approach for testing ELLs. I present a summary of research focusing on three areas: (a) research on the effectiveness of the linguistic modification approach for improving accessibility of assessments for ELLs, (b) research findings on the issues concerning reliability of linguistically modified tests, and (c)

research evidence on the validity of linguistically modified tests.

Research on the Effectiveness of Linguistic Modification Approach for Improving Accessibility of Assessments for ELLs

Findings from studies on the assessment and accommodations of ELL students suggest that the linguistic modification of test items can be an effective and valid accommodation for ELL students. In fact, among many accommodations used in several experimentally controlled studies, the linguistic modification accommodation was the only accommodation that reduced the performance gap between ELL and non-ELL students without compromising the validity of assessments (Abedi, Hofstetter, & Lord, 2004; Abedi, Lord, & Hofstetter, 1998; Abedi, Hofstetter, Lord, & Baker, 2000; Maihoff, 2002).

The effects of some of the linguistic features discussed previously were examined on a sample of 1,031 eighth-grade students in Southern California (Abedi et al., 1997). In this study, the math items for eighth-grade students were modified to reduce the complexity of sentence structures and to replace potentially unfamiliar vocabulary with more familiar words without changing the content-related terminologies (mathematical terms were not changed). The results showed significant improvements in the scores of ELL students and also non-ELLs in low- and average-level mathematics classes, but changes did not affect scores of higher performing non-ELL students. Low-frequency vocabulary and passive voice verb constructions were among the linguistic features that appeared to contribute to the differences. These features contributed to the linguistic complexity of the text and made the assessment more linguistically complex for ELL students.

The findings of this study were cross-validated in another study in which Abedi et al. (1998) examined the impact of linguistic modification on the mathematics performance of English learners and non-English learners using a sample of

1,394 eighth graders from schools with a high enrollment of Spanish speakers. Results were consistent with those of the earlier studies and showed that the linguistic modification of items contributed to improved performance on 49% of the items, with the ELL students generally scoring higher on shorter/less linguistically complex problem statements. The results of this study also suggested that lower-performing native speakers of English also benefited from the linguistic modification of assessment.

Other investigations also provided cross-validation evidence on the effectiveness of language modification approach in improving the validity of assessments for 1,574 eighth-grade ELL students. The effects of the language modification approach on reducing the performance gap between ELL and non-ELL students were examined in another study (Abedi, Courtney, & Leon, 2003) using items from the NAEP and the Third International Math and Science Study (TIMSS). Students were provided with either a customized English dictionary (words were selected directly from test items), a bilingual glossary, a linguistically modified test version, or the standard test items. Only the linguistically modified version improved performance of ELL students without affecting performance of non-ELL students. Maihoff (2002) also found linguistic modification of content-based test items to make assessments more accessible for ELL students. Kiplinger, Haug, and Abedi (2000) found that the linguistic modification of math items improved performance of ELL students in math without affecting performance of non-ELL students. Rivera and Stansfield (2001) compared ELL performance on regular and simplified fourth- and sixth-grade science items. Results of this study showed that linguistic modification approach did not affect scores of English-proficient students, indicating that linguistic modification is not a threat to score comparability.

In general, the research evidence shows linguistic complexity as a major source of measurement error on the assessment outcomes for ELL students. Research findings also suggest that reducing the level of unnecessary linguistic complexity of assessments helps to improve

assessment validity and reliability for these students and makes content-based assessments more accessible for ELL students.

Research Findings on Language as a Source of Measurement Error in Assessments

In classical measurement theory, “Reliability refers to the accuracy or precision of a measurement procedure” (Thorndike, 2005, pp. 109–110) or “consistency of measurement; that is, how consistent test scores or other assessment results are from one measurement to another” (Linn & Gronlund, 1995, p. 81). Reliability is directly affected by measurement error and there are many sources of measurement error that can affect reliability of test scores. Sources of measurement errors in paper-and-pencil tests can be from sources such as test format, pagination, font size, complex charts and graphs, and crowded pages. Unclear test instructions and problems with test administration and scoring may also be sources of measurement error and may impact reliability of assessments. Sources of measurement error often have random distribution and may directly or indirectly affect the validity of assessment, since reliability puts a limit on the validity of assessment (see, Allen & Yen, 1979; Thorndike, 2005, pp. 191, 192). However, there are sources of measurement error that may systematically affect reliability of a test which are often referred to as sources of “bias.” Unnecessary linguistic complexity within assessments is a major source of measurement error which systematically affects the reliability of assessments for ELL students.

To examine the impact of linguistic complexity on the assessment of ELLs, we compared reliability coefficients for ELL and non-ELL students on state assessments across several content areas, including mathematics and science. Since language factors may differentially impact performance of ELL and non-ELL students, we computed reliability coefficients separately for each of the two groups (ELLs and non-ELLs) using internal consistency approach. The main

limitation of the internal consistency approach in classical test theory, however, is the assumption of unidimensionality. That is, the internal consistency approach can only be applied to assessments that measure a single construct (see, for example, Abedi, 1996; Cortina, 1993). Unnecessary linguistic complexity, however, may introduce another dimension into the assessment, making the assessment multidimensional. The higher the level of impact of linguistic complexity of assessment, the more serious is the multidimensionality issue.

To examine the pattern of differences of the internal consistency coefficient between ELL and non-ELL students, data from several locations nationwide were analyzed (Abedi et al., 2003). Access to the item-level data provided an opportunity to examine item-total correlation and unique contribution of each item to the overall reliability of the test.

The results of internal consistency analyses showed a large gap in the reliability coefficients (alpha) between ELL and non-ELL students. The gap in the reliability between these two groups decreased as the level of language demand of the assessment reduced. The reliability coefficients for non-ELL students ranged from 0.81 for science and social science to 0.90 for math. For ELL students, however, alpha coefficients differed considerably across the content areas. In math, where language factors might not have as great an influence on performance, the alpha coefficient for ELL (0.80) was only slightly lower than the alpha for non-ELL students (0.90). However, in English language arts, science, and social science, the gap on alpha between non-ELL and ELL students was large. Averaging across English language arts, science, and social science, the alpha for non-ELL was 0.81 as compared to an average alpha of 0.60 for ELL students. Thus, as elaborated earlier, language factors introduce a source of measurement error, negatively affecting ELL students’ test outcomes while their impact on students who are native or fluent speakers of English is limited (for a more detailed discussion of reliability differences between ELL and non-ELL students, see Abedi et al., 2003).

Research Findings on the Impact of Linguistic Complexity on the Validity of Assessments for ELL Students

“Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). If the proposed use of content-based assessments as a measure of students’ knowledge in content areas such as mathematics and science is confounded by unnecessary linguistic complexity of assessments, then the outcomes of the tests may not be valid. For example, if a test that is intended to measure algebra in grade 8 has a linguistic structure so complex that ELL students have difficulty understanding the questions, then the test actually measures more than what is intended to measure. In this example, algebra content is the relevant or target construct and unnecessary linguistic complexity is the unintended or irrelevant construct. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) caution interpreting the test outcomes for individuals who have not sufficiently acquired the language of the test.

The impact of unnecessary linguistic complexity as a source of construct-irrelevant variance, which may differentially impact the validity of standardized achievement tests between non-ELL and ELL students, was examined in a study by Abedi et al. (2003). A multiple-group confirmatory factor analysis model was used to illustrate how linguistic complexity may impact the structural relationships between test items and the total test scores as well as between test scores and external criteria. In this study, test items were grouped into parcels, and those parcel scores were used to create latent variables. The results of analyses indicated major difference in the structural relationship between ELL and non-ELL students on relationship between test items and to total test scores, and between test scores and

external criteria. The correlations of item parcels with the latent factors were consistently lower for ELL students than they were for non-ELL students. This finding was true for all parcels regardless of which grade or which sample of the population was tested. For example, among ninth-grade ELL students, the correlation for the four reading parcels ranged from 0.72 to 0.78, across the two samples. In comparison, among non-ELL students, the correlations for the four reading parcels were higher, ranging from 0.83 to 0.86, across the two samples. The item parcel correlations were also larger for non-ELL students than for ELL students in math and science. Again these results were consistent across the different samples (see Abedi et al., for a detailed description of the study; see also Kane, 2006, for a discussion of criterion-related validity).

The data presented above clearly suggest that ELL students perform far behind their non-ELL peers. Studies have shown that the distortion caused by construct-irrelevant or “nuisance” variables, such as linguistic biases, may mainly be responsible for such performance gaps (Abedi, 2006).

Practical Steps for Improving Assessments for ELL Students

Formative Assessments to Help Identify Language Issues in the Assessments of ELL Students

Formative assessments can serve as useful tools for teachers and assessment developers to identify areas where ELL students may need language assistance (Abedi, 2009). The outcomes of summative assessments (e.g., state standardized assessments) can also be used formatively and can inform curriculum planning and instructional practices for ELL students. However, such outcomes may have limited utility for ELL students and their parents when it comes to understanding the language issues. Yet most studies on the impact of language factors on the assessments of ELL students are conducted on the data from end-of-year state standardized summative

assessments. While these data are helpful to identify sources of linguistic factors affecting performance of ELL students in general, they may not be helpful for individual student cases. Assessment outcomes would be too little too late for teachers to identify sources of problem and adjust their instructional plans to help ELL students become proficient enough in English to meaningfully participate in the mainstream instruction and assessments (Abedi, 2009).

Outcomes of formative assessments, if designed properly, can provide diagnostic information for teachers to assess areas that the ELL students need language assistance in. Such information may also help state assessment officials and test publishers to examine the accessibility issues in the current assessments and to plan necessary revisions to the current assessments to make them more linguistically accessible for these students.

Accessible Assessments at the Classroom Level Versus Accessible Assessments at the State or National Level

Language factors impact all assessments, whether used locally (in the classrooms), at the district or state level, or at the national level. Reducing

the level of unnecessary linguistic complexity is shown to make those assessments more accessible and helps provides a more clear interpretation of the assessment outcome. However, due to the much larger population of test takers at the state and national levels, more attention is usually paid to the large-scale assessment instruments, which has a larger impact on students population. Therefore, it would be extremely helpful to solicit input from teachers and students in the process of linguistic modification and incorporate those suggestions into the process along with the input from the linguistic modification team.

Guidelines and Recommendations for Creating More Linguistically Accessible Assessments for Students

As discussed earlier in this chapter, the practice of linguistically modifying assessments starts at the beginning of the item development process. A linguistic rubric, which has been developed by researchers, can guide the process (for an example of linguistic rubric, see Abedi, 2006, p. 392). For example, a test item can be rated on its linguistic complexity in a 5-point Likert Scale, 1 being less linguistically complex and 5 being complex. Below is the scoring rubric taken from Abedi, 2006, page 392, for rating “1”, least linguistic complex:

1	<p>Exemplary Item</p> <p><i>Sample features:</i></p> <ul style="list-style-type: none"> ● Familiar or frequently used words; word length generally shorter ● Short sentences and limited prepositional phrases ● Concrete item and a narrative structure ● No complex conditional or adverbial clauses ● No passive voice or abstract or impersonal presentations
----------	--

However, the process can be enhanced by introducing additional checks and balances. Some recommendations to improve the level of effectiveness of linguistic modification approach in making assessments more linguistically accessible for all students, particularly for ELL students, include the following:

1. Use a series of cognitive labs and focus groups on the target population for the newly developed assessments to identify linguistic structures that are considered as complex based on the input from the cognitive lab and focus group participants.

2. Ask a group of experts (linguist, content and measurement experts, and a teacher) to identify which of those linguistically complex statements are content related and which are unnecessary in measuring the focal construct.
3. Ask the group of experts to provide suggestions on how to linguistically modify the test items without altering the focal construct.
4. Ask content experts to review the linguistic revisions to make sure the linguistic structures related to the focal construct have not been altered. In their review of linguistic revisions, experts can use the linguistic modification rubric that has been discussed above.

Summary and Recommendations

There are many different factors that may affect accessibility of assessments for ELL students; among them linguistic factors play a major role on the accessibility of assessments for ELL students. Research on the assessment of ELL students has clearly linked the outcome of assessments with a student's language background. Students with a lower level of English proficiency may perform poorly in content-based assessments, mainly due to problems in understanding assessment questions and a difficulty in presenting their content knowledge in open-ending questions, rather than a lack of content knowledge. Thus, assessment outcomes are confounded with the student's level of English proficiency.

Researchers have illustrated that the linguistic modification of test items is a viable option to make assessments more accessible to ELL students. Under this approach, the linguistic structures of test items that are judged to be unnecessarily complex are modified based on a rubric that identifies different features of linguistic complexities. The distinction between language that is related and language unrelated to the focal construct is made by a team of experts including a linguist, a content expert, and a teacher. After all confounding linguistic features are identified, the team of experts provides suggestions

for revisions of the test items to make them more linguistically accessible to ELL students.

Research on the impact of the linguistic modification of test items discussed in this chapter has demonstrated that this approach is effective in making assessments more accessible to ELL students. ELL students who have received linguistically modified versions of tests have performed significantly and substantially higher than ELL students who were tested under the original version of test. That is, linguistic modification of test items improved the validity of assessments for ELL students. The performance of the non-ELL students under linguistically modified version of assessment remained the same, which is a good indication of the validity of linguistic modification approach since it does not alter the construct being measured.

The linguistic modification of assessments also helps improve the psychometric quality of the assessments. A major source of systematic error can be controlled by reducing the linguistic complexity of items, and the reliability of the assessment can be substantially improved. Summaries of the studies presented in this chapter demonstrate the impact language factors have on the reliability of assessments and how such factors can be controlled in order to improve the reliability of assessments. Similarly, results of studies presented in this chapter showed how linguistic modification improves the validity of tests by reducing the impact of unnecessary linguistic complexity as a source of construct-irrelevant variance.

To summarize, standardized achievement tests that are used for assessment and accountability purposes must control for the cultural and linguistic biases that threaten the validity of interpretation for ELL students. This chapter discussed the sources of threats due to linguistic complexities and approaches of how to deal with them. Finally, recommendations on how to make assessments more accessible to ELL students were provided.

References

- Abedi, J. (1996). The interrater/test reliability system (ITRS). *Multivariate Behavioral Research, 31*(4), 409–417.
- Abedi, J. (2006). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah, NJ: Lawrence Erlbaum.
- Abedi, J. (2008). Measuring students' level of english proficiency: Educational significance and assessment requirements. *Educational Assessment, 13*, 2–3.
- Abedi, J. (2010). Linguistic factors in the assessment of English language learners. In G. Walford, E. Tucker & M. Viswanathan (Eds.), *The sage handbook of measurement* (pp. 377–398). Oxford: Sage.
- Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Technical Report No. 608). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., & Herman, J. L. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teachers College Record, 112*(3), 723–746.
- Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for english language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.
- Abedi, J., Kao, J., Leon, S., Bayley, R., Ewers, N., Herman, J., et al. (2010). Accessible reading assessments for students with disabilities: The role of cognitive, linguistic, and textual features. *Manuscript in preparation*.
- Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of students' language background on content-based assessment: Analyses of extant data*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.
- Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Technical Report No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on english language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.
- Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological Testing*. Washington, DC: American Educational Research Association.
- Baugh, J. (1988). Review of the article twice as less: Black English and the performance of black students in mathematics and science. *Harvard Educational Review, 58*(3), 395–404.
- Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly, 1*(3), 79–132.
- Botel, M., & Granowsky, A. (1972). A formula for measuring syntactic complexity: A directional effort. *Elementary English, 49*, 513–516.
- Celce-Murcia, M., & Larsen-Freeman, D. (1983). *The grammar book: An ESL/EFL teacher's book*. Rowley, MA: Newbury House.
- Chall, J. S., Jacobs, V., & Baldwin, L. (1990). *The reading crisis: Why poor children fall behind*. Cambridge, MA: Harvard University Press.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.
- Finegan, E. (1978, December). *The significance of syntactic arrangement for readability*. Paper presented at the paper presented to the linguistic society of America, Boston.
- Freeman, G. G. (1978). *Interdisciplinary evaluation of children's primary language skills*. ERIC Microfiche, ED157341.
- Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*: Hillsdale, NJ: Erlbaum.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.
- Hall, W. S., White, T. G., & Guthrie, L. (1986). Skilled reading and language development: Some key issues. In J. Orasanu (Ed.), *Reading comprehension from research to practice* (pp. 89–111). Hillsdale, NJ: Erlbaum.
- Halliday, M. A. K., & Martin, J. R. (1994). *Writing science: Literacy and discursive power*. Pittsburgh, PA: University of Pittsburgh Press.
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (NCTE Research Report No. 3). Urbana, IL: National Council of Teachers of English.
- Hunt, K. W. (1977). Early blooming and late blooming syntactic structures. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 91–106). Urbana, IL: National Council of Teachers of English.
- Jones, P. L. (1982). Learning mathematics in a second language: A problem with more and less. *Educational Studies in Mathematics, 13*(3), 269–287.

- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixation to comprehension. *Psychological Review*, 87, 329–354.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30, 580–602.
- Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000). *Measuring math – not reading – on a math assessment: A language accommodations study of English language learners and other special populations*. Paper presented at the presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Larsen, S. C., Parker, R. M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. *Educational Studies in Mathematics*, 21, 83–90.
- Lemke, J. L. (1986). *Using language in classrooms*. Victoria, Australia: Deakin University Press.
- Linn, R. L., & Gronlund, N. E. (1995). *Measuring and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Lord, C. (2002). Are subordinate clauses more difficult? In J. Bybee & M. Noonan (Eds.), *Subordination in discourse*. Amsterdam: John Benjamins.
- MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32, 692–715.
- MacGinitie, W. H., & Tretiak, R. (1971). Sentence depth measures as predictors of reading difficulty. *Reading Research Quarterly*, 6, 338–363.
- Maihoff, N. A. (2002, June). *Using Delaware data in making decisions regarding the education of LEP students*. Paper presented at the paper presented at the council of chief state school officers 32nd annual national conference on large-scale assessment, Palm Desert, CA.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 200–220). Hillsdale, NJ: Erlbaum.
- Orr, E. W. (1987). *Twice as less: Black English and the performance of black students in mathematics and science*. New York: W. W. Norton.
- Pauley, A., & Syder, F. H. (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics*, 7(5), 551–579. doi:10.1016/0378-2166(83)90081-4.
- Rimmer, W. (2006). Measuring grammatical complexity: The Gordian knot. *Language Testing*, 23(1), 497–519.
- Rivera, C., & Stansfield, C. W. (2001, April). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students*. Paper presented at the paper presented at the annual meeting of the American educational research association, Seattle, WA.
- Schachter, P. (1983). *On syntactic categories*. Bloomington, IN: Indiana University Linguistics Club.
- Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Pearson, Merrill.
- Wang, M. D. (1970). The role of syntactic complexity as a determiner of comprehensibility. *Journal of Verbal Learning and Verbal Behavior*, 9(4), 398–404.

Effects of Modification Packages to Improve Test and Item Accessibility: Less Is More

13

Ryan J. Kettler

The final regulations for the *No Child Left Behind* (NCLB) (U.S. Department of Education, 2007a, 2007b) Act indicated that some students with disabilities may take an alternate assessment based on modified academic achievement standards (AA-MAS), a version of the general assessment test that has been modified to increase the validity of test score inferences for students identified with a disability. The AA-MAS policy is intended to help those students with disabilities who are exposed to grade level material but who because of their disabilities persistently fail to obtain proficiency, to better show what they know and can do. This assessment may be used for all eligible students, but only 2% of the entire student population may take the assessment and be reported proficient within a district or state.

Modifications is the term that refers to changes that are made to a test's content or item format, when evidence has not yet been gathered indicating whether the construct measured by the test has been preserved (Kettler, Elliott, & Beddow, 2009; Koretz & Hamilton, 2006; Phillips & Camara, 2006). Most states that are developing AA-MASs are doing so by using packages of modifications (e.g., removing one incorrect answer choice, removing unnecessary words, and

adding white space), rather than individual modifications in isolation, on items from the general examination in order to improve access to the exam for the eligible population. (The reader is directed to Chapter 9, this volume or to Beddow [in press] for examples of items prior to and following the application of a package of modifications.) The need to gather validity evidence for achievement tests that have been modified using these packages has inspired a growing body of research (Elliott et al., 2010; Kettler et al., in press; Roach, Beddow, Kurz, Kettler, & Elliott, 2010). Kettler (in press) described a framework for collecting and interpreting such evidence, and applied it to early research studies in the area; this chapter serves as an update of that research, with a narrower focus on those studies which involve packages of modifications. While research (Russell & Famularo, 2009; Smith & Chard, 2007) has been completed on individual modifications, this work is less representative of the process in which developers of state assessments currently are engaging.

The AA-MAS Policy

The AA-MAS policy is intended to help those students for whom the general achievement test would be too difficult, but for whom an alternate assessment based on alternate academic achievement standards (AA-AAS) would be too easy. The AA-AASs are designed for students with

R.J. Kettler (✉)
Department of Special Education, Peabody College of
Vanderbilt University, Nashville, TN 37067, USA
e-mail: ryan.j.kettler@vanderbilt.edu

significant cognitive disabilities who are unable to meaningfully complete tests, and are, therefore, typically assessed using rating scales, portfolios, or performance assessments (Elliott & Roach, 2007). Students for whom AA-MASs are intended are able to complete tests in standard formats, often with the support of modifications, but cannot be accurately assessed using a general assessment.

The new policy includes three criteria for identifying students for whom an AA-MAS would be appropriate: (a) each student must have an individualized education program (IEP) with content standards for the grades in which she or he is enrolled, (b) each student's disability must have precluded her or him from achieving proficiency as demonstrated on the state assessment or on another assessment that can validly document academic achievement, and (c) each student's progress in response to appropriate instruction and based on multiple measurements is such that even if significant growth occurs, the IEP team is reasonably certain that she or he will not achieve grade-level proficiency within the year (U.S. Department of Education, 2007a). Measurement for statewide proficiency, historically, has not been very precise for students with disabilities who would be eligible (SWD-Es) for an AA-MAS, theoretically because (a) their disabilities increase their sensitivity to barriers to assessment (e.g., long reading passages, confusing graphics, single items spread across multiple pages) that often are present on such tests and (b) the items are too difficult on average to give maximum information on these students (Kettler, in press).

The final regulations allow that AA-MASs may be less difficult, so long as they remain linked to grade-level content standards. Modifications resulting in less difficult items may yield tests that are more informative for SWD-Es, but evidence for such tests must go beyond changes in mean scores and item difficulty to more direct indicators of measurement precision. The *Standards for Educational and Psychological Testing* (the *Standards*, American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education

[NCME], 1999) present an approach to evaluating tests based on evidence regarding test content, response processes, relations to other variables, internal structure, and consequences of testing. (The reader is directed to Kettler [in press] for a review of these forms of evidence as they relate to research that has been done, or may be done in the future, to evaluate AA-MASs.) The approach and evidence types described in the *Standards* should be a guide in determining whether modified tests are more precise measures of the same constructs that are targeted by general achievement tests.

This chapter reviews research from (a) state technical manuals (Louisiana Department of Education, 2009b; Poggio, Yang, Irwin, Glasnapp, & Poggio, 2006 [Kansas]; Texas Student Assessment Program, 2007) and (b) experimental studies (Elliott et al., 2009, 2010; Kettler et al., in press) on the effects of packages of modifications to items and tests. Each of these studies addressed a different set of modification packages, question types, grade levels, and content areas; Table 13.1 depicts the characteristics addressed by the three state test technical manuals and three experimental studies that were primary sources of evidence for this review. Estimates from the general assessment are included for comparison, because the measurement of knowledge and skills using the general assessment taken by students who do not qualify for an alternate assessment remains the "gold standard" by which AA-MASs should be judged (Kettler, in press). Table 13.2 depicts coefficient alpha across the three states by test, content area, and grade level or grade band.

State-Modified Achievement Tests

By October of 2010, three states – Kansas, Louisiana, and Texas – had AA-MASs that had been approved by the federal government. All three states had technical data for their AA-MASs posted on the websites of the departments of education. Data related to these examinations are discussed next.

Table 13.1 Characteristics of AA-MAS studies

	Question types	Examples of modifications studied	Grades		
			Reading/ELA	Math	Science
AA-MAS technical manuals					
KAMM	MC	Remove a response option, select less items and less complex items	3–8, 11	3–8, 10	
LAA 2	MC, CR	Select items from general pool that are less cognitively complex	4–10	4–10	4, 8, 11
TAKS-M	MC, CR	Remove a response option, increase font, change layout, simplify text	3–8, 10	3–8, 10	5, 8, 10
Experimental studies					
CAAVES	MC	Remove a response option, simplify language, add graphic, change layout	8	8	
CMAADI	MC	Remove a response option, embed questions in passages, simplify graphics	7	7	
OASIS	MC	Remove a response option, voice-over, simplify language, simplify graphics			9–12

Note: ELA = English/language arts; AA-MAS = alternate assessment based on modified academic achievement standards; CAAVES = Consortium for Alternate Assessment Validity and Experimental Studies; CMAADI = Consortium for Modified Alternate Assessment Development and Implementation; KAMM = Kansas Assessment of Modified Measures; LAA 2 = Louisiana Educational Assessment Program Alternate Assessment, Level 2; OASIS = Operationalizing Alternate Assessment of Science Inquiry Skills; TAKS-M = Texas Assessment of Knowledge and Skills-Modified; MC = multiple choice; CR = constructed response

Kansas Assessments of Multiple Measures

The Kansas Assessments of Multiple Measures (KAMM) are modified assessments in reading and mathematics developed with the intention that items would be less complex than those used on the Kansas General Assessments (Poggio et al., 2006). In both reading and mathematics, the KAMM include less items than do the general assessments, and all of the items are in a multiple choice format with three response options. The reading test is composed of items independent from the general assessment. The mathematics test is composed of items selected from the general pool and modified, and students are allowed to use a calculator throughout it.

The technical manual for the KAMM includes reliability estimates in the form of coefficient alpha, as well as classification accuracy and classification consistency. These estimates tend to be similar in value to those observed on the general assessment. Coefficient alphas are in an acceptable range (i.e., < 0.80) and lower, but

close to those reported for the general assessment at most grade levels, with the exception of mathematics at the higher grades. The only validity evidence included in the KAMM technical manual is a description of the connection between grade-level indicators and the items selected for the mathematics test.

Louisiana Educational Assessment Program Alternate Assessment, Level 2

The Louisiana Educational Assessment Program Alternate Assessment, Level 2 (LAA 2) is composed of multiple choice and constructed response items selected from the general assessment pool based on data from the general education and special education populations. The LAA 2 covers reading, mathematics, science, and social studies across various content areas, using a smaller number of items than are used for the general assessment. Teams of teachers reviewed the items for appropriateness for the eligible population, potential for bias, content relevance,

Table 13.2 Coefficient alphas across three states by test, content area, and grade level or band

Grade level/band	Kansas		Louisiana		Texas	
	AA-MAS	General ^a	AA-MAS ^b	General	AA-MAS ^c	General
Reading/English/Language arts						
Third	0.86	0.88–0.90	–	0.93	0.82	0.89
Fourth	0.88	0.91–0.92	0.74 (0.83)	0.89	0.84	0.88
Fifth	0.89	0.88–0.92	0.68 (0.89)	0.92	0.87	0.87
Sixth	0.86	0.92–0.93	0.76 (0.92)	0.93	0.85	0.88
Seventh	0.86	0.92–0.94	0.78 (0.94)	0.93	0.86	0.91
Eighth	0.86	0.92–0.94	0.74 (0.84)	0.88	0.88	0.88
Ninth	–	–	0.76 (0.95)	0.95		
Tenth	0.90	0.92–0.93	0.81 (0.89)	0.87	0.86	0.91
Mathematics						
Third	0.87	0.91–0.93	–	0.90	0.84	0.88
Fourth	0.87	0.91–0.92	0.86 (0.90)	0.92	0.82	0.89
Fifth	0.87	0.91–0.92	0.79 (0.81)	0.86	0.85	0.89
Sixth	0.86	0.93–0.95	0.74 (0.80)	0.93	0.74	0.92
Seventh	0.82	0.94–0.95	0.71 (0.77)	0.89	0.76	0.92
Eighth	0.81	0.94–0.95	0.80 (0.85)	0.92	0.76	0.91
Ninth	–	–	0.69 (0.76)	0.91		
Tenth	0.75	0.94–0.95	0.72 (0.79)	0.92	0.69	0.94
Science						
Elementary school	–	–	0.82 (0.86)	0.86	0.84	0.84
Middle school	–	–	0.76 (0.80)	0.88	0.81	0.90
High school	–	–	0.76 (0.80)	0.86	0.79	0.91

^aA range of reliabilities were reported on the Kansas general assessment, because multiple forms were used

^bValues in parentheses are based on a Spearman-Brown estimate and the length of the general assessment

^cOn Texas's AA-MAS, English/language arts was measured at the high school level, and a stratified coefficient alpha was used. AA-MAS = alternate assessment based on modified academic achievement standards

cognitive complexity, and format (Louisiana Department of Education, 2009b).

Reliability evidence for the LAA 2 is reported as coefficient alpha, via an estimate based on the Spearman-Brown formula that projects what alpha would be if the test were the length of the general assessment, and also using a stratified alpha that may more appropriately characterize a test that includes multiple item types. In all 17 comparisons at various grades and content areas, coefficient alpha for the LAA 2 is lower than coefficient alpha for the general assessment, with the difference exceeding 0.10 in 12 comparisons (Louisiana Department of Education, 2009a, 2009b, 2009c). The technical summary includes three reasons for this difference: (a) the

LAA 2 tests are shorter, (b) the LAA 2 population is smaller, and (c) the LAA 2 population is more homogenous (Louisiana Department of Education, 2009b).

While it is unclear how a smaller population would systematically result in lower reliability estimates, it is definitely true that reliability estimates tend to be lower when using a more homogenous population. It is also true that the LAA 2 tests are shorter, and that shorter tests tend to be less reliable, which highlights a critical issue with the strategy that many states are using to develop an AA-MAS. While the policy is intended to improve measurement for eligible students, it is unlikely the reliability of scores representing these students' knowledge

and skills can be increased as the number of items is decreased. The Spearman-Brown estimate is, therefore, not appropriate for comparing reliabilities of shorter AA-MASs with those of longer general examinations, because it obscures the fact that the AA-MASs are less reliable at the lengths that are actually used. However, it is informative with regard to the types of reliabilities that *could* be obtained using AA-MASs that have as many items as the corresponding general assessments. The coefficient alphas for the LAA 2 that were calculated using the Spearman-Brown formula are much closer to the coefficient alphas for the general assessment, and only tend to differ and fall below the acceptable range in mathematics and science at the higher grade levels. The stratified alpha in every case across general assessments and AA-MASs is equal to or greater than the coefficient alpha, but is within 0.03. Validity evidence presented in the technical manual is limited to a description of the iterative process intended to maintain content validity, which featured a diverse group of experts and an emphasis on the statewide content standards.

Texas Assessment of Knowledge and Skills – Modified

The Texas Assessment of Knowledge and Skills – Modified (TAKS-M) is composed of multiple-choice and constructed response items, which originated from the general assessment, and were modified by the Texas Education Agency based on feedback from teachers and from a special task force (Texas Student Assessment Program, 2007). The agency used a list of modification strategies specified by content area to develop AA-MASs in reading/English language arts (e.g., simplify difficult vocabulary, divide section into meaningful units), mathematics (e.g., delete extraneous information, simplify graphics), and science (e.g., delete one part of a compound answer, provide appropriate formulas). The items were then subjected to reviews for appropriateness of modification, maintenance of the original construct, and connection to classroom instruction.

Reliability for the TAKS-M is reported as the Kuder-Richardson 20, which is equivalent to coefficient alpha but for dichotomous data, and as a stratified coefficient alpha, where appropriate. Values for these coefficients are lower in all cases than corresponding values from the general assessment, but tend to be similar and in the acceptable range on reading/English language arts tests and on all content areas at the elementary grade levels (Texas Education Agency, 2009; Texas Student Assessment Program, 2007). Reliability estimates tend to diverge more, with coefficients for the AA-MAS falling below the acceptable range, on mathematics and science tests at higher grade levels. The technical manual for the TAKS-M includes extensive information on the process used to ensure alignment to the content standards, as a method of maintaining content validity. The manual also includes evidence based on relations with other variables, indicating that the reading and mathematics tests are measures of related but separate constructs, and that neither measure was inappropriately related to demographic variables such as gender or ethnicity.

In addition to research to evaluate state AA-MASs, experimental studies of the effects of modification packages on items and tests have been completed as part of three federally funded projects. These studies are discussed next.

Experimental Studies of Modifications

The first large-scale, experimental research study on the validity of packages of test modifications was completed as part of the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES¹). The study influenced future studies by:

¹ CAAVES was a U.S. Department of Education Enhanced Assessment grant codirected by Elizabeth Compton and Stephen N. Elliott. Several studies on item modification were conducted within this multistate project during 2007–2009.

- (a) using packages of modifications designed by teams of educators to yield original and modified conditions of items without changing the connection to grade-level content standards. [The reader is directed to Kettler, Elliott, and Beddow (2009) for a description of this *Modification Paradigm*, as well as for a list of theory-based and research-supported modifications.],
- (b) including groups of eligible students and control groups identified by special educators based on federal criteria,
- (c) incorporating a within-subjects design that counterbalanced the order of conditions and of equal-length forms, and
- (d) analyzing the impact of modifications on reliability and student performance on tests and items.

Replication studies have been conducted as part of the Consortium for Modified Alternate Assessment Development and Implementation (CMAADI²), and Operationalizing Alternate Assessment for Science Inquiry Skills (OAASIS³) projects.

Consortium for Alternate Assessment Validity and Experimental Studies

As part of the CAAVES project, Elliott et al. (2010) administered short sets of items in reading and mathematics to eighth-grade students ($n = 755$) in four states. Each participant was part of one of three groups: students without disabilities (SWODs), students with disabilities who

would not be eligible for an AA-MAS (SWD-NEs), and students with disabilities who would be eligible (SWD-Es). Each student completed sets of equal length in a condition with original items, in which each student read unmodified items silently from a computer screen, and in two conditions with modified items: (a) Modified and (b) Modified with Reading Support. In the Modified condition students read modified items silently from a computer screen, and in the Modified with Reading Support condition some elements of the items (i.e., the directions, stem, and answer choices, but not passages or essential vocabulary) were read aloud to the students using voice-over technology. The two modified conditions in reading had only 72% of the total number of words, when compared to the Original condition, and the modified conditions in mathematics had only 74% of the total number of words.

The researchers used the Spearman-Brown prophecy formula to estimate the reliability coefficients of hypothetical test forms with 39 items, a length more representative of a state achievement test. Table 13.3 includes a summary of coefficient alpha estimates across studies and conditions. All reliabilities across groups, conditions, and content areas were acceptable, ranging between 0.85 and 0.94. The estimates were not meaningfully higher or lower for any group in the Original condition versus the two modified conditions, indicating that the modifications did not change the reliability of the test.

While large differences were not found in reliabilities across groups and conditions in the CAAVES study, it is possible that items in the Original condition were more accessible than those that are typically found on state general assessments (Kettler, in press). The average p-values (i.e., percent of items correct) in the Original condition (SWODs = 64%, SWD-NEs = 51%, and SWD-Es = 34%; Elliott et al., 2010) indicate that the difficulty of the test may have been more appropriate for a group of students with disabilities. An average p-value of 50% for the intended population of test-takers is ideal when trying to maximize discrimination. Therefore, the results of the CAAVES study may have overestimated the reliability of general

² CMAADI was a U.S. Department of Education General Supervision Enhancement grant codirected by Stephen N. Elliott, Michael C. Rodriguez, Andrew T. Roach, and Ryan J. Kettler. Several studies on item modification were conducted within this multistate project during 2007–2010.

³ OAASIS was a U.S. Department of Education Enhanced Assessment grant directed initially by Courtney Foster and ultimately by John Payne. Several studies on item modification were conducted within this multistate project during 2008–2010.

Table 13.3 Coefficient alpha estimates by condition, study, content area, and group

Grade level	Original	Modified	Modified with reading support
CAAVES reading (13 items projected to 39 items)			
SWODs	0.91–0.92	0.93–0.94	0.93–0.94
SWD-NEs	0.90–0.91	0.90–0.91	0.91–0.92
SWD-Es	0.88–0.89	0.88–0.89	0.88–0.89
CAAVES mathematics (13 items projected to 39 items)			
SWODs	0.88–0.89	0.89–0.90	0.89–0.90
SWD-NEs	0.86–0.87	0.86–0.87	0.86–0.88
SWD-Es	0.86–0.88	0.86–0.88	0.86–0.87
CMAADI reading (20 items projected to 40 items)			
SWODs	0.80	0.66	–
SWD-Es	–	0.52	–
CMAADI math (20 items projected to 40 items)			
SWODs	0.84	0.66	–
SWD-Es	0.77	0.77	–
OAASIS science (20 items projected to 60 items)			
SWODs	0.88	–	0.83–0.87
SWD-NEs	0.65–0.84	–	0.79–0.86
SWD-Es	0.65–0.85	–	0.76–0.78

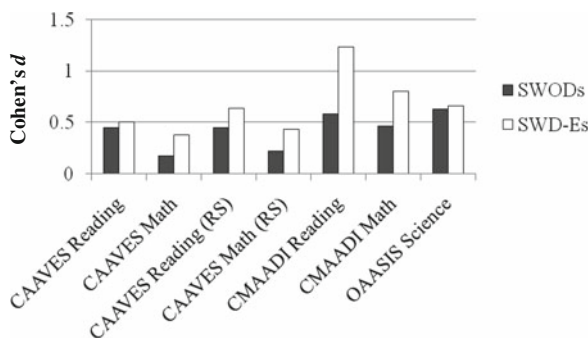
CAAVES = Consortium for Alternate Assessment Validity and Experimental Studies; CMAADI = Consortium for Modified Alternate Assessment Development and Implementation; OAASIS = Operationalizing Alternate Assessment of Science Inquiry Standards; SWODs = students without disabilities; SWD-NEs = students with disabilities – not eligible; SWD-Es = students with disabilities – eligible

achievement tests for SWD-Es, and consequently underestimated the potential improvement in reliability through modification.

Kettler et al. (in press) used an item response theory (IRT) framework on the same data set to test for differential boost based on modifications. Differential boost is a concept that has historically been used to evaluate testing accommodations, based on the idea that an appropriate accommodation should help students who need it more than those who do not (Fuchs et al., 2000; Kettler & Elliott, 2010). Differential boost fits within the domain of evidence based on relations with other variables, and in modification

studies would be apparent when the boost for eligible students is greater than the boost for one or more control groups. In statistical terms, a boost is expressed as an effect size, which reflects the magnitude of the relationship between variables – in this case, the condition (original versus a modified condition) and the test scores. A differential boost is found when the magnitude of boosts is significantly different for different groups. In the CAAVES study, the researchers found significant differential boosts between the Original condition and the Modified condition, indicating that SWD-Es benefited more from the modifications than did SWODs, when differences in ability

Fig. 13.1 Boosts by group across modification conditions, studies, and content areas. rs = reading support



level among groups were controlled in a Rasch model. Significant differential boosts were also apparent in this model when comparing SWD-Es with SWD-NEs in reading, or when comparing the Original condition with the Modified with Reading Support condition in mathematics.

Elliott et al. (2010) analyzed the same data within a classical test theory framework, finding boosts that were significant for all groups between the Original condition and the two modified conditions, but not finding any boosts that were *differentially* significant. Figure 13.1 depicts the boosts experienced by each group across modification conditions and content areas for the three experimental studies reviewed in this chapter. All effect sizes in Fig. 13.1 were computed using Cohen's d , an expression of the difference between mean scores in the two conditions, divided by the pooled standard deviation. Effect sizes for SWD-Es tended to be in the medium or high ranges, while effect sizes for SWODs tended to be in the small range. Collectively, results from the CAAVES study indicate that measurement precision can be maintained and scores can be improved through a modification process that keeps items connected to grade level content standards.

Consortium for Modified Alternate Assessment Development and Implementation

As part of the CMAADI project, Elliott and colleagues (2009) conducted a replication of the CAAVES study using original and modified sets of items from the Arizona Instrument to Measure

Standards. The participants included seventh-grade students representing two groups: SWODs ($n = 106$) and SWD-Es ($n = 46$). Each student completed two 20-item sets, which they read silently to themselves in both conditions, in either reading or mathematics. A MAZE measure of reading was administered to all students in order to determine whether low reading abilities are a likely barrier that keeps SWD-Es from accessing achievement tests. The SWD-E group correctly completed a significantly smaller number of choices on the MAZE; the effect size of the difference between the groups was in the medium range ($d = 0.72$).

The modification process in the CMAADI study appeared to yield better measurement for students who would be eligible for an AA-MAS (Elliott et al., 2009). Estimated coefficient alpha for a 40-item set improved for reading, although it remained below the acceptable range, and stayed nearly equal in mathematics. Item-to-total correlations increased by an average of 0.11 in reading, but only increased by an average of 0.01 in mathematics. A differential boost was observed in both content areas, as the effect of the modification process was significantly larger for SWD-Es than it was for SWODs in both reading and mathematics. These findings provide strong support for the impact of the modification process on reading, while the results for mathematics were mixed.

Operationalizing Alternate Assessment for Science Inquiry Skills

The OAASIS project replicated the design of the CAAVES and CMAADI studies in the context

of a high school biology test. Three groups of students – SWODs ($n = 168$), SWD-NEs ($n = 156$), and SWD-Es ($n = 76$) – were recruited across three states. Each student completed two 20-item sets of online multiple-choice items. The items were read silently by the students in the Original condition, but were read aloud by voice-over technology in the Modified condition. Items in the Modified condition contained, on average, only 70% of the total number of words included in the Original condition. A MAZE measure was administered to a subset of students, with SWODs correctly completing a significantly larger number of choices, compared to SWD-NEs ($d = 0.66$) and SWD-Es ($d = 0.83$). This finding replicated the finding from the CMAADI study, indicating that SWD-Es as a group do not read as quickly or as accurately as do SWODs.

The modification process in the OAASIS study yielded mixed results with regard to measurement precision for eligible students. Coefficient alpha for the two sets of items used in the study differed greatly in the Original condition, but alpha was reduced for one set through the modification process and increased for the other set. The projected coefficient alphas for the modified sets based on 60-item test forms (the median length of the general achievement tests in the participating states) were both slightly lower than the acceptable range. Boosts in total scores for both SWD-Es and SWODs were significant, but the boost for SWD-Es was not significantly larger than the boost for SWODs. These results indicate that the scores of SWD-Es are increased through the modification process, but that measurement precision is not systematically increased.

What Do We Know About Modified Achievement Tests?

The research completed on packages of modifications used to make a test more accessible has provided test developers and special education leaders a number of lessons:

- (a) The federal criteria for AA-MAS eligibility identify a population of students that attains

lower scores on tests across content areas, and reads less quickly and less accurately, compared to the general population of students without disabilities.

- (b) Packages of modifications designed to improve the accessibility of tests result in scores that are higher for both eligible and non-eligible groups of students across content areas.
- (c) This boost is sometimes significantly different in favor of the eligible group, and is sometimes relatively equal across groups, but does not typically favor the non-eligible groups (i.e., data indicate modifications do not *increase* the performance gap).
- (d) The reliability of AA-MASs for the eligible group tends to be similar to the reliability of the general achievement test at the elementary school grade band across content areas and in reading at higher grade bands. The reliability estimates, however, tend to diverge in mathematics and science at the middle and high school grade bands, with AA-MASs not being as reliable as general achievement tests and falling below the standard for acceptability.
- (e) A package approach to modifications appears to help improve the reliability of those reading tests which do not work well for SWD-Es in their original form; results are mixed in mathematics and science.

Collectively, these findings yield the conclusion that in the process of modifying items to make them more accessible, less is often more. The modification process is typically aimed at reducing the content of an item, often resulting in items that have less words, less cognitive load, and less answer choices, and which likely take less time for eligible students to complete. This is an important factor for a group of students who read slowly and therefore likely take longer to complete examinations. The modified tests allow students to be more successful, as quantified by higher scores, and to have an experience that is more similar to the experience of students without disabilities completing a general assessment. The tests developed using packages of modifications often have equal or greater measurement

precision for eligible students. Therefore, it is surely the case that when less is equal (in terms of measurement precision), and simpler (in terms of reading or cognitive load), and allows eligible students to have a more typical testing experience, less is better in the context of item and test accessibility.

Future Needs

Research on packages of modifications made to items and tests to improve accessibility has only been emphasized recently. The movement has been inspired by the final regulations of NCLB (U.S. Department of Education, 2007a, 2007b), which have pushed states that often do not have funds necessary to purchase new pools of items, to instead develop tests by improving items they already own. The findings are likely to become more positive as ineffective modifications are removed from consideration, yielding more effective packages of modifications overall. For example, exploratory analyses conducted as part of the CAAVES and OAASIS studies have indicated that adding graphics to reading tests or that changing passages to bulleted lists on science tests may reduce the measurement precision of these tests when used with the eligible population. Additional replications of the studies that have been completed to date, as well as similar studies on grade levels and content areas that have not been addressed, are critical to developing more accessible tests.

Subsequent studies on the effects of item and test modifications should also incorporate more information about the eligible population and their experience during testing. One example of information that should be collected on the eligible population is an indicator of working memory. The modifications made in the reviewed studies were often aimed at reducing cognitive load, based on the assumption that working memory limitations are one barrier to success on typical achievement tests. Test data on working memory should be gathered on students identified as belonging to eligible groups, in order to determine whether this assumption is true. An

example of data that should be collected on the testing experience is the amount of time taken to complete forms in original versus modified conditions. While it is highly likely that the modified versions consisting of less words and less answer choices take less time to complete, this should be verified and the magnitude of the difference should be quantified.

Conclusions

Tests developed using packages of modifications can already make the testing experience of students eligible for an AA-MAS more similar to the experience of students who take the general examination. While this experience does not appear to come at a cost in terms of reliability in reading, mathematics, or science at the elementary grade band, it may result in less precise measurement in mathematics and science at the middle and high school grade bands. Experimental research findings indicate that measurement precision can be maintained in these content areas and at these grade bands, and perhaps even improved. These findings are positive, and hold promise for greater improvement in the near future, resulting in even better methods that allow all students to show their knowledge and skills through accessible tests.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Beddow, P. A. (in press). Beyond universal design: Accessibility theory to advance testing for all students. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques*. Charlotte, NC: Information Age Publishing.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., et al. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children, 76*(4), 475–495.
- Elliott, S. N., & Roach, A. T. (2007). Alternate assessments of students with significant cognitive disabilities: Alternative approaches, common technical

- challenges. *Applied Measurement in Education*, 20(3), 301–333.
- Elliott, S. N., Rodriguez, M. C., Roach, A. T., Kettler, R. J., Beddow, P. A., & Kurz, A. (2009). *AIMS EA 2009 pilot study*. Nashville, TN: Learning Sciences Institute, Vanderbilt University.
- Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children*, 67(1), 67–81.
- Kettler, R. J. (in press). Holding modified assessments accountable: Applying a unified reliability and validity framework to the development and evaluation of AA-MASs. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques*. Charlotte, NC: Information Age Publishing.
- Kettler, R. J., & Elliott, S. N. (2010). Assessment accommodations for children with special needs. In E. Baker, P. Peterson & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 530–536). Oxford, UK: Elsevier Limited.
- Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, 84, 529–551.
- Kettler, R. J., Rodriguez, M. R., Bolt, D. M., Elliot, S. N., Beddow, P. A., & Kurz, A. (in press). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*.
- Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). United States of America: American Council on Education and Praeger Publishers.
- Louisiana Department of Education. (2009a). *iLEAP 2009 technical summary*. Baton Rouge, LA: Author.
- Louisiana Department of Education. (2009b). *2009 LAA 2 technical summary*. Baton Rouge, LA: Author.
- Louisiana Department of Education. (2009c). *LEAP GEE 2009 technical summary*. Baton Rouge, LA: Author.
- Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733–757). United States of America: American Council on Education and Praeger Publishers.
- Poggio, A. J., Yang, X., Irwin, P. M., Glasnapp, D. R., & Poggio, J. P. (2006). *Kansas assessments in reading and mathematics: Technical manual*. The University of Kansas: Center for Educational Testing and Evaluation. Retrieved November 2, 2009, from <http://www.ksde.org>
- Roach, A. T., Beddow, P. A., Kurz, A., Kettler, R. J., & Elliott, S. N. (2010). Incorporating student input in developing alternate assessments based on modified academic achievement standards. *Exceptional Children*, 77(1), 61–80.
- Russell, M., & Famularo, L. (2009). Testing what students in the gap can do. *Journal of Applied Testing Technology*, 9(4).
- Smith, R. L., & Chard, L. (2007). *A study of item format and delivery mode from the California Modified Assessment (CMA) pilot test*. Educational Testing Service: Statistical Analysis and Research.
- Texas Education Agency. (2009, updated May 7). *Technical digest 2007–2008*. Austin, TX: Author. Retrieved November 5, 2009, from <http://www.tea.state.tx.us>
- Texas Student Assessment Program. (2007). *Texas assessment of knowledge and skills-modified: Technical report*. Austin, TX: Author. Retrieved November 2, 2009, from <http://www.tea.state.tx.us>
- U.S. Department of Education. (2007a). *Modified academic achievement standards: Non-regulatory guidance*. Washington, DC: Author.
- U.S. Department of Education. (2007b). *Standards and assessments peer review guidance*. Washington, DC: Author.

Including Student Voices in the Design of More Inclusive Assessments

14

Andrew T. Roach and Peter A. Beddow

We hardly know anything about what students think about educational change because no one ever asks them. . . . The information is negligible as to what students think of specific innovations that affect them. To say that students do not have feelings and opinions about these matters is to say that they are objects, not humans (Fullan, 2001, pp. 182–189).

Due, in part, to changes in federal policy (e.g., Individuals with Disabilities Education Act [IDEA] and No Child Left Behind [NCLB] Act), the past two decades have seen a dramatic change in the number of students included in state and district accountability systems. According to the National Center on Educational Outcomes (NCEO), in the early 1990s most states reported that fewer than 10% of students with disabilities participated in their states' large-scale assessment. By the year 2000, the average percentage of students with disabilities in the general assessment had risen to 84%, and by 2008 that number had risen to above 95% (i.e., the participation rate required by NCLB).

There are a number of potentially positive consequences of increased participation in state accountability systems: (a) higher expectations

and improved performance in core academic areas; (b) increased access to the general grade-level curriculum; and (c) additional professional development and material resources allocated to improving the performance of students with disabilities.

State assessment data, however, suggest that there has been only modest movement toward “closing the gap” in proficiency between students with disabilities and their peers, and the differences remain very large – ranging from 31.4 percentage points in elementary school reading to 39.7 points in middle school mathematics (Albus, Thurlow, & Bremer, 2009). Surely, these numbers cannot tell the whole story of how inclusive accountability has influenced the educational experiences of students with disabilities. More research is needed to “break open the black box” of inclusive assessments and accountability systems, and to understand the ways in which implementing these programs influences students' knowledge, perceptions, and behaviors.

Concurrent with the movement to include students with disabilities in state tests, a number of researchers have undertaken efforts to apply universal design principles to assessment development (Dolan, Hall, Banerjee, Chun,

A.T. Roach (✉)

Department of Counseling and Psychological Services,
Georgia State University, Atlanta, GA 30302, USA
e-mail: cpsatr@langate.gsu.edu

Portions of this chapter appeared previously in Roach, A. T., Beddow, P. A., Kurz, A., Kettler, R. J., & Elliott, S. N., (2010). Incorporating student input in developing alternate assessments based on modified achievement standards. *Exceptional Children*, 77, 61–80.

& Strangman, 2005; Thompson, Johnstone, Anderson, & Miller, 2005). The goal of these efforts has been to design accessible assessment tasks and delivery systems that support, rather than inhibit, students' ability to show what they know and can do.

In April 2007, the U.S. Department of Education (USDOE) revised regulations under the No Child Left Behind Act and codified the pursuit of accessible assessment strategies by allowing states to develop alternate assessments based on modified academic achievement standards (AA-MAS). According to the USDOE *Non-Regulatory Guidance* (2007), features included in AA-MAS items should facilitate the understanding of students with disabilities, or provide background information and support in ways that do not "compromise the validity and reliability of the test results" (p. 26).

In essence, test developers and policymakers expect students' experiences with, and cognitions while, completing accessibility-enhanced assessments to be *different* from what happens when the same students take existing items on the general large-scale tests. Some information to support this assumption can be gathered from statistical analyses of test results (e.g., differential item functioning), but these methods can only provide quantitative evidence to support test item development. To understand the effects of universal design features intended to enhance item accessibility and students' performance, test developers and researchers must use a variety of methods that tap students' cognitions, problem-solving behaviors, and opinions.

Asking for student perspectives and feedback makes intuitive sense, if test developers' objective is to create the most effective and useful assessment instrument. In other industries, manufacturers regularly conduct consumer-focused evaluations to insure that their products are viewed as meeting customers' needs. Yet, as Fullan states in the epigraph to this chapter, we "hardly know anything about what students think about" educational assessments because test developers, researchers, and policymakers have seldom paused to ask them.

Epistemological and Methodological Frameworks for Including Student Voices

Integrating student voice in the development and validation of inclusive and accessible assessment strategies requires an expansion of the epistemological and methodological frameworks that dominate traditional measurement research. Research in measurement and psychometrics generally has been conducted from a positivist epistemology, which values precision and standardization in data collection, control of external and contextual influences, and limiting researcher subjectivity. Because of this epistemological stance, measurement research emphasizes the use of quantitative methods (i.e., statistical analyses) with large samples of participants (Gallagher, 2009).

Moss (1996) called for inclusion of research methods built on an interpretative framework in measurement research and practice. She suggested "(interpretative approaches) would enable us to support a wide range of sound assessment practices, including those less standardized forms of assessment that honor the purposes teachers and students bring to their work; (and) to theorize more productively about the complex and subtle ways that assessments work within the local contexts in which they are developed and used. . . ." (p. 20). Moss contrasts interpretative methods with a "naturalist" view of social science, which posits that social scientists should use methods similar to those employed by biological and physical scientists to study the natural world, and she suggests a variety of reasons why interpretative frameworks might be more appropriate for investigating different measurement and assessment programs and strategies (e.g., including inclusive and accessible assessments):

1. Assessment data generally consist of symbolic constructions – texts, products, performances, and actions – that reflect the behaviors, understandings, and interpretations of test takers and users.
2. Measurement researchers must inquire about the intentions and perspectives of test takers

and users because they cannot fully understand these from position of an outside, objective (i.e., distant) observer.

3. Interpretations (e.g., test scores, performance descriptions) that measurement researchers and test developers construct can be, and often are, reinterpreted by and subsequently influence the behaviors and beliefs of the students and teachers they describe.
4. A primary goal of measurement research should be to understand meaning in the context in which it is produced and received.

The National Research Council's *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001) also called for an expansion of theoretical and methodological approaches to educational assessment. Pellegrino and colleagues indicated that many widely used tests of educational achievement were built on outdated psychometric and cognitive models that did not reflect the most recent advances in either field. Moreover, they suggested assessments "should focus on making students' thinking visible to both their teachers and themselves so that instructional strategies can be selected to support an appropriate course for future learning" (p. 4). Achieving this goal is particularly relevant for students with (or at risk for experiencing) learning difficulties. Research methods that elicit students' written and verbal responses to tasks are necessary in order to "unpack" student cognition, identify supportive as well as confusing task features, and produce data that support current and future instructional decision making.

In 2001, The Spencer Foundation funded a series of conversations and ongoing collaboration between prominent psychometric and sociocultural researchers on the topic of educational assessment. The goal of this project was "two-fold: (a) to provoke examination of the influence of psychometrics on assessment and, in turn, on concepts of learning and teaching, educational reform, fairness, and 'scientifically based evidence'; and (b) to imagine how the processes of assessment might be better served or even reshaped through collaborative, cross-disciplinary efforts" (Moss, Pullin, Gee, &

Haertel, 2005). This group's examination of how scholarship regarding situated and distributed cognition might influence assessment research and design is particularly applicable to development and validation of inclusive and accessible assessments. In essence, researchers working from a situated/distributed framework view cognition as "not locked privately inside individual heads" but spread across interactions with other individuals and with tools and technologies. "What someone can do in one situation or with other people or tools and technologies can be quite different than what a person can do alone or across a wide array of contexts" (p. 71). These questions about what students can do with different levels and types of support are of the utmost importance in constructing more inclusive and accessible assessments. Unfortunately, traditional psychometric research is not well suited to investigating these sorts of context-bound behaviors; interpretative/qualitative methods (e.g., interviews, participant observation) are needed to develop "thick descriptions" of the effects of accommodations, universal design features, and other supports.

Uses of Student Response Data in the Standards for Testing

Support for integrating student input in developing and validating accessible assessments can be found in test standards as well as various professions' standards for ethics and professional practice. For example, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) are intended to guide the development and validation of testing practices in education and psychology and also include a relatively comprehensive overview of the rights and responsibilities of various stakeholder groups, including test developers and test users. The value of information regarding student responses and perceptions in supporting the development of inclusive and accessible assessments is addressed at multiple points in the *Standards for Educational and Psychological Testing*.

Standard 10.3, in the chapter on testing individuals with disabilities, indicates: “Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications” (p. 106). Because pilot testing often occurs with a smaller sample of participants, collecting information regarding student behaviors and cognitions during testing and their perceptions of assessment tasks may be more manageable than during actual implementation. Gathering student response data during pilot testing allows test developers to identify items with features students perceive as confusing. Identifying items and item features that may unintentionally influence and inhibit the performance of students with and/or without identified disabilities during pilot testing can reduce the unnecessary costs required to “retrofit” test forms and procedures during “live” testing.

The *Standards for Educational and Psychological Testing* suggest information about student response processes and test-taking behaviors can provide evidence to support the construct validity of an assessment. “Questioning test takers about their performance strategies can yield evidence that enriches the definition of a construct...” (AERA, APA, & NCME, 1999, p. 12). When attempting to develop accessible test items, student response data can provide important information about the reasons for observed differences in performance across item types (unmodified vs modified) and student groups (students with and without identified disabilities). The use of concurrent think-aloud protocols and follow-up questioning may allow researchers to “unpack” unexpected results. For example, differential item functioning may indicate a particular item was difficult for students with identified disabilities in comparison to their peers. Recording students’ concurrent verbalizations while solving the item in question and questioning students following completion of the task may illuminate item features that contributed to the observed results. The *Standards for Educational and Psychological Testing* identify the latter as an important

potential contribution of student response data: “Process studies involving examinees from different subgroups can assist in determining the extent to which capabilities irrelevant or ancillary to the construct may be differentially influencing (student) performance” (p. 12). This type of evidence is central in the development of more accessible tests, where test items may include features that are intended to reduce or eliminate construct-irrelevant influences on student outcomes.

Test developers also may collect data about student perceptions to provide consequential validity evidence. One desired outcome of the inclusion of universal design features in test development is that the resulting tests will be more accessible and comprehensible, leading to improved student motivation and sense of efficacy. The *Standards for Educational and Psychological Testing* address this claim, indicating “Educational tests... may be advocated on the grounds that their use will improve student motivation... Where such claims are central to the rationale of testing, the direct examination of testing consequences necessarily assumes even greater importance” (AERA, APA, & NCME, 1999, p. 17). Follow-up questioning and surveys can provide important information about the influence of accessible tests on student motivation and efficacy.

Ethics and Standards for Professional Practice

Developing and implementing inclusive and accessible assessments also could be conceptualized as a form of educational intervention. In this case, information about student experiences and perceptions are essential evidence about the acceptability of these inclusive assessment strategies. Acceptability refers to an individual’s perceptions regarding the appropriateness, fairness, and reasonableness of an intervention (Kazdin, 1981). Evaluating the acceptability of testing accommodations, test item modifications, or the integration of universal design features in test forms and delivery requires surveys or interviews

to understand the perceptions of students who qualify for these inclusive assessments strategies.

As the professionals most likely to make decisions about the provision of inclusive assessment strategies (e.g., testing accommodations) and more accessible alternate forms of assessment, special educators and school psychologists have an ethical responsibility to determine the acceptability of these supports. In *What Every Special Educator Must Know: Ethics, Standards, and Guidelines*, the Council for Exceptional Children (CEC) emphasized the need to consider how students' experiences, perceptions, and beliefs influence their learning and academic performance. For example, CEC's Initial Content Standard III (2008) states:

Special educators are active and resourceful in seeking to understand . . . individuals' academic and social abilities, attitudes, values, interests, and career options. The understanding of these learning differences and their possible interactions provides the foundation upon which special educators individualize instruction to provide meaningful and challenging learning for individuals with exceptional learning needs (p. 51).

Integrating information regarding students' social skills, perceptions, or occupational aspirations into decisions about their participation in large-scale testing calls for an expanded conceptualization of acceptability. Yet, it seems clear that these factors can have a dramatic influence on engagement and achievement on large-scale tests. For example, students with disabilities may less likely to view postsecondary education (or, in some cases, high school graduation) as feasible, leading to reduced motivation and effort on state-mandated high school graduation examinations. Within this context, educators' efforts to provide accommodations or to provide more accessible test forms may not result in passing scores and be (incorrectly) coded as ineffective because the acceptability and perceived utility of these efforts were never gauged with targeted students.

Similarly, school psychologists are expected to collaborate with students to develop supports and interventions like testing accommodations and accessible assessments. According to the National Association of School Psychologists' *Principles for Professional Ethics*:

"School psychologists discuss with students the recommendations and plans for assisting them. To the maximum extent appropriate, students are invited to participate in selecting and planning interventions" (NASP, 2010, II 3.11). Currently, decisions about participation in state assessment programs are made by students' IEP teams. In practice, these decisions often are made by special educators with limited consultation with school psychologists, family members, general educators, and students themselves. Although the ethics codes suggest that educators have a responsibility to include students with disabilities in decision making about how to be included in large-scale assessments, we unfortunately know very little about students' ability to select appropriate supports for taking and succeeding on large-scale assessments. Additional research is needed to determine if students can successfully identify and use accommodations and modifications that result in improved assessment performance.

Beyond these professional standards of practice and ethics, many educational researchers have been influenced by the United Nations Convention on the Rights of the Child (UNCRC) Article 12 that indicates educators and researchers should:

. . .Assure to the child who is capable of forming his or her own views the right to express those views in all matters affecting the child, the views of the child being given due weight in accordance with the age and maturity of the child.

Indeed, UNCRC Article 12 may provide the strongest endorsement for including student perspectives and opinions in the development and validation of the inclusive and accessible assessment strategies as it clearly recognizes children and adolescents' agency and right to self-determination.

Existing Assessment Research That Integrates Student Voice

Several research groups have conducted examinations of student responses to testing, using three primary methodologies: student drawings

(e.g., Wheelock et al., 2000), surveys (e.g., Roach et al., 2010), and various interview strategies (Roderick & Engel, 2001; Triplett & Barksdale, 2005; Roach et al., 2010; Johnstone, Liu, Altman, & Thurlow, 2007; Johnstone, Bottsford-Miller, & Thompson, 2006). Taken together, this body of research highlights the importance of integrating student voices in the design of assessments for students with a broad range of abilities and needs. In this section, we discuss the procedures and results of a set of recently published studies of students' perspectives on assessment. Our review focuses on two data collection methods – student drawings and interviews – that we believe hold the most promise for providing inclusive and accessible forums for students' voices to be heard and honored. We encourage researchers who are interested in student perspectives on inclusive assessments to consider the potential barriers to communication that may be inherent in some data collection methods (e.g., reading survey items, composing open-ended written responses) in designing studies.

Research Using Student Drawings: Two Examples

To understand students' experiences with and opinions of large-scale tests, Wheelock, Bebell, and Haney (2000) asked a sample of high school students to respond by drawing their reactions after participating in the Massachusetts Comprehensive Assessment System (MCAS). The MCAS is a series of paper-and-pencil tests consisting of multiple-choice and constructed response items. The MCAS is considered a high-stakes assessment system in that students are required to pass the test in order to receive a high school diploma. At the time of the study, the MCAS required students to sit for up to 13 one-hour test sessions in English/language arts, mathematics, science and technology, and history/social studies. In all grades, the authors reported the test times exceeded those recommended by the Massachusetts Department of Education.

The authors gathered participants by sending electronic mail to the Massachusetts teacher online, listserve, asking for their help in a study examining student perceptions about the MCAS. They sent a follow-up e-mail to respondents, directing them to ask their students to follow a simple prompt, as follows: "Draw a picture of yourself taking the MCAS." The researchers subsequently received 411 student drawings from 18 teachers in grades 4, 8, and 10, across eight school districts. Drawings were disaggregated by the type of community (i.e., urban, suburban, rural). They coded the drawings according to student postures, testing materials, and the inclusion of other students' or teachers' figures, as well as affective features (e.g., facial expressions that clearly denoted emotion). Additionally, the researchers collected comments written in speech or thought bubbles and captions. Interrater agreement on data coding exceeded 90%. The drawings "provided a variegated picture of how students...view high stakes testing" (Wheelock et al., 2000, p. 6). The majority of drawings depicted the test event as a solitary experience (i.e., 70% of the drawings did not include other figures in the drawings). Nearly two-thirds of the drawings contained indicators of students' emotions or perceptions while taking the test. Of these, the frequency of students who indicated the test was difficult was four times greater than those who indicated the test was easy. Some depicted a combination of hard and easy items. Question marks were depicted in 8.5% of student drawings; they were typically included in thought bubbles. In some, students pictured themselves asking for help from a teacher. For example, one drawing depicted the student asking, "Who was Socrates? Who was Socrates? What kind of question is that?" (Wheelock et al., 2000, p. 8). A greater percentage of urban students reported the MCAS as difficult (15.6%) when compared to rural students (5.7%).

A larger percentage of urban students (16.5%) depicted themselves taking a test that was "too long" compared to suburban and rural students (0 and 3.3%, respectively). One eighth grader drew a picture of herself with steam coming out of her ears as she sat before a test booklet

containing 6,021,000 pages. One fourth grader drew herself taking an even longer test, containing 1,000,956,902 pages, which she labeled “Stinkn’ [sic] test.” Other students’ drawings depicted themselves thinking or saying things such as “Five pages more,” “TO [sic] MUCH TESTING,” “Is it over yet?” and “Not MCAS again!” Several student drawings depicted the student feeling tired or rushed to complete the test.

The researchers coded 13.4% of the drawings as depicting the student experiencing anxiety. One student depicted himself thinking the test was “nerve-wracking.” Others drew themselves praying or wishing for help. Several drawings showed students who feared failing the test and having to go to summer school. Little discrepancies were observed in anxiety drawings across grade levels or community types. Ten percent of the drawings portrayed students feeling anger. One drawing depicted the student marching on City Hall. Another showed the student setting fire to the MCAS. Yet another depicted the student thinking the test was designed to reveal what he had failed to learn. Greater percentages of students in grades 8 and 10 (19.4%) depicted themselves as angry compared to students in grade 4 (6.6%). Four times as many urban students compared to rural students depicted themselves as angry.

Given the importance of test-taker motivation, it was notable that many drawings (5.3%) portrayed the student languishing throughout the course of a day, daydreaming about things unrelated to the test, or sleeping. Similarly, 3.9% of the drawings depicted students as relieved at the conclusion of testing. Some student drawings, by contrast, contained positive depictions (18%). A greater percentage of students in grade 4 (21.5%) depicted themselves as “thinking, solving problems, confident, or working hard” (p. 9) compared to students in grades 8 and 10 (8.3%). Greater percentages of urban and suburban students (21.1 and 20.1%, respectively) compared to rural students (9.7%) depicted themselves as diligent.

Triplett and Barksdale (2005) conducted a similar study in which they asked students in

grades 3 through 6 to “draw a picture about your recent testing experience.” Students then responded in writing to the prompt “tell me about your picture” (p. 237). Their rationale for the study was to discover “what they could learn about the test milieu from the primary stakeholders – the children” (p. 237).

For the total sample, no drawings depicted the student smiling. The majority of the drawings (56%) depicted the student in isolation. Emotions were depicted in 32% of the drawings, with *nervous* being the most frequent emotion depicted or discussed. Students frequently indicated their nervousness was the result of time pressure, not being able to find the correct answer, and failing the test. One student wrote, “I felt as if time was slipping through my fingers. I tried to stay calm....There was only 55 min to complete it. I thought I was a goner!”(p. 253). Another common emotion was *anger*, primarily over the length of the test, its difficulty, social isolation during the test event, and the possibility of failure. Fifteen of the drawings included depictions of fire, including one student who noted, “The school is gonna burn, we’re saved!”(Triplett & Barksdale, 2005, p. 251). Question marks were also a common feature in the drawings, usually with the purpose of indicating confusion. One student drew a detailed battle scene between question marks and light bulbs, centered around an answer sheet. He wrote, “I felt like there was a war going on in my head. The light bulbs won and the question marks lost!”(p. 250). Fourteen drawings included positive statements, using the words *happy*, *glad*, or *liked*; in most cases, the positive statements were connected with the completion or termination of the test event. No student drawings contained positive statements about the test itself.

Research Using Student Interviews: Four Examples

A number of researchers have examined student perceptions of large-scale testing through the use of various interview strategies. Roderick

and Engel (2001) interviewed 102 low-achieving students in grades 6 and 8 about Chicago Public Schools' policies ending social promotion (i.e., the implementation of standards-based tests to guide the determination of students' passage to the next grade). Roderick et al. found students mostly understood the main purposes of tests; indeed, some reported tests were useful and argued they should be used to determine which students "deserve to be retained" (p. 197). Only 4% of students interviewed, however, reported "not worrying" (p. 204) about passing the test. No students in the high-risk portion of the sample reported they did not worry about the test. By contrast, 52% of students in the high-risk group reported "Worrying a lot"; whereas, the percentages were lower for the moderate risk group (34%) and the low risk group (6%). Nine percent of the sample reported they spent time outside of class working on skills to help them pass the test. While a plurality of students for the total sample (53%) reported working hard in school, there was large variance across groups; 60% of the moderate and low-risk students reported working hard, while only 37% of the high-risk students reported working hard. Few differences were observed in any of the results across grades, genders, or race/ethnicity.

Beyond the information on students' opinions of Chicago Public Schools' retention policies, Roderick and Engel (2001) reported several student comments that provided some indication of their perceptions of the accessibility of tests and test items. For instance, when asked if these tests are hard or tricky for him, one student responded, "Hard. . .there's a lot of hard stuff on there that's tricky – that you've got to know" (p. 209). Another student anticipated that the test itself probably was more difficult than the class work leading up to it: "I got bad grades on [the class work] and it's too hard. If it's going to be hard just doing the class work, it's going to be real hard on the test" (p. 210). Another noted the importance of teachers' preparing students by providing practice tests: "(The teacher) plays around saying she doesn't want to see us again next year, that it's time for us to leave She cares about all the children She shows us by

teaching us more stuff and giving us examples of the test"(p. 214).

Based on their analyses, Roderick and Engel concluded creating incentives for low-achieving students through rewards and feedback may improve school effort (but not necessarily test scores). They also identified a group of students for whom incentives appeared to be ineffective. They cited two primary reasons for this: (a) first, some students, even those who valued promotion, felt the goal was unattainable; (b) second, students with low motivation were less likely to find support or encouragement in or out of school, facilitating the students' maintenance of their present trajectories of performance.

A cognitive interview study conducted by researchers at the National Center on Educational Outcomes aimed to use student input to inform the development of "comprehensible and readable" test items (Johnstone et al., 2007, p. 1). Students completed items in original or modified form (e.g., decreased word count, simplified vocabulary, bold font for key words) while being encouraged to think aloud and then participated in videotaped interviews about their experiences. The authors found students did not perceive any difference in item difficulty as a result of decreasing the number of words in stems and adding bold font to key words, although students preferred the bold font. They reported vocabulary of item stems and answer choices was perceived as important, particularly when non-construct-relevant words were included, as well as words with negative prefixes. It should be noted that the authors based their conclusions on a total participant sample consisting of eight students.

A recent study conducted by researchers from the five-state Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) project examined the effects of item modification with students in grade 8. Discovery Education Assessment provided grade 8 mathematics and reading items with extensive operational histories. Assessment experts, content-area specialists, educators, and researchers then collaborated to modify these items using a framework that was subsequently released as the Test Accessibility and Modification Inventory (TAMI;

Beddow, Kettler, & Elliott, 2008). The TAMI guides item writers to isolate the target construct of test items by reducing reading load of passages and stimuli, clarifying the question or directive in item stems, simplifying visuals, and managing the cognitive demands of item elements.

Following the item modification session, the CAAVES team conducted a cognitive interview study with a small sample of eighth-grade students ($n = 9$). The purpose of the cognitive interviews was to examine how students process the kinds of test items that were to be included on the CAAVES field tests. Specifically, the investigators asked students to think aloud as they completed a set of eight test items in either original or modified form (i.e., some students completed the original item while others completed the modified form). Student responses were audio- and video-recorded for transcription later. These transcripts were analyzed with the aim of understanding student perceptions about test items and provoking consideration about further changes that may be necessary to the items to enhance their accessibility, including reverting to their original forms in some areas if results of modifications were determined to be suboptimal. The cognitive interviews revealed several patterns in student views. It should be noted that statistical analyses were not feasible due to the small sample size.

As noted previously, data were intended to provide indicators of student views to facilitate item development as opposed to yielding empirical results that were generalizable to a large population. First, students required fewer research prompts (e.g., “keep talking,” “tell me what you’re thinking”) when engaging the modified items compared to the originals. Analyses of reading fluency indicated the students without disabilities (SWODs) group read more fluently than the students with disabilities who would be eligible (SWD-Es) group (158.3 words correct per minute compared to 86.3 words correct per minute). Further, students across groups spent less time on the modified items compared to the original items.

Most students in the SWD group indicated visuals were helpful and supported their

comprehension of reading passages and questions. Most students in the SWOD, by contrast, reported the visuals made no difference. Most students in both groups reported visuals and graphs were helpful in mathematics.

As per Rodriguez’s (2005) conclusion that three choices are optimal for multiple-choice items, each of the modified items contained three answer choices instead of the four choices found in the original item. Students who completed the modified items were asked about which version they preferred. With but one exception, students in the SWD group perceived no difference in the difficulty of the items due to the number of answer choices. By contrast, the majority of students in the SWOD group indicated three answer choices made the items easier. One student reflected on his preference, as follows: “If you didn’t get the answer right the first time, you . . . only had three choices to go back and look at . . . instead of four” (Roach et al., 2010, p. 10). Additionally, a number of the modified items used bold font to emphasize key terms. When asked to discuss their thoughts about this feature, the majority of students reported the bold font was helpful (though one student noted while the bold font helped him find the answer, it didn’t make the reading passage easier to comprehend).

Feldman, Kim, and Elliott (in press) investigated students’ perspectives on testing, specifically focusing on testing accommodations (which remains the most commonly used strategy to increase the accessibility of tests for students with special needs). Study participants were eighth graders: 24 students with disabilities and 24 without a disability. Although SWDs reported similar levels of pre-test anxiety, positive self-regard, and anxiety compared with SWODs, results indicated SWDs reported significantly lower test self-efficacy compared with their peers without disabilities. SWDs who received testing accommodations, however, showed decreases in post-test self-efficacy compared to students who did not receive testing accommodations; however, this result was not observed for SWODs.

In addition, Feldman and colleagues (in press) found significant positive correlations between pre-test self-efficacy and test performance

($r = 0.34$ for the accommodated condition and $r = 0.45$ for the non-accommodated test). A significant negative correlation also was observed between pre-test anxiety and test performance ($r = -0.33$), but only for students in the non-accommodated condition.

Incorporating Student Voices in Future Assessment Research

In our research on inclusive and accessible assessment strategies, we have found it essential to involve students with disabilities more directly and actively in our work. Given limited research on universally designed assessments, advances in cognitive load theory and mental load analyses (Clark et al., 2006), and ongoing concerns about improved accountability for students with disabilities, it seems logical and appropriate to invite students to be actively involved in the development and evaluation of more accessible tests. The involvement of students is not required by policy, but we believe that it is an essential component of assessment development, and that it will lead to more accessible items and tests. To that end, we proposed some “next steps” for research that include the perspectives of students in support of development of more accessible assessment systems.

There is a need for additional research that illuminates the effects of stress and test anxiety on the test performance of students with disabilities. Nicaise (1995) defined test anxiety as an individual’s physiological, cognitive, and behavioral responses that produce negative feelings in a test situation. An emerging body of research suggests that students with LD experience test anxiety at rates equal to or above their peers. By comparing students’ test performance to self-reported test anxiety, Bryan, Sonnefeld, & Grabowski (1983) found that (a) students with LD reported more test anxiety than their peers; and (b) the level of reported anxiety was a significant predictor of subsequent achievement test scores. Using the *Test Anxiety Inventory for Children and Adolescents (TAICA)*, Sena, Lowe, and Lee (2007) found that students with LD had higher scores than

their peers on two TAICA subscales (Cognitive Obstruction/Inattention and Worry) that are believed to interfere with test performance, while also reporting lower scores on a third subscale (Performance Enhancement/Facilitation Anxiety) that is related to improved achievement. These reports of increased levels of test anxiety on the part students with LD are troublesome because they represent both (a) a negative unintended consequence of large-scale testing (i.e., increased emotional distress) for students with LD; and (b) a potential source of construct-irrelevant variance that undermines score-based inferences regarding the achievement of students with LD (Saliva, Ysseldyke, & Bolt, 2007). Along with our efforts to create more accessible assessments, research is needed to identify effective interventions that will help students with LD manage text anxiety so they can fully demonstrate their skills during high-stakes testing. In addition, investigations of the anxiety-reducing effects of accessibility-enhancing test and item features would provide additional support for their utility and social validity.

Similarly, there is a need for additional research on the effects of student effort and motivation on test performance. Wise and Cotton (2009) suggested valid interpretation of test results requires “both a well-designed test and the student’s willingness to respond with effort to that test. Without adequate effort, test performance is likely to suffer, resulting in the test score underestimating the student’s actual level of performance” (p. 189). This is an area of concern for the implementation of inclusive and accessible assessments because lack of motivation and effort can undermine the effects of test features intended to facilitate student performance. In fact, the effect of decreased effort can be quite large; Wise and Demars’s (2005) review of the research suggested the test of performance of under-motivated students is over half a standard deviation lower than their peers. Research on students’ test effort and motivation generally has been conducted using self-report questionnaires that may be prone to response bias, but computerized testing provides opportunities to collect additional data (e.g., time spent per item, random guessing) that can corroborate these reports. Test item

modifications are often created under the assumption that they will facilitate students' efforts in responding to inaccessible items that would be de-motivating in their unmodified form; research is needed that provides support for this claim.

Conclusions

Accessible assessments are intended to facilitate the inclusion of students with disabilities in testing programs and support these students in demonstrating their knowledge. Therefore, it is essential to demonstrate the relationship of accessible assessments and students' perceptions of, and experiences with, testing. The *Standards for Educational and Psychological Tests* (AERA, APA, & NCME, 1999) support this idea, stating "certain educational testing programs have been advocated on the grounds that they would . . . clarify students' understanding of the kind or level of achievement they were expected to attain. To the extent that such claims enter into the justification for a testing program, they become part of the argument for test use and should be examined as part of the validation effort" (p. 23). Certainly, setting higher achievement standards for students and expecting improved efforts on the part of students (as well as educators and parents) to meet these standards is a key component of theory of action underlying standards-based accountability programs. Understanding whether (and how) this claim holds for students with disabilities and English language learners is important, but it has seldom been examined.

When students' perspectives regarding assessments and modifications are not considered, educators, policymakers, and test developers may work from a "paternalistic" assumption – that is, "acting upon (our) own idea of what's best for another person without consulting that other person" (Marchewaka, cited in Smart, 2001, p. 200). Although there are some cases in which ascertaining student perspectives and preferences would be difficult (e.g., students with significant cognitive disabilities with no reliable mode of communication), we believe most students with

and without disabilities are fully capable of expressing their opinions regarding the accessibility and acceptability of testing practices.

Many researchers and policymakers (including the authors included in this book) are engaged in efforts to improve the quality of assessments for students with disabilities and English language learners. Although the development and implementation of more accessible tests may not result in improved scores for every student, it appears test design practices that affect educational prospects and planning for many students can be improved. Given the omnipresence of high-stakes assessments in our nation's schools, it is important to give the students most likely to be affected by such assessments a voice in the development process. We encourage educators and test developers to include these data as part of their test development and validation efforts.

References

- Albus, D., Thurlow, M., & Bremer, C. (2009). *Achieving transparency in the public reporting of 2006–2007 assessment results* (Technical Report 53). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test accessibility and modification inventory (TAMI)*. Nashville, TN: Vanderbilt University; Boston: Houghton Mifflin.
- Bryan, J. H., Sonnefeld, L. J., & Grabowski, B. (1983). The relationship between fear of failure and learning disabilities. *Learning Disabilities Quarterly*, 6, 217–222.
- Council for Exceptional Children. (2008). *What every special educator must know: Ethics, standards, and guidelines* (6th ed.). Arlington, VA: Author.
- Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *The Journal of Technology, Learning, and Assessment*, 3(7). Retrieved November 7, 2010, from the <http://www.jtla.org> database
- Elliott, S. N. (1986). Children's ratings of the acceptability of classroom interventions for misbehavior: Findings

- and methodological considerations. *Journal of School Psychology, 24*, 23–35.
- Feldman, E., Kim, J., & Elliott, S. N. (in press). The effects of accommodations on adolescents' self-efficacy and test performance. *Journal of Special Education*.
- Fullan, M. (2001). *The new meaning of educational change*. New York: Teachers College Press.
- Gallagher, M. (2009). Data collection and analysis. In E. K. Tisdal, J. M. Davis & M. Gallagher (Eds.), *Researching with children and young people: Research design, methods, & analysis* (pp. 65–88). Thousand Oaks, CA: Sage.
- Johnstone, C., Liu, K., Altman, J., & Thurlow, M. (2007). *Student think aloud reflections on comprehensible and readable assessment items: Perspectives on what does and does not make an item readable* (Technical Report 48). Minneapolis, MN: National Center on Educational Outcomes.
- Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Kazdin, A. E. (1981). Acceptability of child treatment techniques: The influence of treatment efficacy and adverse side effects. *Behavior Therapy, 12*, 493–506.
- Moss, P. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*(1), 20–28.
- Moss, P. A., Pullin, D., Gee, J. P., & Haertel, E. H. (2005). The idea of testing: Psychometric and socio-cultural perspectives. *Measurement: Interdisciplinary Research and Perspectives, 3*(2), 63–83.
- National Association of School Psychologists. (2010). *Principles for professional ethics*. Bethesda, MD: Author.
- Nicase, M. (1995). Treating text anxiety: A review of three approaches. *Teacher Education and Practice, 11*, 65–81.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Research Council, Division of Behavioral and Social Sciences and Education, Committee on the Foundations of Assessment. Washington, DC: National Academy Press.
- Roach, A. T., Beddow, P. A., Kurz, A., Kettler, R. J., & Elliott, S. N. (2010). Incorporating student input in developing alternate assessments based on modified academic achievement standards. *Exceptional Children, 77*, 61–80.
- Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving students to high-stakes testing. *Educational Evaluation and Policy Analysis, 23*(3), 197.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.
- Saliva, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment in Special and Inclusive Education*. Boston: Houghton Mifflin.
- Sena, W. J. D., Lowe, P. A., & Lee, S. W. (2007). Significant predictors of test anxiety among students with and without learning disabilities. *Journal of Learning Disabilities, 40*, 360–376.
- Smart, J. (2001). *Disability, Society, and the Individual*. Gaithersburg, MD: Aspen.
- Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 1, 2009, from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Technical42.htm>
- Triplett, C. F., & Barksdale, M. A. (2005). Third through sixth graders' perceptions of high-stakes testing. *Journal of Literacy Research, 37*(2), 237–260.
- United States Department of Education. (2007, April). *Modified academic achievement standards: Non-regulatory guidance*. Washington, DC: Author.
- Wheelock, A., Haney, W., & Bebell, D. (2000). What can student drawings tell us about high-stakes testing in Massachusetts? *The Teachers College Record, ID Number: 10634*.

Computerized Tests Sensitive to Individual Needs

15

Michael Russell

The presence of computers in our schools creates valuable opportunities to enhance educational testing. The number of computers in schools has changed dramatically over the last three decades. In 1983, schools had, on average, only one computer for every 125 students. Today, that ratio has dropped to one computer for every 3.8 students and nearly every school across the nation has a high-speed Internet connection (Bausel, 2008). In addition, an increasing number of schools are adopting 1:1 computing programs that provide a laptop for every student.

A decade ago, Bennett (1999) predicted that computer-based testing would pass through three evolutionary stages before reaching its full potential. First, Bennett predicted that computers will be used to increase the efficiency of testing. Second, multimedia will be integrated into tests to increase the authenticity of items and tasks presented to students. Finally, computers will be used to deliver tests anywhere and at any time, so that testing becomes more integrated with instruction.

Today, many states are exploring or have begun to transition their testing programs to computer (Bennett et al., 2008). In most cases, however, testing programs that have embraced computer-based testing have done so solely to

increase efficiency. Their goals are simple – improve the efficiency with which tests are distributed, decrease the time required to score multiple-choice answers, and increase the speed with which results are reported. While achieving these goals saves time and money, they do not harvest the full benefits of computer-based testing.

This chapter focuses specifically on the benefits computer-based testing can bring to increasing accessibility, and thus test validity, by customizing the delivery of, interaction with, and response to test items based on the specific needs of each individual student. Given the capability of computers to customize delivery and presentation of tests, this chapter also explores the emergence of a new approach to framing accommodations as adaptations designed to meet four categories of access needs. The opportunity to develop an access needs profile for each student and to then tailor the delivery of tests based on this profile is also discussed. Finally, this chapter shares evidence that is emerging from one testing program that is capitalizing on these benefits.

Background

Over the past decade, several events have created a critical need for computer-based assessments that flexibly meet the accessibility and accommodation needs of individual students. The Individuals with Disabilities Education Improvement Act (IDEA) requires that students

M. Russell (✉)
Boston College, Chestnut Hill, MA, USA; Nimble
Innovation Lab, Measured Progress, Newton,
MA 02458, USA
e-mail: mike@nimbletools.com

with disabilities and special needs receive appropriate accommodations during instruction and that they receive similar accommodations during testing. Universal Design for Learning (UDL) emphasizes the importance of designing curricular materials in a way that increases access for students with a variety of needs. The importance of careful design and consideration for multiple ways of accessing information by students with disabilities are at the foundation of the National Instructional Materials Accessibility Standards (NIMAS, for more information see nimas.cast.org). Response to intervention (RTI) places strong emphasis on the collection and analysis of data to examine the extent to which specific interventions are having a positive effect on a student. The No Child Left Behind (NCLB) Act places strong emphasis on assessment as an accountability tool and requires students with disabilities and special needs participate in state assessment programs. Finally, in the field of testing, there has been a move toward computer-based test (CBT) delivery, in part, to increase the efficiency with which information is collected and returned to educators.

Together, NCLB and IDEA require testing programs to provide appropriate accommodations for students during testing. Appropriate accommodations vary with each individual student's disability, but include such things as Braille copies of reading materials for students who are blind, the reading aloud of written materials for students with dyslexia or other reading-related disabilities, the magnification of materials for students with visual impairments, the use of tools that isolate (or mask) information on a page for students with information processing disabilities, and the use of oversized writing materials or a keyboard for students with fine motor skill disabilities. In a testing situation, these types of accommodations are required so that students are able to comprehend what a test question is asking of them and so that they can better demonstrate their understanding of tested knowledge and concepts. Without access to appropriate accommodations, students are placed at a severe disadvantage in terms of demonstrating their achievement on the tests they

are required to take and, in many cases, must pass to move to the next grade or to graduate.

Focusing specifically on state assessment programs, there is ample evidence that these programs struggle to provide appropriate accessibility and accommodations for students with disabilities and special needs. For example, in February of 2002, a Federal District Court placed a temporary injunction that effectively halted the administration of California's High School Exit Exam to students with disabilities and special needs because there was compelling evidence that appropriate accommodations were not being provided to all students with disabilities and special needs across the state (*Chapman v. California Department of Education*).

Some states have written guidelines regarding the roles and responsibilities of people who assist in the administration of accommodations (e.g., readers, scribes, and sign language interpreters); however, there is great variability in both the breadth and depth of these guidelines (Clapper, Morse, Lazarus, Thompson, & Thurlow, 2005a). A study that included a focus group with blind and visually impaired adults highlights the ramifications of this (Landau, Russell, Gourgey, Erin, & Cowan, 2003). Study participants who previously had read aloud accommodations provided by a human reader pointed out several problems with this type of accommodation, which included: (a) the quality of the readers varied widely; (b) readers occasionally mispronounced or misread words; (c) readers provided intentional as well as unintentional hints to the correct answer; and (d) participants were sometimes reluctant to ask proctors to re-read parts of an item. In other words, while the students were provided with a read aloud accommodation, the accommodation itself was not delivered in a standardized or equitable manner, and likely did not provide students with an appropriate opportunity to demonstrate their achievement. The problem of inappropriate accommodations for students with sensory disabilities was also documented by Bowen and Ferrel (2003), who wrote, "few tests are valid for use with students with sensory disabilities, and the adaptations made by uninformed professionals can result in

both over- and under-estimates of an individual student's potential" (p. 10).

A 2003 Rhode Island Department of Education study also revealed difficulties schools face in providing accommodations. For example, schools had problems providing some of the most basic accommodations due to the large number of students requiring them combined with lack of space, equipment, and staff. Observations and teacher surveys revealed that some schools "bundle" accommodations for groups of students rather than follow individual IEP recommendations. In particular, the accommodations provided to the high school students were sometimes found to be "less than ideal" (Gibson, Haeberli, Glover, & Witter, 2003, p. 3). In addition, the study found significant differences between the daily accommodations that were provided to students during instruction and the accommodations that were available during assessments. Students that were provided with certain instructional accommodations such as one-on-one reading assistance or shorter assignments were not provided with comparable testing accommodations. Furthermore, accommodations frequently recommended for instruction, such as computers and other assistive devices, were rarely used during assessments.

In short, the struggle to provide appropriate and valid test accommodations results from three factors: (a) It is currently too expensive to provide different versions (Braille, magnified versions, etc.) of paper-based test materials that satisfy the accommodation needs of their students; (b) it is too expensive to provide individual test proctors who can read or sign a test; and (c) it is not possible for testing programs to provide accommodations in a standardized manner or to monitor the quality with which accommodations are provided by schools. While issues of standardization apply more to large-scale tests, these issues are relevant for both classroom-based and state assessment programs.

The fundamental problem encountered when attempting to meet federal accommodation requirements stems from the need to adapt standard assessment materials and administration procedures used for most students who do not

need accommodations to meet the unique needs of students who do need accommodations. In the fixed medium of paper, meeting the accommodation needs of students requires the development of multiple versions of test materials. Given the limited space and personnel available in schools, meeting accommodation needs of students who need to work individually with a test proctor is often not feasible.

The sections that follow explore the tremendous potential computer-based test delivery holds for resolving these issues.

Rethinking Test Accommodations

From a universal design perspective, instructional and test accommodations are intended to support each student's access to instructional or test content, interactions with content, and response to content. Accessing content requires information presented in a given form to be internalized by the student. Interactions with content require students to process, assimilate, manipulate, and/or interpret content that has been internalized. Responding requires students to produce an observable product that is the outcome of their interaction with content. During each of these three stages, the degree to which a variety of constructs operate within a student can interfere with the student's ability to access, interact, and respond in a manner that allows the student to demonstrate that construct.

Over the past 30 years, discussions about accommodations have generally focused on the specific method used to meet a need. While methods are important, the essential aspect of a given instructional or test accommodation is the specific need that must be met to support the development of a construct (i.e., learning) or the measurement of that construct (i.e., testing). For example, one commonly used instructional and test accommodation is "read aloud."

Read aloud is a method that presents an auditory representation of print-based content. Most often, the read aloud accommodation is associated with meeting the need of a student who has difficulty accessing print-based text,

either due to a visual impairment or a reading-related disability. The read aloud accommodation, however, can meet a variety of needs, including decoding text presented in a narrative form, visually perceiving text presented in a narrative form, visually perceiving information presented in graphical form, visually perceiving information presented in tabular form, processing information presented in tabular form, and pacing students as they access and process content.

Depending on which of these needs are being met, the way in which the read aloud accommodation is delivered may vary. For example, when the need that must be met focuses on decoding content presented in narrative form, a student only needs text presented in an auditory form, and does not need descriptions of graphics or tables presented in an auditory form. In contrast, a student with a visual impairment may require an auditory presentation of narrative text, graphics, and tables. While both students may be said to have used a read aloud accommodation, in reality the construct needs of each student that are being accommodated differ in important ways.

Categories of Accommodations

Traditionally, accommodations have been classified into five categories, each of which captures the type of change made to instructional or test content, or the conditions under which instruction is provided or a test instrument is administered. These five categories include changes in (a) presentation; (b) equipment and/or materials; (c) response methods; (d) schedule and timing; and (e) setting (Clapper et al., 2005a). When viewed from the perspectives of universal design and accessibility, accommodations can be reclassified into four categories: adapted presentation, adapted interactions, adapted response modes, and alternate representations.

Adapted presentation focuses on changes to the way in which instructional or test content is presented to a student. Examples of adapted presentation include changing the font

size used to present text-based content, altering the contrast of text and images, increasing white space, and reducing the amount of content presented on page. For paper-based tests, adapted presentation of test items often requires test developers or teachers to create special versions of these materials. As described in detail below, a computer-based test delivery system could build tools directly into the delivery interface that allow students to alter the presentation of content without requiring the development of additional versions of test materials.

Adapted interactions focus on changes to the way in which students engage with test content. Examples of adapted interactions include assisting students with pacing, masking content, and scaffolding. For paper-based tests, adapted interactions often require students to work directly with an adult and/or with additional materials, such as templates or masks. For a digital-content delivery system, adapted presentation and interaction tools can be built into the delivery interface and do not require any additional information, versions, or formats for instructional content or a test item.

Adapted response modes focus on the method a student uses to provide responses to instructional activities or assessment tasks. Examples of adapted response modes include producing text orally to a scribe or using speech-to-text software; pointing to answers or using a touch screen instead of circling, clicking, or bubbling; and using assistive communication devices to produce responses. For paper-based instructional activities and assessment tasks, adapted response modes may require students to interact with a scribe to produce a permanent record of their response. For digital instructional activities and assessment tasks, a computer-based test delivery system could allow students to use a variety of assistive technologies connected to the computer (touch screen, single switch devices, alternate keyboards such as Intellikeys, speech-to-text software, eye-tracking software, etc.) that enable students to produce responses. In some cases, however, additional methods that allow an examinee to respond to text content may also be required.

The final aspect of accessibility focuses on alternate representations. As Mislevy and his colleagues (2010) explain, alternate representations change the form in which instructional or test content is presented to a student. Unlike adapted presentations, which manipulate the way in which the same content is presented to an examinee, alternate representations present students with different versions of the test content. Reading content aloud; presenting text-based content in sign language, Braille; tactile representations of graphical images; symbolic representations of text-based information; narrative representations of chemical compounds (e.g., “sodium chloride” instead of “NaCl”) or mathematical formulas; and translating to a different language are all forms of alternate representations. For paper-based instructional and test materials, alternate representations often require the development of different versions or forms of the materials, or the use of translators or interpreters who present altered representations to the student. As described below, a computer-based test delivery system could tailor which representational forms are presented to a student based on his or her individual need or preference.

Thinking about designing and implementing accommodations using this four-category framework is helpful. First, by focusing on the specific need, it is possible to develop more nuanced methods and guidelines for meeting that need. As discussed above, read aloud can be used to meet several different needs, but it can only do so if the information presented orally matches the student’s specific need. Reading alternate descriptions of graphics for a sighted student with a reading-related disability does not meet his or her need and may create distractions. Similarly, reading aloud text without providing oral descriptions of graphics for a visually impaired student also does not meet the student’s need. By focusing on the specific need, specific instructions or methods for meeting that need can be specified and implemented with high degrees of integrity via computer.

Second, focusing on the specific need first, and then on potential methods for meeting that need, simplifies the process of determining

whether the resulting accommodation may violate the measurement of a given construct. In some cases, adapted presentation, alternate representations, and/or adapted response methods may conflict with the valid measure of a construct. For example, the validity of inferences one makes, based on a set of science items that are designed to measure a student’s ability to read and interpret information presented in scientific notation, would decrease if students were able to engage with representational forms that provided narrative descriptions of the scientific notation. Similarly, providing tactile representations for items designed to measure spatial and visualization skills might decrease the validity of inferences based on the tactile representational forms of content contained in those items. However, when the adapted presentation or alternate representational form is independent of the construct being measured, there are a number of ways in which advances in technology may permit the delivery of test items using alternate formats, without sacrificing the validity of the results.

In this section, we explore potential uses of computer-based technology to meet a variety of needs, including audio, signed, alternate language, and Braille access to knowledge representations. In addition, we explore methods for altering the presentation of knowledge representations through magnification, alternate contrast, and masking. Before doing so, however, it is important to recognize two additional advantages that computer-based technology provides, namely standardizing the provision of accessible test instruments and monitoring the use of accessibility features.

As noted above, many accommodations provided for paper-based tests require interactions between the examinee and a test proctor. For example, a read aloud accommodation requires a proctor to read content to a student and a signing accommodation requires a proctor to translate content to a signed representation and to then present that signed representation to the examinee. Similarly, students who require assistance in recording responses in the correct location, maintaining focus, or monitoring their pace may have a proctor assist with these functions.

In all of these cases, the interaction between the examinee and a test proctor may result in undesired interpretations of content, unintentional or intentional cues or hints, or undesired assistance. Computer-based technology provides opportunities to eliminate these undesirable interactions and allows test developers to make careful and thoughtful decisions about the exact manner in which text is to be read or interpreted and presented in sign. Test developers can also provide multiple methods for recording responses, allowing the student to select the method that allows them to record responses in an accurate manner. Finally, tools can be built into a delivery system to provide structured assistance with focus or pacing without introducing unintended assistance. While the end product of a well-designed computer-based test delivery system may provide educators and students with greater flexibility in the presentation, interaction, and response to test content, it also allows the way in which each option is provided to be standardized across students who make use of a given option and for this standardization to be based on careful decisions made during the test development process.

A second challenge that results when accommodations are provided on paper is accurate reporting of the provision of accommodations. For all paper-based tests, the provision of accommodations is recorded by a test proctor or another adult in the school. The degree to which accommodations are reported accurately, however, varies widely. In addition, because the provisions of test accommodations for a paper-based test are reported at the test level, testing programs do not have any information about the actual use of an accommodation for each item. For example, a student may be provided a read aloud accommodation, but may only ask to have text read aloud for a subset of items on a test. A computer-based test delivery, however, allows a testing program to collect accurate information about the use of each access tool or feature for each individual item on a test. This level of detail holds potential to improve the accuracy of information about the provision of accommodations

while also allowing testing programs to conduct detailed analyses about the items for which accessibility features were used more or less frequently. This level of detail allows test developers to examine features of items that may cause confusion for some students or may interfere with the measurement of a given construct due to construct-irrelevant elements of the item. Collectively, the improved accuracy with which computer-based test delivery can provide an accommodation holds potential to improve test validity, while the increased accuracy and level of detail about the actual use of access tools hold potential to assist test developers in improving the quality of test items.

Audio Access to Knowledge Representations

For paper-based test administration, read aloud is one of the most frequently provided test accommodation (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001). In essence, the read aloud accommodation presents test content in an audio form. Depending upon the test administration guidelines, readers are instructed to present only text-based information and to do so by reading text verbatim. For a mathematics test, these guidelines mean that narrative text associated with direction, prompts, or answer options can be read, but any numbers, formulas, or equations cannot be read. Other programs allow readers to present text, numbers, and mathematical and scientific nomenclature in audio form. Still other programs allow alternate descriptions of charts, diagrams, and other visual content to be provided in audio form. In practice, the read aloud accommodation is sometimes provided individually to each student, while, in other cases, an audio representation is presented to a small group of students.

These different guidelines and practices translate into many different forms of the read aloud accommodation. Importantly, each read aloud form meets a different and distinct need. Interestingly these multiple forms may partially

explain the varied findings regarding the effect of read aloud accommodations on students' test performance (Sireci, Li, & Scarpati, 2003).

Needs Met by an Audio Presentation

In a test situation, it is important that each examinee understands the problem presented by an item, the information provided by the item with which s/he is expected to work, and, for multiple choice items, the responses from which s/he is expected to select an answer. Presenting this information in a text-based representation can present challenges for some students with disabilities and special needs.

For example, students with dyslexia or other disabilities related to processing text may experience difficulty decoding information presented in a text-based form. For these students, reading text aloud reduces challenges to accessing information that result from difficulties decoding text. Some students with dyscalculia may struggle processing numbers and mathematical expressions presented in print-based form. For these students, audio representations that accompany the print-based form can help them to accurately internalize information presented in numerical form or using mathematical nomenclature.

Other students with vision-related needs have difficulty perceiving information presented in print-based form. Depending on the level of the vision need, these students may have difficulty viewing information presented in narrative, tabular, or graphic form. To enable these students to access item content, all text-based information may need to be presented in audio form and descriptions of graphics and tables may also be needed. Other students who experience difficulty processing information presented in tabular or graphical form may also benefit from audio descriptions of tables, charts, and diagrams.

Still other students who are developing their English language skills may have difficulty recognizing words presented in text-based form. These students may also benefit from having text-based information read aloud. Finally, students

who experience difficulty pacing themselves as they perform a test may also benefit from having a proctor pace them through the test by reading content aloud. While these students may not have difficulty accessing the item content, reading content aloud may assist them in maintaining a pace that prevents them from becoming distracted or from working too rapidly through test content.

Clearly, some of these needs overlap. For example, students with reading-related or vision-related needs, both benefit from having text-based information presented in audio form. However, there are also important differences among these needs that prevent a single method of providing read aloud support from meeting each of these specific needs. For example, providing descriptions of charts or diagrams for students with a reading-related need may be distracting if they are able to access and process information presented in graphical form. However, these students may benefit from having text that appears in charts or diagrams read to them without a more general description of the image itself. Conversely, providing only audio representations of text contained in an image may be insufficient for enabling a student with a vision-related need to access the concept(s), relationships, or information presented in that image.

Similarly, presenting text contained in each cell of a table in audio form may be sufficient for a student with a reading-related disability or who is developing fluency in English. But, for a student with information processing needs, who has difficulty seeing relationships among information presented in tabular form, more detailed descriptions of the table and its content may be required. For example, consider the table presented in Fig. 15.1. A student with a reading-related disability may benefit from having the column labels and the student names read aloud verbatim ("Student," "Friday," "Bill," etc.), but may not need numbers read aloud. Conversely, a student with dyscalculia may benefit from having the numbers contained in the table read aloud ("one hundred three," "one hundred fifty-seven," "ninety-eight," etc.), but may not need the column labels or student names presented in audio

Fig. 15.1 Table Associated with an Item

How many more apples did Mary pick on Saturday than on Sunday?
Number of Apples Picked

Student	Friday	Saturday	Sunday
Bill	103	157	98
Mary	118	198	134
Steve	87	134	113
Jane	178	243	54

form. Meanwhile, a student with a vision-related or information processing–related need may benefit from a fuller description of the information presented in the table. For example, rather than simply reading the contents of each cell (“Friday,” “Bill,” “one hundred three,” etc.), the relationships represented through the table may be presented (“Bill Picked one hundred three apples on Friday,” “Bill picked one hundred fifty-seven apples on Saturday,” etc.). Presenting information in this manner may enable these students to both access the individual words and numbers, as well as the relationships among each of those elements of the item’s content. Finally, for a student with a low vision need, additional information about the table that orients the student to the table’s purpose and design may need to be presented in audio form before presenting the actual information and relationships contained in the table. For example, the student might be presented with the following audio overview of the table: “This table is titled ‘number of apples picked.’ The table contains four columns and four rows of data. The first column is labeled Student and contains the names of four students who include Bill, Mary, Steve, and Jane. The three remaining columns are labeled Friday, Saturday, and Sunday. Each of these columns contains information about the number of apples picked by each student on each day.” Once oriented to the table’s design and general contents, these students may then be presented with audio descriptions that present the relationships among information presented in each cell of the table.

The audio representations used to meet each of these needs differ noticeably with respect to

the content that is presented to each student. As noted above, presenting a student with an audio representation that is not aligned with the students’ need may result in either distracting the student with unnecessary information or providing students with inadequate access to the content with which the student is expected to work. For this reason, it is important to align the information presented in audio form with the specific need for a given student. As will be discussed in greater detail next, it is also important for the test developer to carefully consider the construct being measured by an item and consider whether the audio representation provided for an item interferes with the measurement of that construct. For example, if an item containing a table is intended to measure a student’s ability to read and interpret information presented in tabular form, providing an audio representation that describes the relationships among information presented in the table will interfere with the measurement of the intended construct. However, if the item is measuring a student’s ability to perform another mathematical function, such as calculating a mean, using the information presented in the table, an audio representation will likely not interfere with the measurement of the intended construct.

In a computer-based environment, matching audio-based representations of item content with individual student needs requires two important steps. First, the full range of audio representations for each element of an item’s content must be developed for the item. For example, for the item depicted in Fig. 15.1, audio-based representations must be generated for each narrative and

numerical element. In addition, an audio description that provides an overview of the table associated with the item must be generated. Finally, audio representations that describe the relationships among information presented in the table cells must be generated.

Second, once these multiple audio representations are established, the test delivery system must be able to tailor the presentation of the audio representations to the specific need of each student. For example, for a student with a reading-related disability, the system may limit the audio presentations to only words presented in text-based form. For a student with dyscalculia, the system may limit audio presentations to only the numbers contained in the table. For a student with information processing needs, the system may limit the audio presentations to descriptions that capture the relationships among information contained in the table cells. Finally, for a student with vision-related needs, all forms of audio representations may be presented.

To accomplish this tailored audio content delivery, a test delivery system may require users to establish an access needs profile that specifies the specific needs that must be met while the student interacts with item content. The profile would then drive the representational form that is available for each student as each item is presented to the student.

Signed Access to Knowledge Representation

Students who communicate in sign language may read below grade level. For items that measure constructs unrelated to reading and contain information presented in narrative form, reading skills may interfere with the student's ability to access item content. As described above, the read aloud accommodation is intended to reduce this challenge by presenting text-based information in audio form. For students who are deaf or hard of hearing, however, audio-based representations are insufficient for providing access to text-based information. For these students, signed representations are required.

Similar to read aloud, there are multiple needs to consider when creating signed representations of test content. Of primary importance is the signed representation itself. Depending on a student's prior experience, content might be presented in American Sign Language (ASL) or as Signed English. ASL and Signed English both employ hand gestures and symbols to represent words, phrases, and expressions. American Sign Language, however, is a unique form of communication that has its own set of rules for grammar and often employs word order that differs noticeably from spoken English. In contrast, Signed English employs the same grammatical rules as spoken English and effectively represents words through hand gestures. Although most people who communicate in ASL can also communicate in Signed English, people who communicate in Signed English often are not familiar with ASL.

In addition to the form of sign that is employed, it is important to consider whether a student also relies on "mouthing" and/or partial auditory input to access content. Mouthing is the movement of the mouth to either say or mimic the pronunciation of words as they are signed. For students who developed oral language skills prior to developing a hearing-related need, mouthing may assist in interpreting or processing information presented in sign. Similarly, partial auditory input occurs when a person simultaneously presents information in sign and speaks that information aloud. Again, for students who communicate in sign and have partial hearing, auditory input may assist in accessing and processing information presented in sign. Finally, while both ASL and Signed English employ a variety of hand gestures to represent knowledge, facial expressions and body movement are often employed to provide supplementary information such as emotion, dialogue, and time.

To best meet the access needs of students who communicate in sign and read below grade level, it is important to match the specific way in which a signed representation is provided to the method with which the student is accustomed to communicating. When a student works individually with an interpreter, this matching can be accomplished by assigning an interpreter that is accustomed to

communicating information in the same manner with which the student is familiar. Research, however, provides evidence that the use of human interpreters presents several challenges to the provision of a signed accommodation and for test validity.

For example, the use of assistants who vary in their signing ability and in their familiarity with the test content results in the inequitable provision of the signing accommodation to students across a testing program. Johnson, Kimball, and Brown (2001) found that many interpreters were unfamiliar with mathematics vocabulary, especially at the higher grades. In turn, the variability in proficiency of the signing assistants is likely to have a negative effect on test validity and the comparability of test scores for students who use an access assistant (Clapper, Morse, Thompson, & Thurlow, 2005b).

One strategy employed by a few states, such as South Carolina and Massachusetts, that overcomes the problems presented by the use of signing assistants is the provision of the signing accommodation in a videotaped format. For this form of the accommodation, a single signer is used to present the test content to all students. Prior to producing a videotaped recording of the signer, the testing program works with the signer and other experts familiar with the signing of the test content to assure that the material is presented accurately and clearly. A single recording is then made and distributed to all students requiring a signing accommodation. In turn, the provision of high-quality signing is standardized across the entire testing program.

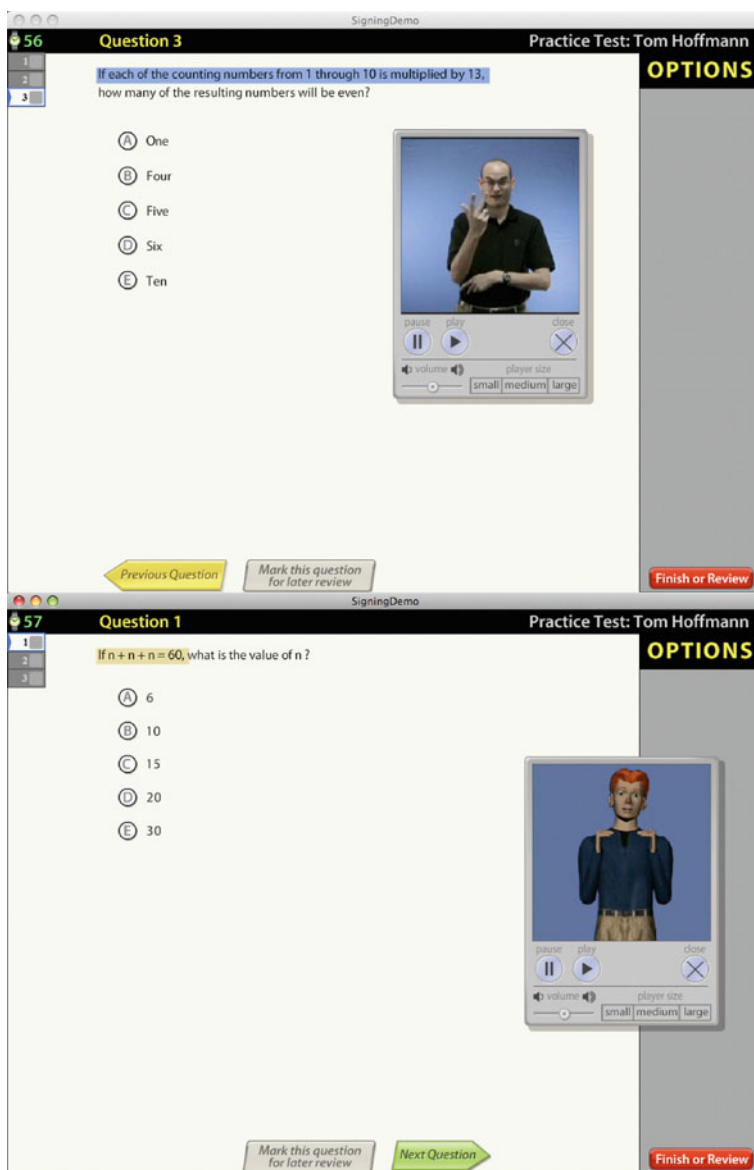
While standardization is highly desirable for a standardized testing program, the use of videotaped recordings are unsatisfactory for four reasons. First, it creates a physical separation between the test content and the presentation of that content. In most cases, the recorded accommodation is played on a television screen that sits in front of the classroom while the student works on a test booklet on his desk. This physical separation makes it difficult for students to move between the test item presented on paper and the signing of the text. Second, the use of a VHS or DVD machine to play the signed version requires

a teacher or the student to use a remote control device to reverse or fast-forward through the tape in order to replay specific portions of the tape. This inefficiency increases the time of testing and causes the focus to move away from the tested content onto the use of the controller. Third, in some cases, the video is played to a group of students simultaneously, requiring individual students to make a request in front of their peers to have a portion of video replayed and to then force all students in the room to re-watch that section. Fourth, when played for a group of students, the distance between the video and the student can make it difficult for students to clearly view the signed delivery.

Computer-based delivery of a test, however, holds promise to overcome problems associated with providing the signing accommodation either by a human or through a VHS or DVD machine displayed on a television screen. By embedding signed video directly in a computer-based test, the video and the test content can be presented in a consistent manner to all students and the signed representation can be thoroughly reviewed and approved prior to the day of testing. As shown in Fig. 15.2, embedding signed video in a computer-based test also allows the examinee to view a signed presentation of test content in very close proximity to each other and to the student (e.g., the video can be displayed within inches of the text-based representation of the item). Students can also be provided with individualized control over the size of the video displayed on their computer screen. Segments of video can be linked to blocks of text or portions of an item (e.g., each answer option) such that a student can click on the text and the associated video is played automatically, thus eliminating the need to use a controller to fast-forward or reverse through video. Finally, because the video is played on each individual computer, students may view portions of a video as many times as needed without affecting the test experience of other students.

Computer-based delivery of signed accommodations also provides an opportunity to employ signing avatars to present the signed accommodation. Avatars are human-like digital figures that can be programmed to move like a human. Using

Fig. 15.2 Signed presentation of test item by a recorded human and an avatar



avatars provides opportunities for students to customize specific aspects of the signed accommodation. Avatars base their movements on a script. A single script can be used to make multiple avatar characters move in exactly the same manner. This, then, allows students to select the character they like best from among a set of characters. Thus, a student could select a character that is male or female, short or tall, brown haired or blond haired, dark skinned or light skinned, and so on. Giving students the choice of character

may help increase their connection with the character and in turn their motivation during testing. Similarly, students could also control the background color to optimize the contrast between the avatar and the background so that they can best view signed presentation of the test content.

In addition, students can be given control over specific features of the signed presentation. For example, a student who reads lips while communicating via Signed English could activate mouthing, which allows the avatar's lips to move

in a manner that imitates the speaking of words. Similarly, a student with partial hearing who both views Signed English and obtains verbal cues could activate sound that works in conjunction with the signing. Finally, a student who is equally fluent in ASL and Signed English could switch between the two at any time during testing. Each of these features of avatars may help students tailor the provision of signing accommodations to meet their specific individualized needs.

Alternate Language Access to Knowledge Representation

As noted above, audio presentation can assist some students who are developing English language proficiency access text-based content. In some cases, however, the use of unfamiliar terms, phrases, or expressions are difficult for these students to access even when presented in audio form. In these cases, item content may need to be presented in an alternate language.

Like read aloud and signing, alternate language presentation of content may take different forms. One option is to allow students to access translations of individual words or phrases. On paper, this typically occurs by allowing a student to use an English-to-Heritage Language dictionary to look up words. One shortcoming with this strategy is that many words have multiple meanings and the burden is placed on the student to select the appropriate translation given the context in which the word is being used. On computer, key words in an item that may be unfamiliar to a student might be linked using hyperlinks to the appropriate translated word, given the context in which the word or phrase is used. This computer-based method may be advantageous for two reasons. First, it greatly reduces the amount of time a student would otherwise spend searching for a word using a Heritage language dictionary. Second, it assures that the translated word appropriately conveys the same meaning given the context in which the word is used.

A second strategy is to provide students access to a translated version of the entire item. On

paper, this can be cumbersome because multiple versions of a test must be created, distributed, and matched to a student. On computer, however, a student needs profile can be used to connect the student to the appropriate version. In addition, the student can be provided the option to switch between the original language and translated version of an item.

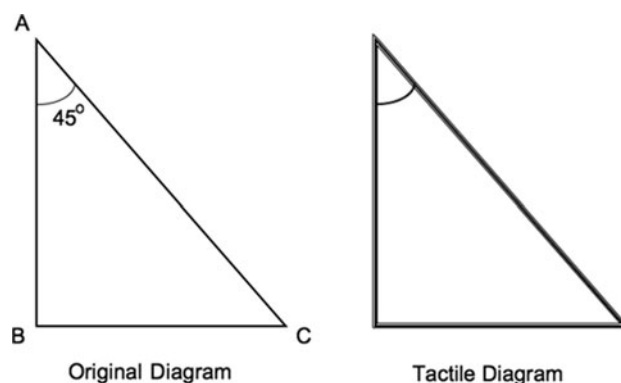
When presenting translated versions of items on computer, it is important to consider whether the student may also benefit by having other elements of the test delivery interface presented in a translated version. These elements may include navigation buttons, menu options, and directions.

Tactile Access to Knowledge Representation

Students who are blind or have low vision may benefit by accessing information in tactile form. For text-based information, content might be presented in Braille for students who are Braille readers. While some graphical content might be conveyed through audio descriptions, other content may be difficult to describe accurately. In such cases, tactile representations may provide greater access.

For paper-based tests, both Braille versions of text-based content and tactile versions of graphics are commonly provided. Computer-based technologies, however, also allow Braille representations of text-based content to be presented using refreshable Braille displays. Similarly, recent advances in peripheral devices also provide opportunities to present graphics in tactile form with supplemental information presented in audio form. Specifically, the Talking Tactile Tablet (TTT) allows students to place sheets containing tactile representations of images onto a touch-sensitive screen (Landau et al., 2003). While exploring the tactile representation, the student can press various features of the image and the computer can provide information about that feature in audio form. For example, Fig. 15.3 displays an item containing a graphic in its original form and in its tactile form. When using the tactile representation on the TTT, the student

Fig. 15.3 Original and tactile representation of image



may press on the top most angle of the figure and the computer could play an audio file that says, “Angle A, 45 degrees.” In this way, regardless of whether the student is a Braille reader, he could use his sense of touch to explore the image while also accessing all text-based information intended to provide additional information about the image.

Adapted Presentation of Item Content

The preceding sections focused on providing alternate representations of item content. This section focuses on altering the presentation of content to make it more accessible. When focusing on adapted presentation, it is important to recognize that the representational form of the content is not being changed. Instead, the presentation of a given representational form is being altered to meet a given need. In many cases, the needs met by altering presentation relate to vision or information processing needs. In this section, three types of adapted presentation will be explored: magnification, altered contrast, and masking. Similar to alternate representations, however, within each category of adapted presentation, there are multiple options that must be considered in order to meet each student’s specific need.

Magnification. Students with vision-related needs may benefit from working with larger versions of text and graphics. On paper, this is typically accomplished by providing a large-print copy of test materials that presents text in a

larger font and larger versions of images. An alternate approach is to allow students to view content using a magnifying glass. On computer, a test delivery system could provide students with several methods for magnifying content. One method is to allow students to adjust the size of the font used to present text-based content. While this method provides student flexibility in determining the font size, there are potential problems that occur when one relies on changing font sizes to increase access for low-vision students. First, changing font size often does not affect text that appears in charts, diagrams, and other images. As a result, students may experience difficulty accessing this text. Second, changing font size often alters the layout of text. This is particularly problematic when an item refers to text in a specific line of a passage or when text is presented in tables. Third, mathematical and scientific nomenclature is often presented using embedded images, the size of which are not affected by changes in font size. Similarly, graphical content is also not affected by font size changes.

Another method for increasing magnification is to provide students with a digital magnifying glass. Like its physical counterpart, a digital magnifying glass could be placed over any content displayed on the screen to render a larger image. A digital magnifying glass might also be resized to cover a larger portion of the screen and/or its magnification level could be adjusted. Unlike adjusting font size, a digital magnifier allows all content to be enlarged. However, for a student who requires all content to be magnified at all

times while taking a test, a digital magnifier may not be a good option because the magnifying glass would need to be repositioned repeatedly to view all content.

An alternative to a digital magnifying glass is to magnify the entire testing interface. For low-vision students, this is advantageous for several reasons. First, rather than manipulating a magnifying glass, they are able to focus on a single area of the screen and move content into that area for viewing. Second, magnifying the entire interface provides the same level of access to all elements of a test item as well as navigation buttons and menu options. Third, a significantly larger amount of magnification can be provided because the entire screen real estate is used to display content. One downside of enlarging the entire interface, however, is that content may be pushed off screen, so methods that allow students to easily move content on- and off-screen are required.

In each case, the method used and the level of magnification should be matched with the student's individual need. Like read aloud, this can be accomplished by establishing a user profile that defines the specific need for each student, and then tailors the tools and environment to those needs.

Alternate Contrast. Students with vision needs and reading-related needs may benefit from presenting content with alternate contrasts. For example, students with low vision may be able to better perceive content when contrast is increased. This can be accomplished by presenting text using a yellow font on a black background. As a second example, some students with Irlen's Syndrome experience difficulty perceiving text. While the cause of Irlen's Syndrome is unknown, individuals with the syndrome see words that are blurry or that appear to move. Placing a colored overlay on top of text helps stabilize and improve clarity of text.

There are three primary methods for altering contrast. First, the foreground and background colors of text and images can be altered. Yellow text on a black background is one combination commonly used to increase contrast. Several other color combinations, however, are effective for some students. Second, reverse

contrast changes the color pattern with which content is presented. For black text on a white background, reverse contrast presents white text on a black background. For other colors, reverse contrast changes the color to its counterpart of the opposite end of the color spectrum. For example, content presented with a yellow hue would be presented with a blue hue. Third, a color overlay could be placed over all content to change the tint with which content is displayed. For example, placing a red overlay over white text and blue-colored images would make all text appear with a pink hue and images with a purple hue. While altering contrast is effective for meeting many types of access needs, altering color or contrast must be done with care when items contain colored images and these colors are referenced in the item prompt or options. For example, an item that refers to a specific colored section in a pie chart, with each section colored differently, might become problematic if a colored overlay or reverse contrast alters the colors used in the pie chart. Thus, like magnification, the specific method and color combinations used to improve access for each student should be matched to the student's needs profile.

Masking. Some students with information processing needs may benefit by limiting the amount of information that is presented to them at a given time (Bahar & Hansell, 2000; Messick, 1978; Riding, 1997; Tinajaro and Paramo, 1998; Witkin Moore, Goodenough, & Cox, 1977; Richardson & Turner, 2000). One technique for limiting the amount of presented information is masking. On paper, masking involves placing an object over content that is not the current focus of the student. In some cases, sticky notes, paper, or a hand is used to mask content. Masking templates are also available and allow students to create a frame through which content is viewed.

On a computer, digital tools can be used to mask content and the delivery of items can be designed to present blocks of information in a predefined manner. For example, some students benefit by reading a prompt and working on a problem before they view answer options for a multiple-choice item. A computer-based test delivery interface can support this need in a

couple of ways. First, an answer-making tool could block response options until the student opts to reveal them. In essence, when a student first views the item, the prompt is shown, but a solid box is placed over each answer option. At the student's choosing, each answer mask can be toggled on or off to reveal or obscure that option. Alternatively, a delivery interface could present the student with the item prompt and then require the student to enter an open-ended response before revealing the multiple-answer options. While this method is more prescriptive, it helps guide students through the process of forming their own answer before engaging with answer options.

An alternate method for masking content focuses on separating content based on its representational form. For example, some items contain a combination of text, mathematical or scientific nomenclature, and graphics. To assist students in focusing on and processing the content conveyed through each of these representational forms, masking can separate the presentation of each form. This can be accomplished by displaying narrative content and blocking other representational forms when the student first views an item. The student could then choose to reveal or hide each representational form of content at any time while working on the item.

A third method of masking is to provide tools that enable students to create custom masks. These tools can take the form of electronic sticky notes that the student can place anywhere on the screen and resize as desired. The notes can then be repositioned or toggled on or off as desired. A digital framing tool could also be used to create a custom mask through which content is viewed. In essence, a digital frame covers the entire screen. The student is then able to define the size and shape of the window through which content is viewed. The window can then be positioned anywhere on the screen by the student as s/he works on the item. When sized so that the window is one line high and the width of a passage, the digital frame can serve as a line-by-line reader. In some cases, a digital frame can be combined with magnifying to create a large-print single-line mask. As touch-sensitive screens become more

commonplace, there is also potential to capitalize on this technology to allow students to create custom masks dynamically by dragging their fingertip over sections of an item they would like to temporarily hide.

Like magnification and alternate contrast tools, the masking tools available to the student can be tailored based on the students' need profile. Unlike magnification and alternate contrast, however, the use of masking tools often requires training and multiple practice opportunities prior to use during an operational test.

Implications for Test Validity

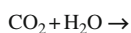
Test accommodations are intended to increase test validity by reducing the influence of non-tested constructs on the measure of the tested construct. As discussed throughout this chapter, a variety of access and interaction needs can interfere with the measure of a given construct. By tailoring the availability and use of access tools and alternate representations, the influence of these non-tested constructs may be reduced, thus improving the validity of resulting test scores.

When deciding which access and interaction tools and which alternate representations should be made available to students during testing, it is essential to consider carefully the construct that is intended to be measured. When the intended construct overlaps with the construct addressed by an access tool or representational form, then the provision of that tool or representational form will have a negative effect on test score validity. However, when the measured construct and the construct addressed by an access tool or representational form are independent, test score validity should either be unaffected or positively affected.

For example, the validity of inferences based on a set of items designed to measure whether a student can read, interpret, and use information presented in a table to solve a problem should not be negatively affected by the use of magnification, alternate contrast, or masking tools. Similarly, validity should not be negatively affected by reading aloud the text or numerical

Fig. 15.4 Sample item stems containing scientific nomenclature

Item 1



Item 2

Increasing amounts of CO₂ in the atmosphere is believed to contribute to global warming...

elements contained in tables. However, providing audio descriptions that express the relationships among elements in tables would decrease validity because such descriptions remove the need for students to recognize and understand relationships among data in the table.

Similarly, depending upon the construct being measured, providing an audio representation of scientific nomenclature may or may not reduce test validity. For example, consider the two items in Fig. 15.4. Both items include the expression CO₂, which employs scientific nomenclature to represent carbon dioxide. The first item is intended to measure whether students can interpret scientific nomenclature to complete a chemical equation. The second item is intended to measure students' knowledge and understanding of global warming. The ability to read and understand scientific nomenclature is an essential component of the construct measured by the first item, but is not a component of the construct measured by the second item. For this reason, providing an audio-based narrative representation that presents CO₂ as "carbon dioxide" would be problematic for the first item, but not the second.

To determine whether or not a given access tool or representational form has potential to violate the construct being measured, two issues are important. First, it is essential that the construct being measured is carefully and fully defined. As part of the definition process, it is important to describe what is and what is not intended to be measured. Second, it is essential that the construct being addressed by a given access tool or representational form is clearly specified. In doing so, it is useful to focus on the access need met by a given tool or representational form. Upon comparing the construct being measured and the construct addressed by an access tool or

representational form, one can determine whether the use of that tool or representational form will negatively impact test validity.

The Future Is Today

Over the past decade, several events have created conditions that allow us to reconceptualize test accommodations. The growing presence of computers in schools makes the delivery of tests on computer a realistic option for many schools and their teachers. Once committed to computer-based test delivery, digital technologies allow test developers, whether they be classroom teachers or large-scale testing programs, to capitalize on the flexibility of computers to tailor the test-taking experience to maximize each student's opportunity to demonstrate what they know and can do. To be clear, the goal of flexible, tailored test delivery is not to produce higher test scores or to increase the number of students who score above a certain point (e.g., proficient). Instead, the aim is to tailor conditions so that test scores provide a measure of student knowledge and skill that accurately reflects their current achievement. However, since many construct-irrelevant factors are believed to negatively affect the test performance of many students, more accurate measures for these students will often result in higher scores.

As described in detail earlier, computer-based test delivery provides an opportunity for us to shift away from thinking of test accommodations as changes made by test developers or test proctors for a small percentage of students with disabilities, and instead to think about altering the presentation, interaction, response modes, and representations used to engage all students in

tasks designed to reveal what each student knows and can do. In other words, rather than thinking of test accommodations as changes required for a small percentage of the test-taking population, computer-based test delivery coupled with principles of universal design permit us to flexibly tailor the test-taking experience for all students.

Providing the level of flexibility and individual tailoring described above in a paper-based world would be unaffordable and create an administrative nightmare. By investing in careful planning and development, however, flexibly tailored test-taking experiences on computer is achievable. For example, through funding from the U.S. Department of Education, a consortium of states led by New Hampshire Department of Education are currently implementing tailored computer-based test delivery for increasing numbers of students (<http://nimbletools.com/necap/index.htm>). This work started in 2006, when New Hampshire began working with software developers to develop a universally designed test delivery system. The first version of the test delivery software employed to deliver a mathematics test to high school students provided only three types of support – access to narrative content in audio form, magnified presentation of content, and navigation using assistive communication devices. To assure that the audio representations employed for the audio form did not violate the construct measured by a given item, considerable effort was invested in developing detailed scripts that prescribed the exact way in which text-based content was read aloud. As part of this process, scripts were reviewed by experts in mathematics and by item writers, and all audio versions of items were deemed appropriate given the constructs measured. Students who might benefit from these tailored supports were given the option of taking the state test on computer. In response, the state experienced nearly a tripling of students who opted to exercise their right to test accommodations. In addition, students who used the tailored delivery system for a given accommodation performed significantly higher than students who opted for a paper-based version of an accommodation (Russell, Higgins, & Hoffmann, 2009).

This work has since been expanded to include several additional methods for tailoring the presentation of and interaction with, representational forms of test content. In the spring of 2009, New Hampshire, Vermont, and Rhode Island gave students the option of taking their eleventh-grade science test on paper or on computer using a universally designed test delivery interface. The computer-based interface provided a variety of access and interaction supports, including audio access to narrative content, audio access to graphical content, audio access to all interface elements, navigation and response using any assistive communication peripheral device, magnification, multiple methods of altered contrast (including color overlays, reverse contrast, alternate font and background colors), multiple masking tools, auditory calming, and scheduled breaks. In the 92 schools that opted to deliver the test on computer to their students, the states again saw nearly a tripling in the number of students who opted to employ one or more tailored test delivery tool. In addition, regression analyses indicate that students who used one or more of the tailored delivery tools performed better than students who opted to receive a matched accommodation on paper.

While these findings are from just two early implementations of tailored computer-based test delivery, they provide three lines of preliminary evidence regarding the benefits of tailored delivery. First, these two implementations demonstrate that it is feasible and practical to develop tailored test delivery systems and to use those systems to administer large-scale tests. This is important because it demonstrates that the ideas presented above can be, and have been, translated into practice, and can be widely adopted in schools today. Second, both studies provide evidence that many students who opt not to employ one or more accommodation for a paper-based test will opt to have their needs met using a tailored computer-based test delivery system. In research performed to date, decisions about the use of access tools were not based on IEP status. Instead, teachers exposed students to the access tools through a tutorial and practice tests. In some cases, teachers then gave all students the option of using

tools if they felt the tools would be useful for them. In other cases, teachers limited the option to those students who they believed were likely to benefit from the use of a given tool. In both cases, however, the research conducted to date provides evidence that many students may not be taking advantage of opportunities to reduce the influence of construct-irrelevant factors through accommodations on paper but are willing to do so through tailored test delivery on computer. Third, these implementations provide preliminary evidence that tailored test delivery has a positive effect on student performance and does allow many students to better demonstrate their knowledge and skills.

Going forward, research is needed to better understand the decisions students and teachers make about the accessibility tools they opt to use during testing. As this understanding evolves, it is also important to develop resources that will help teachers and students make more informed decisions about tool use and assist in building a student's access needs profile. Finally, teachers and students will need support in updating and modifying a profile as a student's skills, knowledge, and facility with specific representational forms and access tools evolve. Once these tools and procedures are in place, it seems logical that increased benefits of tailored test experiences will result.

Clearly, more research is needed before definite claims can be made about the benefits of tailored computer-based test delivery. Nonetheless, the potential to improve test score validity by delivering tests using computer-based systems that are sensitive to individual needs shows great promise. As tailored test delivery is adopted more broadly, it will become important for testing programs to carefully define the constructs a test is intended to measure to inform decisions about which tailored delivery options will enhance test score validity and which may decrease validity. In addition, to fully capitalize on tailored test delivery, item writers will need to carefully consider and develop alternate representations that meet the needs of a diverse body of test takers. By fully defining constructs and developing appropriate alternate representational forms of item content,

testing programs will be better positioned to capitalize on tailored test delivery to provide valid measures of test performance for all students.

References

- Bahar, M., & Hansell, M. H. (2000). The relationship between some psychological factors and their effect on the performance of grid questions and word association tests. *Educational Psychology, 20*(3), 349–364.
- Bausell, C. V. (2008). Tracking U.S. trends. *Education Week*, March 27, 2008.
- Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practices, 18*(3), 5–12.
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9). Retrieved January 21, 2010, from <http://www.jtla.org>
- Bielinski, J., Thurlow, M., Ysseldyke, J., & Fieidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items*. Minneapolis, MN: National Center on Educational Outcomes.
- Bolt, S. E., & Thurlow, M. (2004). Five of the most frequently allowed testing accommodations in state policy. *Remedial and Special Education, 25*(3), 141–152.
- Bowen, S., & Ferrell, K. (2003). Assessment in low-incidence disabilities: The day-to-day realities. *Rural Special Education Quarterly, 22*(4), 10–19.
- Clapper, A. T., Morse, A. B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. L. (2005a). *2003 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Clapper, A. T., Morse, A. B., Thompson, S. J., & Thurlow, M. L. (2005b). *Access assistants for state assessments: A study of state guidelines for scribes, readers, and sign language interpreters* (Synthesis Report 58). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Synthesis58.html>
- Gibson, D., Haeberli, F. B., Glover, T. A., & Witter, E. A. (2003). *The use of recommended and provided testing accommodations* (WCER Working Paper No. 2003–2008). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.
- Johnson, E., Kimball, K., & Brown, S. O. (2001). American sign language as an accommodation during standards-based assessments. *Assessment For Effective Intervention, 26*(2), 39–47.

- Landau, S., Russell, M., Gourney, K., Erin, J., & Cowan, J. (2003). Use of the talking tactile tablet in mathematics testing. *Journal of Visual Impairment and Blindness*, 97(2), 85–96.
- Messick, S. (1976). Personality consistencies in cognition and creativity. In S. Messick & Associates (Eds.), *Individuality in learning* (pp. 4–22). San Francisco: Josey-Bass.
- Mislevy, R. J., Behrens, J. T., Bennett, R. E., Demark, S. F., Frezzo, D. C., Levy, R., et al. (2010). On the roles of external knowledge representations in assessment design. *Journal of Technology, Learning, and Assessment*, 8(2). Retrieved January 21, 2010, from <http://www.jtla.org>
- Richardson, J., & Turner, T. (2000). Field dependence revisited I: Intelligence. *Educational Psychology*, 20, 255–270.
- Riding, R. (1997). On the nature of cognitive style. *Educational Psychology*, 17(1), 29–49.
- Russell, M., Higgins, J., & Hoffmann, T. (2009). Meeting the needs of all students: A universal design approach to computer-based testing. *Innovate*, 5(4).
- Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Research Report 485). Amherst, MA: Center for Educational Assessment.
- Tinajero, C., & Paramo, M. (1998). Field dependence-independence cognitive style and academic achievement: A review of research and theory. *European Journal of Psychology of Education*, 13, 227–251.
- Witkin, H., Moore, C., Goodenough, D., & Cox, P. (1977). Field-dependent and field-independent cognitive styles and their educational implications. *Review of Educational Research*, 47(1), 1–64.

The 6D Framework: A Validity Framework for Defining Proficient Performance and Setting Cut Scores for Accessible Tests

Karla L. Egan, M. Christina Schneider,
and Steve Ferrara

Standard setting is commonly considered one of the final steps in the test-development cycle, occurring just prior to the release of test results, and it is usually narrowly defined, referring only to the workshop where cut scores are determined. Instead, standard setting should encompass the entire test-development cycle. It is more than a workshop; rather, standard setting is a multistep process where a state makes policy decisions about the rigor for achievement expectations, explicates descriptions of levels of achievement, and gathers stakeholders to recommend cut scores that separate students into achievement levels.

The standard setting process should begin with states writing achievement-level descriptors (ALDs), which are a means for state education agencies to communicate their expectation for student performance to local education agencies and other stakeholders. In addition, states should use ALDs to guide writing items and setting cut scores. This practice enables the test to be designed so that it supports the test score interpretations (i.e., what students should know and be able to do in relation to the content standards) intended by the state. By writing ALDs early, a state explicates a policy statement and fleshes out the content-based competencies that describe academic achievement as well as its aspirations

for the state's schoolchildren. Item writers can use ALDs to guide item development so that the items on the test better align to what the state means by Proficient.

The purpose of this chapter is to introduce a framework that structures standard setting as a multistep process in which the development and refinement of ALDs are emphasized. This framework should aid practitioners, such as state education agencies, in defining Proficient performance (and, by extension, other achievement levels), and designing and implementing standard setting for the alternate assessments of modified academic achievement standards (AA-MAS) and alternate assessments of alternate academic achievement standards (AA-AAS). Although there is much overlap with standard setting for unmodified grade-level assessments, issues pertinent to standard setting for the AA-MAS and AA-AAS are highlighted throughout the chapter.¹ The framework is designed around key events in the multistage standard

¹ Accessibility issues on unmodified grade-level assessments are not addressed in this chapter. States that provide students appropriate accommodations on an unmodified grade-level assessment that are also implemented during classroom instruction have made the test accessible to students with disabilities. That is, the items and test measure the same construct for students with disabilities as for their nondisabled peers. For standard setting, this means that panelists should be able to discuss assessment constructs without discussing accessibility and that the same recommended cut scores are appropriate for both sets of students.

K.L. Egan (✉)
CTB/McGraw-Hill, Monterey, CA 93940, USA
e-mail: karla_egan@ctb.com

setting process. Validity evidence is collected in each phase of the framework so that practitioners are able to build an argument to support the uses of the test scores for the AA-MAS and AA-AAS. Named the 6D Framework, it consists of six phases: Define, Describe, Design, Deploy, Deliver, and Deconstruct.

ALDs and Standard Setting Terminology

Before delving into this framework, however, it is important to define commonly used standard setting terms that may be new to novice standard setters. First is the term *achievement level*. Even before the advent of No Child Left Behind (NCLB), achievement levels were commonplace in K–12 testing. Almost every state has achievement-level designations, such as Basic, Proficient, and Advanced. Achievement levels are simply the labels associated with each level of performance, and they provide a normative-type judgment of an examinee’s test performance (Haertel, 1999). If test performance is labeled Proficient instead of Basic, it implies that the student possessed more knowledge and skills than did her or his peers who attained the Basic status. Sometimes achievement levels are called *performance levels*.

The term *achievement standard* is sometimes used interchangeably with the term *cut score*. In this chapter, the two are not interchangeable. Achievement standards are levels and descriptions of performance (i.e., ALDs), and cut scores are the specific points on the test scale that separate students into achievement levels. Cut scores are the numeric operationalization of the ALDs and are established during a cut-score recommendation workshop.

Finally, the term *standard setting* is often used to refer to the workshop where educators recommend cut scores. In this chapter, standard setting refers to a multiphase process that is used by states to define ALDs and their accompanying cut scores. The term *cut-score recommendation workshop* refers to a workshop where educators recommend cut scores.

A Validity Framework for Standard Setting for All Standards-Based Assessments

Validity refers to the appropriateness of the uses or interpretations of test scores, as opposed to the test score itself. In the case of the standard setting process, the cut scores, the ALDs, their relationship, and the intended interpretations are validated. It also may be necessary to validate the cut scores and ALDs of the AA-MAS and AA-AAS in relationship to those of the unmodified grade-level assessments.

To validate the cut scores, ALDs, and their relationship, states must collect various types of evidence throughout standard setting process as well as input from various stakeholder groups. Practitioners can use the 6D Framework to provide a logical structure for planning and implementing their AA-MAS or AA-AAS standard setting, in addition to collecting validity evidence. After providing general descriptions for each of the phases of the 6D Framework, detailed descriptions are provided in subsequent sections in this chapter.

Define: States define the population taking the test, the content standards underlying the test, the relationship of the AA-MAS or AA-AAS to the unmodified grade-level assessment, and the intended uses of the ALDs. All such decisions affect how ALDs are written.

Describe: States develop ALDs that illustrate the policy expectations regarding what students in each achievement level should know and be able to do.

Design: States plan the workshops associated with the multiphase standard setting process so that educators, policy makers, and state education staff can provide input into the cut scores. States also write items and construct tests during this phase.

Deploy: States implement the multiphase standard setting workshops, and the decision-making agency in the state (e.g., state board of education) approves the final cut scores.

Deliver: States deliver score reports along with interpretation guides, including ALDs, to local education agencies, teachers, and examinees and their parents.

Deconstruct: States and their contractors collaborate in collecting the documented validity evidence from the prior steps and compiling it into the standard setting technical report. States explicate the evidence confirming (and disconfirming) the relationship between the ALDs and cut scores in this phase.

Oftentimes the phases of the 6D Framework will occur in the order delineated above; however, it is possible for some phases to occur out of order. For example, it is possible that states will complete the standard setting technical report before delivering score reports. It is also likely that states will conduct various validity studies long after completing the program and standard setting technical report. The remainder of this chapter examines the phases of the 6D Framework in more detail.

Define

Within this first phase of the 6D Framework, states must define the scope of the assessment, deciding who and what will be tested. They must also define the number, intended purposes, and rigor of the ALDs. In large part, the advice about standard setting in this chapter and in the literature is applicable to unmodified grade-level assessments as well as the AA-MAS and AA-AAS. The unique considerations associated with the AA-MAS and AA-AAS are the students themselves and the designs of the tests (Egan, Ferrara, Schneider, & Barton, 2009). This section discusses the student populations and the relationship between the AA-MAS or the AA-AAS and the unmodified grade-level assessment. In addition, it discusses the desired rigor of the ALDs, the means for increasing accessibility, the intended uses of the ALDs, and the involvement of stakeholders. This section ends with a discussion of validity evidence.

Identify the Examinee Population

With unmodified grade-level assessments, the target population is obvious: almost all students in the state. With the AA-MAS and AA-AAS, the decision of who will take these tests is less obvious. Students with mild and moderate disabilities who take an AA-MAS are believed to be able to learn most of what their non-disabled, grade-level peers learn, but with more difficulty, over longer periods of time, perhaps in less depth, and perhaps with less demonstrable mastery over content and skills. It is not clear, however, exactly who constitutes this group, and the definition varies from state to state. Examinees taking the AA-AAS have significant cognitive disabilities and test on general content standards that have been extended downward (e.g., Browder, Wakeman, & Jimenez, n.d.).

Identify the Relationship to Unmodified Grade-Level Assessments

To design a meaningful standard setting process, states must also identify the intended conceptual and psychometric relationship between the unmodified grade-level assessment and the AA-MAS or AA-AAS. The AA-AAS is likely to be a distinct assessment, unlinked to the unmodified grade-level assessment. In some cases, states may want a psychometric or conceptual relationship between the AA-MAS and AA-AAS.

The AA-MAS, on the other hand, is required to target the same general content standards as the unmodified grade-level assessment. To fulfill this requirement, most states are developing an AA-MAS derived directly from the unmodified grade-level assessment. Typically, the items in the unmodified grade-level assessment are modified to reduce the cognitive complexity of the skills that the items target through such techniques as scaffolding items, bolding and underlining, and providing graphic organizers to help examinees organize their thinking (Wothke, Cohen, Cohen, & Zhang, 2009). In addition, the AA-MAS often comprise fewer items and are targeted to measure

student skills at a lower ability level than the unmodified grade-level assessment (which is typically unable to provide much information about what these students know and can do).

Even though states are designing the AA-MAS and unmodified grade-level assessment to measure the same standards (and sometimes using some of the same items), the relationship between the two assessments is not obvious. States may intend that the AA-MAS (a) predict performance on the unmodified grade-level assessment, (b) link vertically to the unmodified grade-level assessment, or (c) be distinct from the unmodified grade-level assessment. If the intended relationship is predictive, then, for example, Proficient performance on the AA-MAS might indicate that the state expects these students to achieve scores at a Basic level on the unmodified grade-level assessment. Predictive relationships can be established empirically, through prediction studies, or conceptually, following vertical articulation procedures (e.g., Lewis & Haug, 2005). If the goal is to link the AA-MAS to the unmodified grade-level assessment vertically, then cut scores from the unmodified grade-level assessment can be located on the AA-MAS score scale. Vertical linking must be established through a successful vertical linking study (e.g., Young, 2006). If the state intends the relationship to be distinct, then consideration must be given to choosing achievement-level labels and writing ALDs that will allow this to happen.

The left side of Table 16.1 illustrates predictive and linked relationships between ALDs for the unmodified grade-level assessments and the AA-MAS. Aligning Proficient on the AA-MAS with Below Basic on the unmodified grade-level assessment is intentional in this illustration. Most unmodified grade-level assessments provide some useful psychometric information about student achievement near the bottom of the test scales, so examinees who achieve the Proficient level on the AA-MAS should be ready to demonstrate what they know and can do on the unmodified grade-level assessment. The right side of Table 16.1 illustrates a distinct relationship between the ALDs for the two assessments in which a state will not link performance on

Table 16.1 Predictive or linked relationship and distinct relationship between ALDs for the unmodified grade-level assessments and AA-MAS

Predictive or linked		Distinct	
<i>Unmodified grade-level assessment</i>	AA-MAS	<i>Unmodified grade-level assessment</i>	AA-MAS
Advanced		Advanced	
Proficient		Proficient	
Basic	≈ Advanced	Basic	
Below basic	≈ Proficient	Below basic	
		Basic	
		Below basic	Advanced Proficient Basic Below basic

the AA-MAS to performance on the unmodified grade-level assessment. In this case, the state offers no expectation of how student performance on the AA-MAS may translate into performance on the unmodified grade-level assessment. States must define these relationships between the assessments before they write ALDs.

Identify Desired Rigor for ALDs

Before writing ALDs, states should decide the rigor that they want to associate with the ALDs for the AA-MAS and AA-AAS through generic policy statements. These statements will provide guidance and direction when writing ALDs. Generic policy statements do not describe specifically what students should know and be able to do within each achievement level; rather, the generic policy statements set the tone for the type of ALDs the state would like developed during the ALD-writing workshop. The tone used in the generic policy statements may have a significant impact on how the ALDs are written. For example, if the state is looking for minimal competency standards, the generic policy statements might talk about students “mastering basic skills.” On the other hand, if the state is looking for challenging ALDs, the statements might talk about “mastering challenging content.”

Identify Means for Increasing Accessibility

Prior to developing the AA-MAS, states should identify the approaches that will increase accessibility. This will be difficult because research on item modifications is limited, with most modifications based on the rationale (as opposed to empirical research) that the modifications will increase accessibility (Elliott et al., 2010; Kettler et al., in press; Welch & Dunbar, 2009). Early findings are emerging. For example, recent research from a series of item design studies and a pilot test of item modifications showed that graphic organizers appear to reduce the impact of working memory deficits, while bolding key text appears to aid attention for reading items (Wothke et al., 2009).

Identify Intended Uses of ALDs

Test development, cut-score recommendation, and scale score interpretation are three interconnected processes for which ALDs provide guidance (Schneider, Egan, Siskind, Brailsford, & Jones, 2009). The three processes work together to build a unified assessment system. During the *Describe* phase of standard setting, the state creates ALDs to support these three intertwined uses.

Stakeholder Involvement

A wide variety of stakeholders (e.g., politicians, concerned citizens, local educators, and/or parents) may be brought together to provide a state education agency feedback about the types of assessment relationships and ALD uses that would benefit students and educators. For example, the policy makers who crafted NCLB decided that all children must be assessed, regardless of disability. Once politicians enacted the law, each state education agency was required to determine exactly how to meet these requirements within their state. In the case of the AA-MAS, the states need to determine who and what

to assess. State education agencies may undertake a decision-making process where stakeholders make recommendations to the state education agency through a series of workshops or committee meetings.

Validity Evidence

It is necessary to collect validity evidence within each phase of the 6D Framework; much of this evidence is focused on the fidelity of implementing the standard setting procedure itself (Cizek, 1996). Called procedural evidence of validity, it “focuses on the appropriateness of the procedures used and the quality of the implementation of these procedures” (Kane, 1993, p. 14). The value of this type of evidence cannot be understated. Although it alone cannot support the validity of the ALDs and cut scores resulting from the standard setting process, its absence will undermine arguments to support their validity.

This means that practitioners will find that they collect the same types of procedural evidence in different phases, including documentation on process planning and panelist recruitment and selection as well as process evaluations. Instead of repeating these types of evidence in each section, they will be explained for the Define phase and should be assumed for the remaining phases where workshops or meetings occur.

Process Planning

Whenever the state holds a committee meeting or workshop where key decisions regarding the standard setting process will be made, the state should create a detailed plan that explicates the steps to be taken during that committee meeting or workshop. During the Define phase, for example, a state may hold a committee meeting to decide the examinee population for the AA-MAS. The plan for this committee meeting should be practicable, meaning that it should be easy to implement and understandable to the

general public (Hambleton & Pitoniak, 2006). A well-articulated plan will show the logic and rationales for the process that is to be followed in the committee meeting or workshop.

Panelists and Committee Members

As part of process planning, states should explain how workshop panelists or committee members will be recruited, including specific information such as race, gender, or area of expertise that may be targeted. It is important that the state can show a legitimate attempt to recruit a panel or committee that is representative of the state's diversity. Following the meeting, the state should summarize demographic information about the panel or committee as well as information on their expertise.

Process Evaluations

These are especially important to all committee meetings and workshops implemented in the 6D Framework. Process evaluations are attitudinal surveys administered to workshop panelists to collect information on their views of the validity of the procedure as well as their levels of agreement with the final workshop product. Process evaluations should be collected throughout the committee meeting or workshop to gauge panelist attitudes and understanding. Positive evaluations will strengthen the validity arguments associated with a workshop. Negative evaluations can undermine an entire workshop; therefore, it is important to collect feedback often in order to address negative attitudes early.² Panelists may report that they understood the task, thought the process was fair, but then report that they are not able to defend the final workshop product. If such results are found for a minority of panelists, this should not undermine the workshop recommendations.

² See Cizek, Bunch, and Koons (2004) for an example of a workshop evaluation.

Describe

In the Describe phase of the 6D Framework, ALDs are developed. It seems obvious that states should make decisions about what to assess at the beginning of the test-development cycle. Less obvious is that states should also determine *how much* students are expected to know in regard to the content standards. This means that states should define the knowledge, skills, and abilities (KSAs) in the ALDs (i.e., decide what it means to be Proficient) prior to the beginning of the test-development cycle so that they serve as a foundation for item development. This section defines the types of ALDs, discusses defining Proficiency on the AA-MAS and AA-AAS, and provides practical advice for writing ALDs.

Types of ALDs

In the past, ALDs described the typical or borderline student (Hambleton, 2001). Such descriptors were usually framed in the context of the cut-score recommendation workshop, ignoring that ALDs have purposes beyond the cut-score recommendation workshop. Achievement-level descriptors can be used for intertwined purposes (test development, cut-score recommendations, and score reporting), and different types of ALDs align to each purpose. There are three distinct stages of ALD development that support the three uses of ALDs:

- First, states should develop *target ALDs* to specify their expectations for students at the threshold of each achievement level. The target ALDs define the state's policy and content-based expectations (i.e., what it means to be Proficient). They are the lower-bound descriptions of an achievement level and guide the cut-score recommendation workshop. These descriptors target the skills all Proficient students should have in common.
- Next, states should develop *range ALDs*, which are expansions of the target ALDs. Along with the target ALDs, states should develop range ALDs prior to item writing because they should guide the item-writing

process. The range ALDs reflect the KSAs expected of examinees in the Proficient range. Because these descriptors target a range of students, not all Proficient students should be able to demonstrate all KSAs in the descriptions.

- Last, states should produce *reporting ALDs* once they adopt final cut scores. These ALDs represent the reconciliation of the target ALDs with the final cut scores. The target ALDs reflect a state's *expectations* of student performance, while the reporting ALDs reflect *actual* student performance. Because the final cut scores reflect policy considerations, in addition to the content considerations found in the target ALDs, it is necessary to reconcile the target ALDs with the KSAs reflected by the final cut scores. The reporting ALDs define the appropriate, validated inferences stakeholders may make about examinee KSAs based upon the student's test score. In other words, the reporting ALDs support score interpretation.

The range ALDs are the expectations regarding what students *should* know across the range of Proficiency, while the target ALDs focus on the skills students *should* possess to just enter Proficiency. The target ALDs are a specific subset of range ALDs. The target and range ALDs address the KSAs that a state *expects* Proficient students to know and be able to do. These expectations are based on aspirations for student performance, rather than *actual* student performance (i.e., reporting ALDs). The development of the target and range ALDs is akin to the work being done in the area of evidence-centered design (Bejar, Braun, & Tannenbaum, 2007; Plake, Huff, & Reshetar, 2009) because of the recommendation that ALDs be used to define the intended inferences about students, and the items be developed specifically to elicit the KSAs explicated in the ALDs.

Once a state finalizes cut scores, it is necessary to review the target ALDs based on the KSAs students demonstrated on the test. It is important to remember that target ALDs reflect expectations, which sometimes do not align with the reality of student performance. The reporting ALDs should reflect KSAs demonstrated by students on the

assessment. The Deploy section in this chapter discusses the refinement of target ALDs into reporting ALDs.

Defining Proficiency

Before discussing a method for writing ALDs, it is necessary to comment further on the relationship among the AA-MAS, AA-AAS, and the grade-level assessment. For the AA-AAS, the test measures the extended content standards, which means that the AA-AAS is established separately from the grade-level assessment. For the AA-MAS, the relationship to the grade-level assessment is more complicated because the state must define a linked, predictive, or distinct relationship between the two assessment types. Table 16.2 shows an example of a linked relationship where the grade-level Proficient definition has been translated into an AA-MAS achievement level called Approaching Expectations. Both sets of achievement-level descriptors are based upon the grade-level standards and both define the intended inferences that can be made about the student KSAs. As may be seen, however, the ALDs may vary in relation to grade-level standards. The way in which Proficiency is defined on the AA-MAS will very much depend on the way the state education agency plans to link the AA-MAS to the grade-level assessment and how clearly the state defines how grade-level standards are to be tested in the item specifications.

ALD-Writing Workshop

A workshop where a diverse group of stakeholders come together to create ALDs is one of the various ways to write descriptors. The proposed ALD-writing workshop is a multi-day event where the ALD committee of stakeholders participates in multiple rounds of discussion. The workshop should begin with an orientation where representatives from the state agency describe the test's target population, the test-development cycle, and how the ALDs function within that cycle. Following this, the panelists are trained

Table 16.2 Bridging the unmodified grade-level assessments and AA-MAS target ALDs through a linked relationship

Unmodified grade-level assessment	AA-MAS
What <i>Proficient</i> students <i>should</i> be able to do	What <i>Approaching Expectations</i> students <i>should</i> be able to do
<ul style="list-style-type: none"> find the area of rectangles and irregular figures drawn solve multistep problems find the correct ordered pair for a point on the coordinate plane add money amounts under \$100 add or subtract fractions with like denominators add or subtract fractions with unlike denominators find the range of a set of whole numbers where no whole number is greater than 50 find the probability of simple events extend the patterns of geometric shapes compare, order, and simplify fractions simplify mathematical expressions by using the order of operations rules 	<ul style="list-style-type: none"> find the area of rectangles and irregular figures drawn on a grid solve multistep problems when steps are scaffolded find the correct ordered pair for a point on the coordinate plane when steps are scaffolded add benchmark money amounts under \$100 (20, 30, 40, etc.) add or subtract common fractions with like denominators when graphics of fractions are shown add or subtract common fractions with unlike denominators when graphics of fractions are shown find the range of a set of whole numbers where no whole number is greater than 10 find the probability of simple events when illustrated by graphics extend short patterns of geometric shapes compare, order, and simplify fractions when graphics of fractions are shown simplify mathematical expressions when order of operations is sequential

on the method for writing both target and range ALDs. The following section provides a description of the proposed method for writing ALDs.

The proposed ALD-writing workshop has four rounds. In the first round, the AA-MAS and AA-AAS committees should:

- study and discuss the generic policy statements
- study and discuss the ALDs for the unmodified grade-level assessments
- discuss and annotate the content standards

In the second round, the AA-MAS and AA-AAS committees should compile and edit the ALDs. In the third round, a panel exchanges ALDs with a panel from a different grade level and provides feedback. In the final round, panelists consider the revisions and finalize the ALDs.

A meta-committee should meet immediately after the final round. The meta-committee is a cross-grade panel that will examine, edit, and revise the ALDs for coherency and articulation across the grade levels. The results of the meta-committee are provided to the state education agency for further consideration.

Method for Writing Range and Target ALDs

One method for beginning the ALD-writing process is to request that panelists parse the state content standards (or extended content standards) into achievement levels. For example, the content standards are parsed into Basic, Proficient, or Advanced. For these achievement levels, panelists will annotate the content standards to indicate whether the skills are expected of the just Proficient (P-), average Proficient (P), or highly Proficient (P+) student. The content standards provide ALD writers a framework for categorizing KSAs into achievement levels and provide parameters for the content the state education agency has deemed important for students. Throughout the parsing and annotation process, panelists study and discuss the content standards as well as their expectations for student achievement.

Once parsed, the annotated content standards must be compiled into range and target ALDs. Figure 16.1 shows an example of how a content standard may be annotated and transformed into ALDs. These ALDs should be written so that the language of the ALDs is accessible

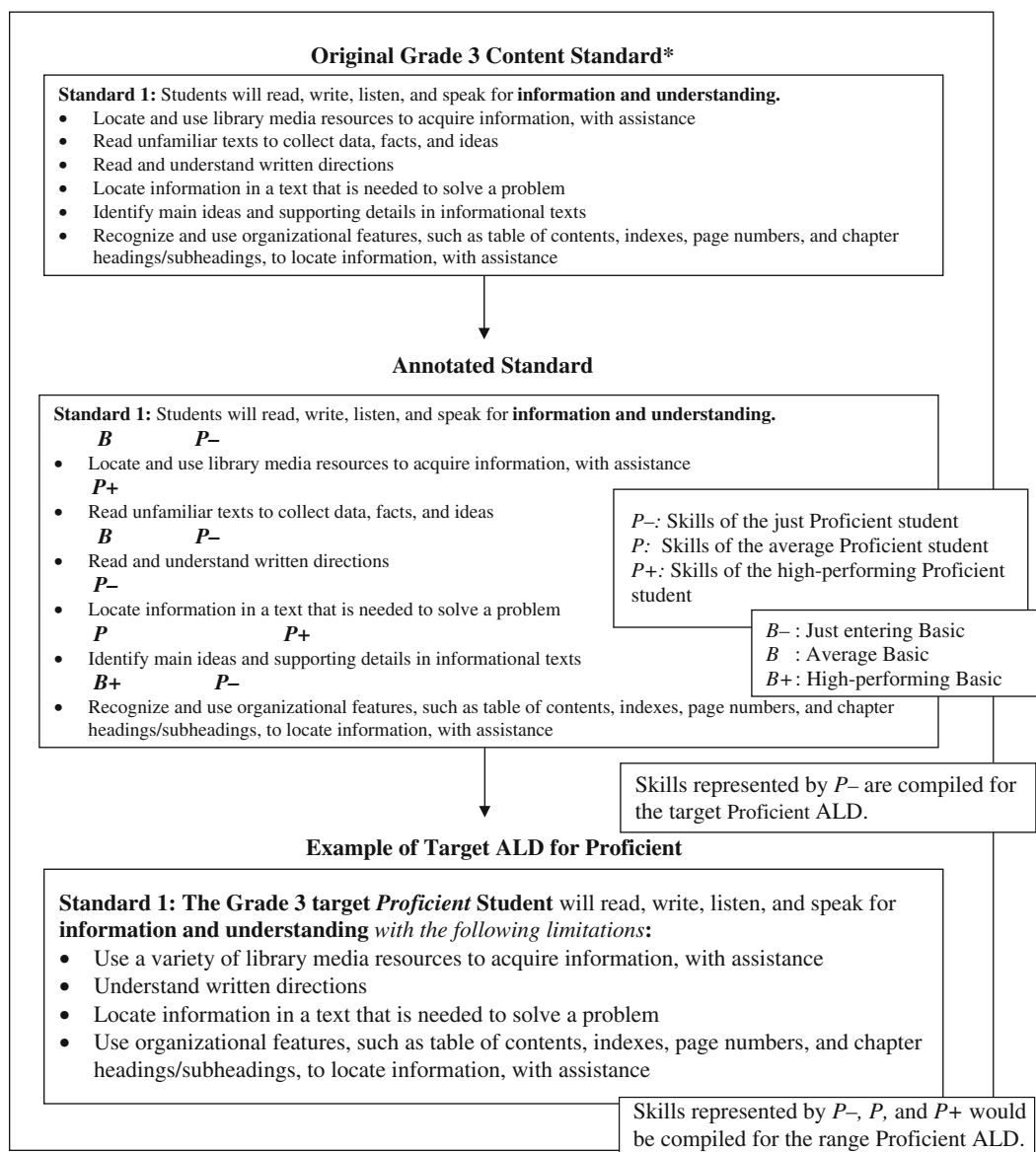


Fig. 16.1 Creating target and range ALDs

*Example content standard from New York States (2005) Grade 3 English Language Arts Core Curriculum (<http://www.emsc.nysed.gov/cia/ela/elacore.pdf>)

and clear. Mehrens and Lehmann (1991) provide guidance for developing measurable content objectives that can easily be adapted and extended to develop well-written ALDs.

ALDs should contain explicit, action verbs. When explicit verbs are used, teachers, parents, and stakeholders have a better sense of what students can actually do, whereas descriptors

that use implicit verbs require that stakeholders decode their meaning. For example, writing “The Level 2 student recognizes the main idea” does not lead to a clear conceptualization of what the student should do to demonstrate his or her skill. Rather, “The Level 2 student is able to tell or show the main idea” provides a better conception of actual student behavior used to measure

the skill. [See Mehrens and Lehmann (1991) for other examples for clarifying the language using implicit verbs.]

ALDs should provide contextual or scaffolding characteristics. Contextual or scaffolding characteristics of items can elicit intended examinee knowledge and skills. For example, students may be able to respond to mathematics questions that incorporate adding and subtracting positive and negative integers, only when number lines are present. This information can describe differences in examinee knowledge and skills that teachers can use to support instruction.

Validity Evidence

The range ALDs should guide item writing. If this is the case, then the state should show how the newly developed items align to the different achievement levels. Once the test is administered and cut scores are finalized, the state can re-examine how well the items aligned to their intended achievement level. This empirical evidence will show how well (or poorly) item writers were able to align items to the desired ALDs.

Design

During the Design phase of the 6D Framework, a standard setting plan is created that addresses the breadth of the process, meaning that the plan should encompass the cut-score recommendation workshop, the review by policy makers, and a method for refining the target ALDs (to the degree that is necessary). This phase acknowledges the importance of multiple stakeholder groups in the standard setting process, and it recognizes that several different types of workshops may be held before the cut scores are finalized.

The various groups often view the cut scores with competing perspectives. For example, teachers at a cut-score recommendation workshop may view the cut scores from a content-driven perspective, while administrators at a policy review may view the cut scores in terms of the number of students who will need remediation. For the same set of cut scores, state-level policy makers

may view the cuts from a reasonability perspective, considering how the performance data will be received by the public. Each group brings a unique view of the data, and these views need to be reconciled.

In the Design phase, the state will decide how the views of the different groups will inform one another. Typically, the teacher group recommends cut scores first, followed by the policy review, then the state review. If it is known in advance that state-level policy makers have a certain entrenched view (e.g., our state must have NAEP-like achievement standards), then it seems reasonable to share that view with the other groups.

To understand what methodologies can be used to establish teacher-recommended cut scores, it is necessary to consider the test design. In many ways, the test design will drive the methodology that will be used to set cut scores. In this section, possible AA-MAS and AA-AAS test designs are briefly discussed to set the stage for methodologies that can be used for teacher workshops. The section ends with a brief overview of a methodology for a policy review.

AA-MAS Test Designs

The most popular design for the AA-MAS is tests comprised of multiple-choice and constructed response items. Almost all planned and implemented AA-MAS are derived from corresponding unmodified grade-level assessments. In redesigning an unmodified grade-level assessment for use as the state AA-MAS, it is popular to reduce the total number of items as well as the overall difficulty of the test. States have only recently begun to address subtler modification issues, such as taking steps to reduce the cognitive complexity of items by bolding or underlining portions of the items.

AA-AAS Test Designs

Ferrara, Swaffield, and Mueller (2009) identify three designs for the AA-AAS currently used in K–12 assessments: assessment portfolios,

Table 16.3 Standard setting methods relevant for the AA-MAS and AA-AAS test designs

Methodology	Test design	Sample references
Bookmark	Multiple-choice/constructed response; Multiple choice only	Cizek and Bunch (2007, chapter 10); Mitzel, Lewis, Patz, and Green (2001)
Item-descriptor (ID) matching	Multiple-choice/constructed response; Multiple choice only	Cizek and Bunch (2007, chapter 11); Ferrara, Perie, and Johnson (2008)
Modified Angoff	Multiple-choice/constructed response; Multiple choice only	Cizek and Bunch (2007, chapter 6)
Body of work	Constructed response only; portfolio; performance task; rating scale	Cizek and Bunch (2007, chapter 9); Kingston, Kahl, Sweeney, and Bay (2001)
Profile sorting	Constructed response only; portfolio; performance task; rating scale	Jaeger (1995)
Reasoned judgment	Constructed response only; portfolio; performance task; rating scale	Roeber (2002)

performance tasks, and rating scales. Assessment portfolios contain collections of student academic work, video and audio recordings of students performing academic tasks, and other evidence of their performance in relation to the extended content standards.

Performance tasks generally focus on several related extended academic standards and contain anywhere from 3 to 15 items. Items often are scaffolded. Administration of assessment tasks usually is supported by manipulatives such as real objects (e.g., soil or rocks for science), response cards for nonverbal students, and other communication systems and assistive devices used by students during classroom activities.

Rating scales typically include considerable numbers of items linked to extended content standards. Teachers and other observers rate students as they complete tasks, collect student work samples, or interview other teachers and adults. Rating scales are typically multilevel (e.g., performance can be classified as nonexistent, emerging, progressing, or accomplished). The types of supporting evidence that must be provided typically are prescribed.

Methodologies for Recommending Cut Scores

The cut-score recommendation methods used for the unmodified grade-level assessments are also appropriate for the AA-MAS and AA-AAS. Item-based, cut-score recommendation

methods (e.g., Angoff, Bookmark, Item-descriptor matching) are appropriate for tests comprised of multiple-choice and constructed response items (including essay prompts) or rating scales. Methods requiring that panelists make proficiency judgments based upon student work samples (e.g., Body of Work, Profile Sorting) or scoring rubrics (e.g., Reasoned Judgment) are suitable for portfolio assessments that organize collections of student work as evidence of achievement. It is beyond the scope of this chapter to detail the cut-score recommendation methods, but there is an extensive, easily accessible literature on a variety of methods.³ Table 16.3 matches several cut-score recommendation methods with the test design for which they are frequently recommended. Table 16.3 also cites sample references that provide information on the cut-score recommendation methods, where the interested user can find more information on a particular methods.

Methodology for Policy Review of Cut Scores

Although policy reviews of cut scores are a frequent occurrence in the standard setting process for K–12 assessments, a standard method is not

³ Egan, Ferrara, Schneider, and Barton (2009) overview several cut-score setting methodologies for use with AA-MAS.

used. In some cases, policy reviews consist of a state superintendent along with one or two political operatives adjusting the cut scores to better suit policy needs. In other cases, a more formal process is followed that allows educators with an understanding of policy (e.g., district superintendents, school principals) to recommend changes to the teacher-recommended cut scores. In this process, a committee of 10–20 may participate as follows:

- Round 1: Review impact data (percentage of students in each achievement level) from the teacher workshop along with the target ALDs. Additional data may include impact data from other state, national, or international tests. Discuss the appropriateness of the impact data within and across grades and across content areas. Recommend changes to the cut scores.
- Round 2: Review the recommended changes to the cut scores. Re-open discussion on the appropriateness of the cut scores within and across grades and content areas. Recommend changes to the cut scores. Articulate a rationale for the changes.

The recommendations of the policy review may then be provided to the state's decision-making agency. Although policy considerations have the most weight during this phase of the process, it may be helpful to policy makers to see how the content in the ALDs may be modified by changes to the cut scores.

Validity Evidence

When the cut-score recommendation workshop is implemented, the state should plan to collect three types of evidence internal to the process itself: within-method consistency, intrapanelist consistency, and interpanelist consistency (Hambleton & Pitoniak, 2006). These types of evidence examines the consistency with which panelists are able to translate the target ALDs into cut scores (Kane, 2001).

Within-Method Consistency. This type of evidence examines what happens when the cut-score recommendation method is replicated across groups and is indicated by the standard error of

the cut score. If the recommendations are relatively consistent across groups, then the resulting standard error will be small. Similarly, discrepant recommendations may result in unacceptably large standard errors.

Intrapanelist Consistency. This type of evidence examines the consistency of judgments made by the panelist across rounds. It is anticipated that panelists will adjust their cut scores across rounds as they learn new information and gain better understanding of the target ALDs.

Interpanelist Consistency. This type of evidence examines the consistency of judgments across panelists. Most cut-score recommendation methods use consensus-building activities, where panelists are expected to engage in structured dialogues about their expectations of student performance. It is expected that the panelist cut scores will converge across rounds.

Deploy

Although the Define, Describe, and Design phases occur across months or years, the Deploy phase often occurs very rapidly. In the Deploy phase of the 6D Framework, panelists are recruited, meeting space is reserved, and the standard setting design is implemented. It is during this phase that the state implements the teacher and policy workshops, and the state decision-making agency enacts the final cut scores. An often-ignored step in this phase is an empirical validation of the target ALDs. It is crucial to the accurate interpretation of student performance that the target ALDs align well with the final cut scores. To ensure this, the state should revisit the target ALDs once final cut scores are determined.

Refining Target ALDs into Reporting ALDs

Target ALDs describe what students at the cut score *should* know. Ideally, panelists use the target ALDs to develop their cut-score recommendations so that the cut scores on the test scale

reflect the information articulated in the target ALDs. Over the course of a multiphase standard setting, the target ALDs may no longer reflect what students *are expected to know* based upon final cut scores; thus, when final cut scores are established, the target ALDs do not reflect the final cut scores. One reason for this is that policy considerations often influence a state's final cut scores. When a policy-based review follows a standard setting, policy experts are guided by different considerations, including the cross-grade logic of the impact data (e.g., Does the percentage of students at or above Proficient in each grade make sense?), available funding, political implications of cut scores (e.g., would they be considered “easy” compared to NAEP), and past performance in the state (Schneider et al., 2009). In short, the policy experts may consider a variety of reasons to adjust cut scores and, most often, they do not consider the target ALDs.

Once target ALDs no longer reflect the cut score, they do not provide a valid interpretation of the meaning of the achievement levels. The target ALDs can be refined into reporting ALDs by adjusting the content to reflect the final cut scores. This process may be accomplished by mapping the test items in order of difficulty, separating them by the cut scores, and summarizing the KSAs found in each achievement level. This may mean that some content is removed and some is added to the reporting ALDs (Bourque, 2009; Schneider, Egan, Kim, & Brandstrom, 2008; Schneider et al., 2009). As new test forms are introduced, the reporting ALDs can be refined further to reflect the new content.

Validity Evidence

In the Deploy phase, at least three workshops are implemented: the cut-score recommendation workshop, the policy review workshop, and the ALD refinement workshop. Given the role of ALDs for reporting, ALD refinement is an important validation step for a standard setting process. Traditionally, practitioners have collected evidence to show that the cut scores from the

cut-score recommendation workshop reflected the target ALDs. Now, it is necessary to turn this idea on its head and ensure that the reporting ALDs reflect the final cut scores. To do this, states should show how items align to each achievement level once cut scores are finalized. This can be accomplished by gathering stakeholder feedback to ascertain whether the KSAs of the items located around the cut scores are consistent with the descriptors for each achievement level.

Deliver

In this phase of 6D Framework, the results of the standard setting process are released to the public in the form of score reports to parents, teachers, and administrators. Often, shortened versions of the reporting ALDs will appear on the score report, and longer, more informative reporting ALDs will be on the state's website. When score reports from the AA-MAS and AA-AAS are delivered, it is important to communicate how those results should be interpreted in light of the grade-level assessment. If the term *Proficient* is used to label student performance on all three assessments, it is important to communicate the differences in what the term means for each assessment.

Using Reporting ALDs

Reporting ALDs should contain guidance so that teachers and parents understand how to interpret and use the ALDs. For example, states may explain whether the reporting ALDs are cumulative so that Proficient students will most likely possess the KSAs found in the achievement levels below Proficient. Additionally, teachers can use the reporting ALDs to better understand the general differences in student performance within a content area. Teachers should not use the reporting ALDs as shortcut teaching frameworks nor should they use them as a mini-curriculum because the reporting ALDs do not encompass the breadth of the content standards.

Validity Evidence

Validity evidence for the Deploy phase consists of the score reports and test results released to schools, and the interpretative guidance with suggestions for using the reporting ALDs. An interesting study would involve surveying educators to inquire how well the ALDs describe student performance, given how their students were classified on the test.

evaluated for the degree to which they support the conclusion that there is a good relationship between the ALDs and the cut scores. States and their contractors assemble this validity evidence into the technical report. Because technical reports tend to have a confirmationist bias, it may be beneficial to seek an outside perspective on the relationship between the range ALDs, the reporting ALDs, the cut scores, and the intended interpretations.

Deconstruct

In this phase of the 6D Framework, the various types of validity evidence collected throughout the standard setting process are compiled and

Technical Report

The technical report is an important piece of evidence for the procedural validity of the standard setting. Figure 16.2 suggests an outline for a

I.	Executive Summary
II.	Chapter 1: Planning
	a. Panelist recruitment
	b. Committee formation
	c. Panelist demographics and expertise
III.	Chapter 2: Execution
	a. Training Description
	i. Validity evidence: training materials; objective tests of panelist understanding of the cut score method; evaluations
	b. Round 1 Description
	i. Validity evidence: Round 1 votes; objective tests of panelist understanding; Round 1 evaluations
	c. Round 2 Description
	i. Validity evidence: Round 2 votes; Round 2 evaluations; intrapanelist consistency; interpanelist consistency; standard error of cut score
	d. Round 3 Description
	i. Validity evidence: Round 3 votes; Round 3 evaluations; intrapanelist consistency; interpanelist consistency; standard error of cut score
IV.	Chapter 3: Policy Review of Cut Scores
	a. Committee selection
	b. Process
	c. Final recommendations
V.	Chapter 4: Approval Process
VI.	Chapter 5: Refinement of Target ALDs into Reporting ALDs
	a. Process
	b. Final reporting ALDs
VII.	Chapter 6: Summary

Fig. 16.2 Outline for a cut-score recommendation technical report

technical report for the cut-score recommendation workshop. Technical reports should be created for each important panelist workshop held during the standard setting process. The technical reports should be developed as the authoritative source to understand the overall process undertaken to set cut scores and to review the technical evidence associated with the standard setting. The audience for technical reports includes members of technical advisory committees or peer review committees.

Within the narrative of the technical report, the author should directly address any anomalies that may have occurred during the course of the standard setting. For example, a grade-level panel may request an additional round of voting and discussion, and the motivation for this additional round should be discussed. The technical report should also address any seemingly aberrant results in the data. For example, a grade-level committee may indicate disagreement with the cut-score recommendation process in the evaluations. This disagreement may result from other sources (e.g., a state law that drives cut-score placement) as opposed to disagreement with the process itself.

Outside Perspective

As an additional source of validity evidence, a state education agency could hire a consultant to review the technical report and the relationship between the ALDs and the cut scores. The consultant could specifically look for disconfirming pieces of evidence that were undetected previously. The purpose of this review is not to undermine the results of the standard setting; rather, it is to build further support by finding the disconfirming evidence and addressing it.

Validity Evidence

The technical reports will serve as an important source of evidence for procedural validity for the standard setting process. Much of the validity evidence discussed throughout this chapter, such as the recruiting process and the panelist

demographics, will be summarized in a technical report.

Discussion

This chapter describes the 6D Framework for designing a standard setting process that encompasses all phases from writing ALDs to setting and finalizing cut scores. The purpose of this framework is to help practitioners collect validity evidence throughout a multistage standard setting process that supports the intended interpretation of test scores. Given that ALDs have such an important role in the assessment development and reporting process, they should be developed carefully. Although the measurement field has been calling for increased attention to the ALD-development process, change in current practice has been slow.

Implementing the 6D Framework

The 6D Framework should be implemented in conjunction with the test-development cycle. Table 16.4 provides an overview of each phase in the framework. It is important to understand that the Define phase and range ALDs from the Describe phase remain static once developed. They feed into the test-development cycle but are not part of the cycle itself. Thus, the content standards, general policy statements, target examinee population, and range ALDs should not change throughout the life of the testing program. The rest of the phases and the target ALDs are part of the test-development cycle. Information from each test administration can be used to modify the reporting ALDs so they better describe student performance in the achievement levels.

There are two reasons that range ALDs are not updated based on the final cut scores. First, the range ALDs guide item writing and test development. As such, they cannot be moving targets from year to year or the underlying test construct will also change, thereby jeopardizing the process used to equate test scores from year to year. Second, the range ALDs reflect the entire range of performance in the Proficient achievement level; thus, the range ALDs should be more

Table 16.4 Steps in the 6D framework for designing and developing the AA-AAS and AA-MAS standard setting process

6D phase	Standard setting tasks	Validity evidence
Define	<i>Define target examinee population;</i> <i>Describe relationship to unmodified grade-level assessments;</i> <i>Create generic policy statement</i>	Academic KSAs that students are expected to learn in each content area <ul style="list-style-type: none"> • For AA-MAS, these are the grade-level content standards • For AA-AAS, these are extended from the grade-level content standards Definition of relationship to the unmodified grade-level assessment Definition of the population of students who take the assessment Intended uses of ALDs Process plans, committees, evaluations
Describe	<i>Plan and hold ALD-writing workshop;</i> <i>Create target ALDs;</i> <i>Create range ALDs</i>	Description of ALD-writing workshop Overview of panelist recruitment and demographics from ALD-writing workshop Definition of target achievement for the students just entering the Proficient achievement level and other levels Definition of the range of KSAs expected of examinees at the Proficient achievement level and other levels Alignment of items to range ALDs
Design	<i>Design cut-score recommendation workshop</i>	Description of cut-score recommendation workshop Appropriate cut-score recommendation method for the test design
Deploy	<i>Recruit panelists;</i> <i>Hold cut-score recommendation workshops;</i> <i>Hold policy-review workshop;</i> <i>Obtain state approval of cut scores</i>	Description of panelist recruitment effort Panelist demographics Comparison of panelist demographics to state student demographics Panelist evaluations of workshops Within-method consistency, intrapanelist consistency, interpanelist consistency Final cut scores Alignment of reporting ALDs to final cut scores
Deliver	<i>Distribute score reports and interpretation guides</i>	Reporting ALDs delivered to public
Deconstruct	<i>Implement validity studies;</i> <i>Write technical reports</i>	Description of workshops from other 6D phases Detailed results from workshops (e.g., individual panelist recommendations) Outside review of validity evidence

robust to changes in the cut scores than are target and reporting ALDs, which focus on the very narrow range of KSAs that are right at the cut score.

What to Do When Test Development Has Already Occurred

Because standard setting has long been considered the end of the test-development cycle, many states may find themselves in the situation where test development has occurred prior to the development of ALDs. In this event, it is still important to develop the ALDs prior to the standard setting, using the content standards. Ideally, the ALDs should be developed in a separate workshop

from the cut-score recommendation workshop. This allows for community input and for the state agency to revise and approve the ALDs prior to the cut-score recommendation workshop. Because these ALDs were not used to guide the test-development process, it will be important to periodically review the ALDs to ensure that the test has not drifted from the content of the ALDs. It may also be necessary to update the ALDs so that they continue to accurately reflect the KSAs of the students in each achievement level.

Conclusions

The main purpose of this chapter was to introduce the 6D Framework, which demonstrates the centrality of ALD development and

refinement to both a multistage standard setting process and the test-development process. Three types of ALDs were introduced: range, target, and reporting. States use range ALDs to guide item writing and test development, target ALDs to guide the cut-score recommendation workshop, and reporting ALDs to help stakeholders interpret test scores.

The multistage standard setting process also considers that various members of the community have a stake in what the ALDs reflect and where the cut scores are set. Defining standard setting as a multistage process should increase the transparency of the process for all stakeholders involved.

Finally, it is hoped that the 6D Framework will aid practitioners in the design and implementation of the multistage standard setting process and will provide guidance on the types of validity evidence to collect during each stage of the process. The 6D Framework should aid practitioners in producing valid and defensible cut scores and ALDs.

References

- Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 31–63). Maple Grove, MN: JAM Press.
- Bourque, M. L. (2009). *A history of NAEP achievement levels: Issues, implementation, and impact 1989–2009*. Retrieved February 2, 2010, from <http://www.nagb.org/who-we-are/20-anniversary/bourque-achievement-levels-formatted.pdf>
- Browder, D., Wakeman, S., & Jimenez, B. (n.d.). *Creating access to the general curriculum with links to grade level content for students with significant cognitive disabilities*. Retrieved October 23, 2008, from <http://www.naacpartners.org/products/presentations/national/OSEPLEadership/9020.pdf>
- Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, 15, 13–21.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, 23, 31–50.
- Egan, K. L., Ferrara, S., Schneider, M. C., & Barton, K. E. (2009). Writing performance level descriptors and setting performance standards for assessments of modified achievement standards: The role of innovation and importance of following conventional practice. *Peabody Journal of Education*, 84, 552–577.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., et al. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children*, 76(4), 475–495.
- Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The item-descriptor (ID) matching procedure. *Journal of Applied Testing Technology*, 9, 1–20. Retrieved February 11, 2009, from http://www.testpublishers.org/Documents/JATT2008_Ferrara%20et%20al.%20IDM.pdf
- Ferrara, S., Swaffield, S., & Mueller, L. (2009). Conceptualizing and setting performance standards for alternate assessments. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 93–111). Baltimore: Paul Brookes Publishing.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18, 5–9.
- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Washington, DC: American Council on Education.
- Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15–40.
- Kane, M. (1993). *The validity of performance standards*. Unpublished manuscript developed for National Assessment Governing Board. Retrieved September 9, 2008, from ERIC Document Reproduction Services (ED365689).
- Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum Associates.
- Kettler, R. J., Rodriguez, M. R., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (in press). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*.
- Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*

- tives (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum Associates.
- Lewis, D., & Haug, C. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education, 18*, 11–34.
- Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. Fort Worth, TX; Harcourt Brace College Publisher.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates.
- Plake, B. S., Huff, K., & Reshetar, R. (2009). *Evidence-centered assessment design as a foundation for achievement level descriptor development and for standard setting*. Paper presented at the National Council of Measurement in Education, San Diego, CA.
- Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 20, 2010, from <http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html>
- Schneider, M. C., Egan, K. L., Kim, D., & Brandstrom, A. (2008, March). *Stability of achievement level descriptors across time and equating methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.
- Schneider, M. C., Egan, K. L., Siskind, T., Brailsford, A., & Jones, E. (2009, April). *Concurrence of target student descriptors and mapped item demands in achievement levels across time*. Paper presented at the National Council of Measurement in Education, San Diego, CA.
- Welch, C., & Dunbar, S. (2009). Developing items and assembling test forms for the alternate assessments based on modified achievement standards (AA-MAS) In M. Perie (Ed.), *Considerations for the alternate assessment based on modified achievement standards (AA-MAS): Understanding the eligible population and applying that knowledge to their instruction and assessment* (pp. 195–234). (A white paper commissioned by the New York Comprehensive Center in collaboration with the New York State Education Department.) Retrieved January 31, 2010, from <http://www.cehd.umn.edu/nceo/AAMAS/AAMASwhitePaper.pdf>
- Wothke, W., Cohen, D., Cohen, J., & Zhang, J. (2009). *2% AA-MAS working group Spring 2009 pilot study technical report*. Retrieved January 31, 2010, from <http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=229&ContentID=62021&Content=75362>
- Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469–486). Mahway, NJ; Lawrence Erlbaum Associates.

Part IV
Conclusions

Implementing Modified Achievement Tests: Questions, Challenges, Pretending, and Potential Negative Consequences

17

Christopher J. Lemons, Amanda Kloo,
and Naomi Zigmond

As we write, thousands of educators across the country are working with Individual Education Program (IEP) teams to determine the testing fate of their students enrolled in special education. The teams must decide which students will be required to participate in the challenging, rigorous, and lengthy regular accountability assessment (either with or without accommodations) and which students will be given the chance to demonstrate grade-level proficiency on a possibly less-challenging, shorter, alternate assessment based on modified academic achievement standards (AA-MAS). In the previous few years, hundreds of people, including staff from state departments of education, educators, test developers, parents, and advisors, have spent numerous hours debating the merits of a modified exam and working to develop an appropriate assessment and to select the correct group of students for whom this test would be most appropriate and beneficial.

The rationale for developing such a test is to improve accessibility for students with disabilities who may need specific accommodations and modifications to the regular assessment to allow them to demonstrate their knowledge of grade-level academic content. As Elliott, Beddow, Kurz, and Kettler ([Chapter 1](#), this volume) discuss, enhancing the accessibility of assessment systems is critical to establishing test validity and

ensuring equitable access to educational opportunities provided based upon assessment results. It is in this spirit, at least according to some, that the AA-MAS was initially proposed – it was argued that the accountability system at the time did not permit a small proportion of students with disabilities the opportunity to demonstrate their proficiency in grade-level content. Thus, the US Department of Education (USDE) called for the development of another test that would level the playing field for this group of students. It was intended that this test would improve the accountability system, enhance the instruction provided to these students, and reduce the burden to many school districts of students in special education who failed to meet the high adequate yearly progress (AYP) standards of No Child Left Behind (NCLB).

Unfortunately, it appears that creating a *different* test in the spirit of enhancing equal access may be leading us down a road of unintended consequences. Perhaps, just as in *Brown v. Board of Education*, separate tests are as inherently unequal as separate schools. The purpose of this chapter is to consider whether the potential benefits of creating a different test – the AA-MAS – outweigh the possible negative consequences. While we are highly supportive of both the value of enhancing access to the general education curriculum and assessment for students in special education and the work of colleagues aimed at achieving this goal, we feel that a further examination of the benefits and consequences of the

C.J. Lemons (✉)
University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: lemons@pitt.edu

AA-MAS may be helpful to state leaders and test developers as they move forward in its development. More specifically, we address questions that are currently being raised related to the AA-MAS: Will more students be able to demonstrate “proficiency” on a modified exam? Will fewer schools be penalized on AYP now that an additional 2% of students in special education may be counted as proficient? Will instruction change for students who are now going to be held accountable for something other than full grade-level academic standards? Will all of the time and money spent to create this modified test actually have been worth it?

A Brief History of NCLB, Accountability, and the AA-MAS

Since the beginning of the millennium, the high-stakes accountability requirements of the No Child Left Behind Act (NCLB, 2001) have shaped the face of public education in the United States. The law set ambitious goals for student achievement – all students were expected to be proficient in reading and math (as measured by performance on end-of-year grade-level-standards-aligned state assessments) by the year 2014. And, *all* meant *all*. Such rigorous standards for success communicated a “no excuses” model for education in which every school and every teacher was expected to do what needed to be done to ensure that every student, including those with disabilities, mastered grade-level content. NCLB directives gave teeth to the provisions for full participation of students with disabilities in district and state assessments of the Individuals with Disabilities Education Act (IDEA, 1997) and further strengthened the growing national commitment to have students with disabilities fully included in general education instruction. NCLB, IDEA 97, and IDEA 2004 made it clear that students with disabilities were to be held responsible for the same academic content and performance standards as their grade-level peers.

With only 14 years to achieve the ambitious goal of 100% student proficiency, the academic progress of students with disabilities quickly became the target of public and political

scrutiny. On the one hand, this attention promoted improved educational opportunities for school-aged populations of a group of students historically ignored and undervalued by the educational mainstream. On the other hand, it also placed incredible pressure on special educators, students, and their families to achieve what many considered impossible. Many educators argued that expecting all students to master age-based grade-level academic standards was unrealistic. After all, the students who receive special education and related services do so because their disability negatively impacts typical progress and achievement in school (IDEA 2004, § 300.8 (a)).

The AA-AAS

Recognizing the reality that achieving grade-level proficiency was unrealistic and inappropriate for a very small proportion of students (namely, those with significant cognitive disabilities) and yet striving to uphold the tenets of IDEA 1997, NCLB initially permitted states to develop an alternate assessment based on alternate academic achievement standards (AA-AAS). Standards for proficiency were aligned with academic content standards to promote access to the general education curriculum, but were “redefined” to reflect the academic skills and knowledge that were determined to be of the highest priority for the target group of students with severe disabilities (See NCLB 2001, §200.1 (d)). Again, in the spirit of full inclusion, the law assumed that the vast majority of students with disabilities should take the regular assessment alongside their non-disabled peers. So, although no cap was placed on the number of students with significant cognitive disabilities districts could assign to the AA-AAS, only 1% of the district’s overall proficient population that counted toward adequate yearly progress could come from students scoring proficient on the AA-AAS.

The AA-MAS

Despite the flexibility afforded by the AA-AAS, some leaders, educators, parents, and advocates

argued that there were many students with disabilities who were “left behind” by the mandated accountability system. These students did not have significant cognitive disabilities and were therefore ineligible to take the AA-AAS. However, they did have disabilities that negatively impacted grade-level achievement so substantially that mastery of grade-level content, particularly within one academic year, was seen as highly unlikely if not impossible. These persistently low-performing students with disabilities were commonly referred to as “the gap kids” because neither assessment option was an appropriate measure of their knowledge and skills. In April 2007, the US Department of Education responded to these concerns by permitting states to develop another test – an alternate assessment based on modified academic achievement standards. Legislators intended this additional flexibility to apply to a second small group of students with disabilities. This small group were those students “who can make significant progress, but may not reach grade-level achievement standards within the same time frame as other students, even after receiving the best-designed instructional interventions from highly trained teachers” (USDE, 2007, Part 200). The test was to cover the same breadth and depth of grade-level content standards as the regular assessment yet was to be less rigorous than the regular assessment (USDE, 2007). Similar to the regulations applied to the 1% test, schools could only count up to 2% of the tested population as proficient toward AYP based on performance on the AA-MAS.¹

Federal regulations do not explicitly specify which students in special education should participate in the AA-MAS. Instead, the federal guidance stipulates only that students who participate in the AA-MAS (a) have a disability

that precludes grade-level achievement; (b) have IEP goals aligned to grade-level academic content; (c) experience persistent academic failure despite high-quality instruction; and (d) are unlikely to achieve grade-level proficiency within a year’s time even if significant growth occurs (USDE, 2007). Individual states are charged with the daunting tasks of using these liberal criteria to zero in on the appropriate target population. Furthermore, states are to develop and disseminate explicit guidelines to aid IEP teams in assigning “the right” students to this new test. To complicate matters further, the USDE provided little specificity as to the achievement span states should consider when conceptualizing the AA-MAS target population. “The students who participate in assessments under this option are not limited to those who are close to achieving at grade level or who are relatively far from achieving at grade level” (Title I—Improving the Academic Achievement of the Disadvantaged, 2007, p. 17749). This range of possibilities appears to apply to all non-proficient students with disabilities and to contradict the aforementioned criteria detailed in the Non-Regulatory Guidance requiring persistently low academic performance.

Pennsylvania’s GSEG Project

The apparent vagueness of the federal guidance led many states to pull together a group of advisors to assist in determining whether their state would develop and implement the AA-MAS. And, if so, to whom it would be targeted, what it would look like, and what it would measure. In 2007, the Pennsylvania’s Bureau of Special Education was awarded a General Supervision Enhancement Grant (GSEG) by the US Department of Education’s Office of Special Education Programs to engage in research and inquiry about issues related to the development and implementation of the AA-MAS in the state. The primary goals of the grant were to (a) identify the target population for the test, and (b) to investigate those students’ opportunity to learn grade-level content so as to inform and improve

¹ “Under specific limited circumstances, States and LEAs may exceed the 2% cap. The 2% cap may be exceeded only if a state or LEA is below the 1% cap for students with significant cognitive disabilities who take the AA-AAS. For example, if the number of proficient and advanced scores on the AA-AAS is 0.8%, the State or LEA could include 2.2% of the proficient and advanced scores on AA-MAS in calculating AYP” (Modified Academic Standards, Non-regulatory guidance, 2007, p. 36).

the state's framework for standards-aligned instruction and assessment. The GSEG team, which included the chapter authors, was one group charged with providing guidance to state leaders as they moved forward with the development and implementation of the AA-MAS in Pennsylvania. The team conducted various activities to gather information related to defining the target population. These activities and the guidance developed based on the findings from the activities are discussed next.

GSEG Activities

Survey

The first information-gathering activity conducted by the GSEG team was a small-scale survey of special education teachers in schools identified as key research sites for the grant (see Table 17.1). Survey results are summarized in Tables 17.1, 17.2, 17.3, and 17.4. The survey's purpose was to help describe the opportunities students with IEPs in grades 5, 8, and 11 have had to learn the eligible grade-level content of the regular grade-level test (the Pennsylvania System of School Assessment, *PSSA*) in special education or general education settings, or both. Survey

Table 17.1 Pennsylvania GSEG survey: summary of response data

<i>Survey response data</i>	<i>Count</i>
Number of school districts represented	6
Number of schools represented	20
Number of surveys distributed	110
Number of teacher respondents	109
Fifth grade	52
Eighth grade	32
Eleventh grade	25

questions were divided into four topic areas: (a) *student information* for which teachers were asked to provide academic achievement information about a target student or students whom they considered to be “persistently low performing” and a candidate suited for the AA-MAS based on the criteria detailed in federal guidance (see Table 17.2); (b) *teacher information* for which teachers were asked to provide data about their professional experience and professional development related to the AA-MAS and standards-aligned instruction for students with IEPs (see Table 17.3); (c) *opportunity to learn data* for which teachers were asked to quantify the target students' opportunity to learn the eligible content of the state reading test (see Table 17.4); and (d) *IEP goals and instructional access* that spoke to

Table 17.2 Pennsylvania GSEG survey: summary of targeted student data

<i>Topic area: Targeted student information</i>	<i>Count</i>	<i>%</i>
Disability category		
Specific learning disability	78	72
Mental retardation	11	10
Emotional/behavioral disorder	8	7
Autism	3	3
NOS (Not otherwise specified)	9	8
Grade-level academic content standards-aligned IEP		
Yes	69	63
No	15	14
Unsure	20	18
Qualitative analysis of teacher reports of student competency with grade-level reading skills		
Number of narrative descriptions detailing student mastery of grade-level content	9	8
Number of narrative descriptions detailing student progress/growth on grade-level content	107	98
Number of narrative descriptions detailing deficient student progress with grade-level content	109	100

Table 17.3 Pennsylvania GSEG survey: summary of teacher data

<i>Topic area: Teacher information</i>	<i>Count</i>	<i>%</i>
Total teaching experience		
Teaching 1–5 years	33	30
Teaching 6–10 years	22	20
Teaching 11–15 years	11	10
Teaching 16–20 years	16	15
Teaching > 20 years	27	25
Experience teaching reading/language arts		
Teaching 1–5 years	48	45
Teaching 6–10 years	20	19
Teaching 11–15 years	6	6
Teaching 16–20 years	15	14
Teaching > 20 years	17	16
Experience teaching students with disabilities		
Teaching 1–5 years	34	31
Teaching 6–10 years	23	21
Teaching 11–15 years	17	16
Teaching 16–20 years	12	11
Teaching > 20 years	23	21
Advanced/additional degrees		
Education beyond bachelor's degree	93	85
Provisional teaching certificate	32	29
Permanent teaching certificate	72	66
Dually certified in elementary education	51	47
Dually certified in secondary education-English/LA	22	20
Professional development/training in Standards-aligned Instruction/IEPs for students with disabilities		
Attended training monthly	0	0
Attended training several times per year	61	56
Attended training once per year	16	15
Never attended training	23	21
Unsure if they have attended training	9	8
Reported that training impacted instructional practice	57	52
Overall preparedness to align instruction/IEPs with grade-level academic content standards		
Extremely prepared	44	40
Somewhat prepared	49	45
Minimally prepared	5	5
Unprepared	11	10
Unsure	0	0
Eleventh-grade teacher reported preparedness to align instruction/IEPs with grade-level academic content standards		
Extremely prepared	4	16
Somewhat prepared	6	23
Minimally prepared	5	20
Unprepared	10	40
Unsure	0	0

Table 17.4 Pennsylvania GSEG survey: snapshot of targeted eleventh-grade students

<i>Topic area: Targeted student information</i>	<i>Count</i>		<i>%</i>					
Primary instructional setting for reading instruction								
General education classroom	23		94					
Special education setting (resource room, learning support, pull-out support, etc.)	2		6					
Opportunity to learn Grade-level content								
Assessment anchor descriptor	<i>Not Taught</i>		<i>Exposure</i>		<i>Mastery</i>		<i>Unsure</i>	
	Count	%	Count	%	Count	%	Count	%
<i>Understands fiction appropriate to grade level:</i>								
R11.A.1.1 Identify and apply the meaning of vocabulary	17	68	5	20	1	4	1	4
R11.A.1.2 Identify and apply word recognition skills	18	72	5	20	2	8	–	–
R11.A.1.3 Make inferences, draw conclusions, and make generalizations based on text	20	80	3	12	–	–	2	8
R11.A.1.4 Identify and explain main ideas and relevant details	18	72	5	20	1	4	–	–
R11.A.1.5 Summarize a fictional text as a whole	18	72	5	20	2	8	–	–
R11.A.1.6 Identify, describe, and analyze genre of text	18	72	5	20	2	8	–	–
<i>Understands nonfiction appropriate to grade level:</i>								
R11.A.2.1 Identify and apply the meaning of vocabulary in nonfiction	17	68	5	20	–	–	2	8
R11.A.2.2 Identify and apply word recognition skills	18	72	5	20	2	8	–	–
R11.A.2.3 Make inferences, draw conclusions, and make generalizations based on text	20	80	3	12	–	–	2	8
R11.A.2.4 Identify and explain main ideas and relevant details	18	72	5	20	1	4	–	–
R11.A.2.5 Summarize a nonfictional text as a whole	18	72	5	20	1	4	–	–
R11.A.2.6 Identify, describe, and analyze genre of text	16	64	5	20	2	8	2	8
<i>Understands components between and within texts</i>								
R11.B.1.1 Interpret, compare, describe, analyze, and evaluate components of fiction and literary nonfiction	22	88	3	12	–	–	–	–
R11.B.1.2 Make connections between texts	19	76	5	20	1	4	–	–
R11.B.2.1 Identify, interpret, describe, and analyze figurative language and literary structures in fiction and nonfiction	21	84	2	8	1	4	1	4
R11.B.2.2 Identify, interpret, describe, and analyze the point of view of the narrator in fictional and nonfictional text	19	76	4	16	1	4	1	4
R11.B.3.1 Interpret, describe, and analyze the characteristics and uses of facts and opinions in nonfictional text	18	72	5	20	2	8	–	–
R11.B.3.2 Distinguish between essential and nonessential information within or between texts	18	72	5	20	2	8	–	–
R11.B.3.3 Identify, compare, explain, interpret, describe, and analyze how text organization clarifies meaning of nonfictional text	20	80	2	8	–	–	3	12

Not Taught = respondents reported that the grade-level skill(s) addressed by the assessment anchor was/were “Not Taught” to the target student; *Exposure* = respondents reported that the student was “Exposed To” the grade-level skill(s) addressed by the assessment anchor; *Mastery* = respondents reported that the grade-level skill(s) addressed by the assessment anchor was/were taught “To Mastery”; *Unsure* = respondents reported they were “Unsure” as to whether or not any level of instruction had occurred for the grade-level skills addressed by the assessment anchor

the students' level of academic functioning, the alignment of their IEP goals, and nature of their individualized instruction overall.

Respondents overwhelmingly (72%) identified their lowest-performing students with specific learning disabilities as the population of students in need of an alternate assessment that better captures what those students know and can do (Table 17.2). Overall, survey findings raise serious concerns about target students' opportunity to learn grade-level academic content. When asked to briefly describe the reading skills that targeted students had "mastered," only 8% reported that their lowest-performing students with IEPs had mastered any grade-level skills whereas a majority (98%) reported that students were instead "making progress/demonstrating growth" with the support of significantly adapted and modified instructional techniques and materials (Table 17.2). Analysis of teachers' narrative descriptions of students' academic strengths and weaknesses revealed that targeted students were making the most progress on grade-level skills related to Language Arts goals such as spelling, writing mechanics, and grammar.

Conversely, grade-level reading-related skills, specifically comprehension and vocabulary tasks, were highlighted as most troublesome for students. In fact, 64% of the present level of academic performance details on the target students' IEPs summarized reading assessment data, evidencing that fifth-grade students were achieving approximately two grade levels below peers. In comparison 60 and 72% of the IEPs for targeted eighth- and eleventh-grade students, respectively, suggested they were 3–4 years behind. Additionally, while 63% of teachers overall reported that students' reading IEP goals were aligned with grade-level standards (Table 17.2), analysis of actual IEP goals revealed that over 90% of those goals focused on improving instructional-level reading fluency and 58% of those goals focused on applying grade-level comprehension skills to instructional-level text. In contrast, 14% indicated that target students' IEP goals were not aligned with grade-level standards given how far below grade level the students were performing (Table 17.2). Instead, these IEPs

avored use of significantly modified and adapted instructional materials to promote growth and progress at the students' individual instructional level as measured by progress-monitoring and curriculum-based measurement. Interestingly, 18% of survey participants reported that they were "unsure" as to whether or not students' IEPs were standards-based (Table 17.2). The primary reason noted for this confusion was a lack of professional development/training in standards-based IEP development.

Positively, the majority of teacher respondents reported feeling "extremely" (40%) or "somewhat" (45%) prepared to develop standards-aligned IEPs and deliver standards-aligned instruction (Table 17.3). However, a closer examination of response trends revealed that elementary-level teachers reported feeling more knowledgeable about and better equipped to teach grade-level content to students than did eighth- or eleventh-grade teachers. Specifically, only 16% of the eleventh-grade respondents rated themselves as feeling "extremely" prepared to align instruction to grade-level academic content. Another 23% rated themselves as feeling only "somewhat" prepared, while the remaining 60% reported feeling "minimally" or "unprepared" to align their curricula and instruction with assessment anchors and content standards (Table 17.3). (Given the large number and broad scope of Pennsylvania's academic content standards, the state developed a subset of "Assessment Anchors," which focused on the grade-level standards assessed at each grade level on the state test to better equip teachers to prepare students for the test.) Moreover, when asked to review assessment anchors and eligible reading content for the eleventh-grade PSSA and indicate the level of instruction (i.e., "Not taught," "Exposure," "Mastery," or "Unsure") they provided to targeted students in special education, no more than 3 teachers out of 25 rated any grade-level content as "Taught to Mastery" (Table 17.4). Five teachers rated the majority of eleventh-grade content as instructed at the level of "Exposure" (Table 17.4). Teachers' narrative notes indicated that students were introduced to the big ideas related to the eleventh-grade content standards through significantly

modified or adapted instructional materials and assignments.

IEP data reinforced these assertions indicating that below-level text, text readers, and books on tape were required as part of the specially designed instruction provided to students to promote curricular access. The remaining eleventh-grade surveys (ranging from 16 to 21 respondents for any given assessment anchor) reported that over the bulk of the eligible eleventh-grade content was “Not Taught” to the target students (Table 17.4). Teacher comments suggested instead that those students were working on mastering reading and language arts goals directly related to transitional planning for post-school life. Respondents’ narrative notes explained that had assessment anchor language not included “appropriate to grade level text,” they could have rated all content as being instructed at the level of “Exposure” or “Mastery.” All eleventh-grade respondents expressed some level of concern that the state assessment anchors and eligible test content related to comprehension and reading skills require that those skills be applied to “grade level text.” (In Pennsylvania, as in many other states, “grade level text” equates to permissioned passages selected from a variety of eleventh-grade reading material, considered to be “authentic” literature, such as excerpts from novels, classic literature, and high school texts/anthologies.) Respondents reported that unless below grade-level reading passages were used, it would be highly unlikely for the students they saw as the target population to demonstrate proficiency, regardless of what types of accommodations were applied to the test.

Ninety-four percent of the eleventh-grade teachers who responded to the surveys indicated the general education setting was the primary place of instruction for all reading instruction provided to the target students (Table 17.4). Three out of four of these reported co-teaching with the general education Reading/Language Arts teacher; the remaining teachers indicated that they consulted with the general education teacher to plan for differentiated instruction to support the target students’ learning during reading time without special education teacher support.

This is somewhat disconcerting given the limited research base supporting the positive effects of the co-teaching/consultation model and differentiated instruction on the academic learning of students with disabilities (especially those with mild disabilities) who are fully included in the general education classroom (Fuchs & Fuchs, 1998; O’Sullivan, Ysseldyke, Christenson, & Thurlow, 1990; Zigmond & Matta, 2004).

Overall, survey results demonstrated that respondents clearly felt that the AA-MAS would be most appropriate for a small group of students who had previously performed at the lowest performance level (below basic) on the state assessment. However, the respondents also expressed great concern regarding how to make the test challenging and aligned to grade-level standards, and at the same time “doable” for a group of students who are significantly behind a majority of their grade-level peers. Further, the limited connection between grade-level standards and instruction provided to students the respondents identified as potential AA-MAS takers raises questions regarding the appropriate content for this assessment.

Focus Group

To extend the information provided by the survey, a series of stakeholder focus groups were convened statewide. In all, 110 participants (including parents of students with disabilities, general and special educators, state-level personnel, content area teachers, administrators, school psychologists, curriculum specialists, teacher trainers, and related service personnel) reviewed the federal guidance about the AA-MAS to discuss issues related to identifying the target population, developing modified academic achievement standards, and anticipating the implications and potential impact of the AA-MAS on educational and assessment experiences of students with disabilities. Three major themes arose from these discussions. First, a majority of focus group participants agreed that the students most appropriate for the test are those who are “far below” grade level. The rationale for this choice centered

on many of the points already raised in this chapter and in other publications on the 2% option. Most salient, however, was the belief that targeting only the lowest achievers would accomplish two goals: (a) it would guard against over-identification of students for this “special education test” and (b) it would prompt the state to focus its efforts on improving the regular assessment in terms of universal design and effective accommodations so that more students could be proficient on the regular assessment. However, concern that a “grade-level” test would still be too difficult for very low-performing students was very high, especially if reading passages on the AA-MAS were identical to (or equal in reading level to) those on the regular test. Nonetheless, some argued that any efforts to make the test more accessible and the testing experience more amenable for students so used to failure would be a welcome boost of self-confidence and decrease in stress level when taking the test – all realistic benefits given the unrealistic likelihood of proficiency.

Next, focus group participants’ concerns revolved around the notion of “pretending” that some students are proficient fifth graders (or sixth graders, or seventh graders, etc.) when they are in fact significantly struggling with academic content 2 or 3 years below grade level. Participants had a difficult time resolving the fact that while the AA-MAS may result in improved testing experiences for severely struggling students with disabilities, it may not result in improved instructional experiences for those students. These sentiments are echoed in the literature base (see Elliott, Kettler, & Roach, 2008; Marion, 2007).

Finally, discussion about developing modified academic achievement standards revealed that participants were greatly confused by changes in the federal requirements related to the rigor of the AA-MAS. All participants enthusiastically supported the idea that the modified achievement standards would reflect reduced breadth or depth of grade-level content. Paring down and prioritizing the academic skills addressed on the assessment paralleled what they viewed as crucial everyday instructional practices to promote student learning (Carnine, 1994; Kame’enui,

Carnine, Dixon, Simmons, Coyne, 2002). In fact, it is what most of the special educators said that they were trained to do. Therefore, the revisions to federal guidance deleting the references “reduced breadth or depth” and inserting the phrase “modified academic achievement standards must be challenging for eligible students, but may be less difficult than grade-level academic achievement standards” (§ 200.1 (e) (1)(ii)) were seen as being quite perplexing. Confusion over the terminology “less difficult” abounded. Frustrations with the apparent disconnect between the assessment requirements and real-world instructional practices were clear.

Analysis of PSSA Performance Trends for Students in Special Education

A third activity that the GSEG group conducted to assist in making recommendations about the AA-MAS was a trend analysis. Three consecutive years (2006, 2007, 2008) of end-of-year PSSA performance data were gathered for four cohorts of students in special education (enrolled in grades 3, 4, 5, and 6 during the 2006–2007 school year). Data were gathered on students who had an IEP at any point during the 3-year window and who had at least 2 years of reported PSSA scores.

The analysis was a simple examination of movement between proficiency levels across years. Results, displayed in Table 17.5, are presented only for reading; performance for math followed a similar pattern. First, movement out of the below-basic classification happens rarely; 71–88% of students who perform in the below-basic range repeat this level of performance the following year. The percentage of students who do move from below-basic to proficient or advanced in 1 year is fairly small (with percentages around 4% excluding the high of 10% for the sixth-grade cohort’s 2007–2008 performance). In comparison, an average of 29% of students who score in the basic range moved up to proficiency the following year. Unfortunately, an average of 33% of students moved down to the below basic range in the subsequent year. These data provided some

Table 17.5 Analysis of PSSA reading performance trends for children in special education

<i>Grade 3 cohort</i>					
2007 Performance					
		<i>Below basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>
2006–2007					
2006 Performance	<i>Below basic</i>	4990 (79%)	1029 (16%)	310 (5%)	16 (0.3%)
	<i>Basic</i>	700 (28%)	931 (38%)	775 (31%)	61 (3%)
	<i>Proficient</i>	218 (7%)	628 (21%)	1652 (56%)	454 (15%)
	<i>Advanced</i>	28 (2%)	59 (5%)	470 (39%)	656 (54%)
2008 Performance					
		<i>Below basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>
2007–2008					
2007 Performance	<i>Below basic</i>	4753 (88%)	554 (10%)	104 (2%)	8 (1%)
	<i>Basic</i>	1127 (49%)	821 (36%)	339 (15%)	21 (1%)
	<i>Proficient</i>	359 (14%)	846 (34%)	1129 (45%)	163 (7%)
	<i>Advanced</i>	28 (4%)	59 (7%)	383 (48%)	331 (41%)
<i>Grade 4 cohort</i>					
2007 Performance					
		<i>Below basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>
2006–2007					
2006 Performance	<i>Below basic</i>	6420 (88%)	679 (9%)	165 (2%)	5 (0.1%)
	<i>Basic</i>	1843 (50%)	1340 (36%)	496 (13%)	12 (0.3%)
	<i>Proficient</i>	567 (16%)	1254 (35%)	1557 (44%)	194 (5%)
	<i>Advanced</i>	32 (3%)	73 (6%)	647 (54%)	445 (37%)
2008 Performance					
		<i>Below basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>
2007–2008					
2007 Performance	<i>Below basic</i>	5903 (74%)	1658 (21%)	428 (5%)	16 (0.2%)
	<i>Basic</i>	633 (22%)	1141 (40%)	999 (35%)	68 (2%)
	<i>Proficient</i>	144 (7%)	426 (20%)	1149 (53%)	441 (20%)
	<i>Advanced</i>	7 (2%)	9 (3%)	121 (33%)	226 (62%)
<i>Grade 5 cohort</i>					
2007 Performance					
		<i>Below basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>
2006–2007					
2006 Performance	<i>Below basic</i>	7768 (77%)	1989 (20%)	347 (3%)	18 (0.2%)
	<i>Basic</i>	849 (27%)	1390 (44%)	843 (27%)	64 (2%)
	<i>Proficient</i>	219 (8%)	661 (24%)	1372 (49%)	528 (19%)
	<i>Advanced</i>	13 (2%)	31 (6%)	155 (29%)	339 (63%)
2008 Performance					
		<i>Below basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>
2007–2008					
2007 Performance	<i>Below basic</i>	5883 (74%)	1712 (22%)	346 (4%)	9 (0.1%)
	<i>Basic</i>	931 (26%)	1552 (43%)	1081 (30%)	59 (2%)
	<i>Proficient</i>	130 (6%)	446 (20%)	1260 (57%)	375 (17%)
	<i>Advanced</i>	3 (0.5%)	24 (4%)	181 (28%)	445 (68%)

Table 17.5 (continued)*Grade 6 cohort*

		2007 Performance			
2006–2007		<i>Below basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>
2006 Performance	<i>Below basic</i>	7202 (81%)	1423 (16%)	208 (2%)	18 (0.2%)
	<i>Basic</i>	1547 (36%)	1858 (43%)	880 (20%)	60 (1%)
	<i>Proficient</i>	261 (9%)	876 (29%)	1481 (49%)	411 (14%)
	<i>Advanced</i>	15 (2%)	34 (4%)	306 (34%)	545 (61%)
		2008 Performance			
2007–2008		<i>Below basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>
2007 Performance	<i>Below basic</i>	6761 (71%)	1780 (19%)	787 (8%)	201 (2%)
	<i>Basic</i>	1016 (23%)	1375 (31%)	1739 (39%)	381 (8%)
	<i>Proficient</i>	115 (4%)	324 (11%)	1290 (45%)	1159 (41%)
	<i>Advanced</i>	12 (1%)	10 (1%)	123 (13%)	770 (84%)

evidence that if any students in special education are likely to move from one of the failing categories of performance into a passing one, it would most likely be students who scored in the basic level – a possible reason to exclude such a student from taking the AA-MAS.

Recommendations from the GSEG

Based on the results of the survey, focus groups, examination of performance trends, and ongoing discussions between the GSEG team members, the group developed a set of recommendations for state leaders as they moved forward with development of Pennsylvania’s modified exam. The GSEG team provided these recommendations to the Pennsylvania Bureaus of Special Education (BSE) and Accountability and Assessment (BAA). A summary of the recommendations is provided next.

Students who have previously performed in the lowest performance category on the regular assessment should be eligible to take the AA-MAS. According to federal guidance, the modified assessment was intended for “a group of students with disabilities who can make significant progress, but may not reach grade-level achievement standards within the same time

frame as other students, even after receiving the best-designed instructional interventions from highly trained teachers” (Title I—Improving the Academic Achievement of the Disadvantaged, 2005). The GSEG team interpreted this to mean that the AA-MAS is intended for a group of students who, despite the best efforts of their teachers, cannot reasonably be expected to achieve on grade-level performance within one academic year. The team recommended that students performing in the “low below basic range” (the lowest possible performance category) on the state assessment are the students most likely to fall within this category. It is unlikely that any of them, even those who make significant academic progress in 1 year, will achieve proficiency. Selecting this lowest-performing group of students with disabilities, not those who are performing slightly below expectations (i.e., “almost made proficient”), is consistent with federal guidance and within state mandates to hold students with disabilities to the highest possible standards of performance.

Zigmond and Kloo (2009) debated the consequences associated with targeting “almost proficient” students versus “nowhere near proficient students” for the AA-MAS. The dangers of over-assignment to the test were central to our concerns for permitting students close to achieving

at grade level to participate in the AA-MAS. This group of students logically fits the USDE's designation of students likely to make significant progress, just not enough progress within a year's time. However, if those students participate in the AA-MAS, so will their lower-performing peers because it would be impossible to justify offering a less difficult, more accessible test to students close to mastering grade level but a more difficult, less accessible test to their severely struggling classmates. As such, the intention that only a small percentage of students (i.e., about 2%) would take part in the modified assessment would be missed. Further, depending on standard setting, it is possible that more students than intended could achieve proficiency on the AA-MAS – and, these “extra” students would not count toward AYP.

Allowing all (or the majority of) non-proficient students with disabilities to take the AA-MAS creates a special education track that is fundamentally misaligned with the intent of IDEA and efforts for inclusive practices. The GSEG team suggested that if all or most students with disabilities who perform below proficient are allowed to take the AA-MAS, the new test would become the “special education test” and all students in special education would be held to modified or “easier” achievement standards. This could result in teachers prioritizing some skills over others, spending more/less time on certain skills and concepts than others, and differentiating between which skills students with IEPs should master versus those non-disabled peers should master. This would set back decades of progress, and the ideals of IDEA and NCLB for promoting high expectations for students with disabilities in the spirit of educational accountability, standards-based reform, and full inclusion would be diminished (Samuels, 2007; Zigmond & Kloo, 2009).

In contrast, targeting the population of students who are very far from meeting proficiency standards captures a smaller number of students. Unfortunately, such consistently depressed academic performance suggests that these students achieve significantly below grade level. Much research, and our own survey data from teachers, show that this achievement gap only con-

tinues to widen as students progress through school, especially for students with learning disabilities who constitute the largest proportion of low-performing students with disabilities (Wagner, Newman, Cameto, Levine & Garza, 2006; Schiller, Sanford, & Blackorby, 2008). Although it is difficult to rationalize that they are the students the USDE expects to eventually “catch up” to their peers, it is this population of students for whom a less rigorous assessment seems the most ethically and educationally appropriate given their significant learning struggles (Zigmond & Kloo, 2009). That said, these students (particularly in the upper grades) are so far behind academically that if the test measures only grade-level content, it may be absurd to deem them “proficient” based on what they can actually do with the grade-level material – a point highlighted by our teacher interviews. Thus, if this is the target population, the strict adherence to grade-level content may need to be reconsidered.

It will be easier to rationalize/justify (i.e., explain to parent groups, advocates, and school personnel) the use of a “different” test for the lowest-performing group of students with IEPs than for students currently scoring in the basic range. GSEG team members suggested that students scoring in the lowest performance range are those whose disability truly precludes grade-level achievement of proficiency and whose persistent academic difficulties make them unlikely to achieve grade-level proficiency within 1 year even with intense, research-based, remedial instruction. These are the students who perform hardly above the chance level on the multiple-choice sections of the reading, math, and science tests. Some may argue that this population of students needs a modified test so they can demonstrate what they know and can do. However, many of these students are likely already demonstrating what they can do with grade-level content – which is not much. Thus, we argue that while the goal of enhancing the accessibility of achievement tests is important, it may also be necessary to assess other, likely below-grade level, skills if we are to create a test on which this group of students can honestly demonstrate “proficiency.”

It seems much more difficult to justify redefining what it means to be proficient and providing a “different” or “easier” test to a group of students who are almost proficient. Students in the basic range are “almost” proficient. The performance of this group of students might be improved by more intensive instruction and by more attention paid to making the standard PSSA more accessible through application of more principles and better principles of universal designs for learning (UDL). Focusing efforts on improving the regular PSSA would result in a better measure of student ability for students scoring in the high basic range than the AA-MAS.

Targeting students scoring in the low below-basic group positions districts to increase the percentage of students counting toward AYP. Because at least some of the students with IEPs, mostly those scoring in the basic range, will achieve proficiency on the state exam during the next school year, the GSEG team recommended that placing students into the modified assessment, who might achieve proficiency on the standard assessment if they and their teachers really did expend maximum effort, could actually reduce the proficiency count for students with IEPs, not increase it. If you assign to the AA-MAS both students very unlikely to be proficient next year (i.e., below-basic students) *and* students who have some possibility of being proficient next year (i.e., basic students), the total percent of proficient students based on the AA-MAS would be equal to 2% of the tested population. However, if the students who had a possibility of being proficient were assigned to the regular assessment and they did achieve proficiency, schools could count a number of AA-MAS scores equivalent to 2% of the tested population *plus* the students who achieved proficiency on the regular test.

In other words, if the modified assessment is targeted toward the lowest-performing group of students with disabilities who take the regular assessment, schools will be able to count as proficient a number of these students equivalent to 2% of the tested population *and* any higher-performing students who achieve proficiency on the regular assessment. If, instead, the highest-performing group of students with disabilities is allowed to take the AA-MAS, the maximum

number of students who will be counted as “proficient” will be capped at 2%. This number reflects the reality that (a) IEP teams are unlikely to allow higher-performing students to take the AA-MAS, but require the regular test for lower-performing peers, and (b) even if lower-performing students are required to take the regular test, their likelihood of scoring as proficient is quite low. Put differently, the higher the performance level of students who are allowed to take the AA-MAS, the lower the total number of students with disabilities who will count as “proficient.”

An underlying assumption of this argument is that schools would have a sufficient number of below-basic students who could achieve proficiency on the modified assessment to reach the maximum 2% count. Based on our review of school data, we believe that schools will have an ample number of below-basic students. However, we acknowledge the reality that the design of each state’s AA-MAS and the related standards setting in each state will influence which students will be able to achieve proficiency on each state’s individual test (i.e., the more closely related to the regular assessment, the less likely that the most far-behind student will be able to demonstrate proficiency). However, regardless of the specific test, it is fairly clear that wherever the cut-point for assigning students into the AA-MAS is set, schools will only be able to only count around 2% of students below that cut-point toward AYP.

This fact may seem obvious, but it is important to consider the consequences of allowing a larger number of students to take the AA-MAS. Focus groups participants believed that regardless of which group of students is identified as the target population, that *all* students with disabilities scoring at that level and below it would also be identified. Specifically, if students scoring in the basic range are eligible, all students not achieving proficiency (thus, those in the below-basic range) will also be eligible by definition. As one participant who is a special education teacher and parent of a student with a disability commented, “One wouldn’t expect students who are close to proficiency to take a ‘modified’ test while their lower-performing peers are expected to take the regular test.” Approximately 69%

or 97,000 students with IEPs in Pennsylvania were not proficient on the 2008 PSSA. Thus, IEP teams would be likely to allow more than 60% of students with IEPs (that's almost 6% of the students being tested each year) to take the AA-MAS. The USDE is clear in its guidance that the AA-MAS is intended for a "small group of students with disabilities." Targeting students in the basic performance categories would result in an "over-identification" for participation in the AA-MAS. This would likely impact a much larger percentage of students than was intended by the legislation.

Additional Considerations and Unintended Consequences

The recommendations of the GSEG leadership team provided guidance to the state that the target population should be a group of students in special education who have been performing consistently at the lowest level of performance on the regular state assessment. This guidance, however, did not address some of the larger issues with which school districts are currently grappling. We next address some of the most common reasons given for including a modified assessment as another option in the testing system and provide an argument for why some of these do not appear to be justifiable. Additionally, we examine what may be unintended consequences of focusing on the development of an alternate test to improve accessibility as opposed to improving the regular assessment and considering the learning expectations for students in special education.

Commonly Provided Rationale for the AA-MAS

The 2% test will help us meet AYP. One rationale for the development of the AA-MAS was to create a better measure that would allow more students to demonstrate what they know. However, it appears that another strong motivation for developing the AA-MAS option was

to provide schools with a way to decrease the negative impact of the test performance of students in special education on adequate yearly progress. Under NCLB's AYP system, schools must show that increasingly higher percentages of *all* students are passing the state assessment. This includes students in the subcategory groups of special education, English language learners, minorities, and students from low-income families. As the percentage of students required to demonstrate proficiency has increased, a larger number of schools have failed to meet AYP solely based on the performance of their special education students. The 2% test seems to be an appeasement to district and state leaders who have protested the notion that students enrolled in special education should take the same assessment as their typically developing peers (Lazarus & Thurlow, 2009). The fact that states in the planning phases of developing an AA-MAS were permitted to automatically begin counting this additional 2% as a proxy supports the claim that improving AYP counts was a prime motivation for the modified assessment (Thurlow, 2008).

A deeper look into the impact of the 2% on whether or not a given school will meet AYP by deeming a small group of previously non-proficient special education students as now proficient reveals that the benefit will actually accrue to only a few schools, that the benefit may only help for a few years as the AYP target goal increases, and that the benefit will likely not help any school meet the requirement of having *all* students with IEPs proficient in math and reading by 2014. We note that this final point may be moot contingent upon the reauthorization of NCLB and the restructuring of the 100%-proficient-by-2014 policy. However, regardless of how AYP is reconceptualized, the benefits from the 2% option in terms of "AYP math" for any given school appear scant.

We acknowledge that many states are using additional calculations including confidence intervals, proxy counts, and growth models to calculate proficiency. For demonstration purposes, we are presenting a simplified version of calculating AYP. We are presenting a best-case

scenario in which the full 2% can be counted as proficient and, we acknowledge that some schools may be able to count a few more students as proficient depending upon the number counted under the AA-AAS (see Kettler & Elliott, 2009) – regardless, this doesn't change our overall point. In the most basic sense, the value of the AA-MAS in terms of meeting AYP is based on two numbers: the percentage of the student population enrolled in special education and the proficiency rate of this group of students. The lower the proportion of special education students to the general population, the greater the

boost given by the AA-MAS. The “boost” is a simple calculation: 2% of the student population from among those enrolled in special education. Thus, a school with a 15% special education population would be able to increase their proportion of special education students meeting proficiency by 2/15 or 13.3%. And, the greater the proficiency of the special education students, the more likely that the boost will be enough to meet the proficiency benchmark set by NCLB. The relationship between these two variables is illustrated in Table 17.6. In this table, the benefit in terms of AYP for schools with dif-

Table 17.6 Hypothetical examples of AYP benefits of a 2% increase in proficient special education students from the AA-MAS

Row	Total student population	% SpEd*	# SpEd	Actual % SpEd scoring proficient	Possible % SpEd proficient if full 2% from AA-MAS count	Resulting increase in % SpEd proficient	2% Helped school meet AYP
1	700	6	42	30	63	33	NO
2	700	8	56	30	55	25	NO
3	700	10	70	30	50	20	NO
4	700	12	84	30	47	17	NO
5	700	14	98	30	44	14	NO
6	700	16	112	30	43	13	NO
7	700	18	126	30	41	11	NO
8	700	20	140	30	40	10	NO
9	700	6	42	40	73	33	YES
10	700	8	56	40	65	25	NO
11	700	10	70	40	60	20	NO
12	700	12	84	40	57	17	NO
13	700	14	98	40	54	14	NO
14	700	16	112	40	53	13	NO
15	700	18	126	40	51	11	NO
16	700	20	140	40	50	10	NO
17	700	6	42	50	83	33	YES
18	700	8	56	50	75	25	YES
19	700	10	70	50	70	20	YES
20	700	12	84	50	67	17	NO
21	700	14	98	50	64	14	NO
22	700	16	112	50	63	13	NO
23	700	18	126	50	61	11	NO
24	700	20	140	50	60	10	NO

*SpEd= Students receiving special education services

fering proportions of special education students and special education student proficiency are compared. A stable population of 700 is used for the comparisons. The first point is that the 2% assessment gives the largest boost to schools with the lowest percentage of special education students. Schools with 6% of their population in special education (rows 1, 9, and 17) get an increase of 33.3% whereas schools in which 20% of students are in special education (rows 8, 16, and 24) only gain an additional 10% to count toward AYP. This realization is disconcerting because it appears that the benefit will not go toward many of the most struggling schools – many of the schools with the highest percentages of special education students are also schools serving greater numbers of minority students and students from economically disadvantaged families (see Skiba et al., 2008).

A second point is that even with the additional students counting toward AYP, only those schools that are already fairly close to the AYP target will surpass the goal with the additional count. In this demonstration, the AYP target is set at the 2011 level for mathematics, 67% of students achieving proficiency. In this case, there are only four schools for whom the boost made a difference (rows 9, 17, 18, and 19) – all with lower percentages of special education students and all with

higher levels of proficiency than that currently exhibited by many of our schools.

Another way to look at the impact is to compare the benefit for five actual campuses. Data from five Pennsylvania schools that have been renamed for purposes of this chapter are presented in Table 17.7. The campuses have varying proportions of special education students (ranging from about 12% to about 35%) and varying levels of special education student proficiency (ranging from about 10 to 46%). However, in none of these campuses does the 2% addition bump them up into meeting the 2011 AYP targets of 72% proficient in Reading and 67% proficient in Math. If this had been the 2010 school year (with slightly lower proficiency targets), the addition would have helped Robinson Elementary meet the target in mathematics only. Thus, the additional 2% of students from the IEP pool does not appear to help many schools make their AYP targets. And, higher SES schools that already have a relatively low number of special education students who are already performing near proficiency will likely be the ones to see the benefit. Schools with a large proportion of special education students and schools with low proportions of special education students scoring proficiently will see no AYP benefit from the AA-MAS. In many ways, the benefit of the 2% test is unfairly

Table 17.7 Potential AYP benefits to five pennsylvania schools from a 2% increase in proficiency from the AA-MAS

<i>School</i>	<i>Subject</i>	<i>Total # tested</i>	<i>Total # tested SpEd*</i>	<i>% SpEd</i>	<i>Actual % SpEd scoring proficient</i>	<i>Possible % SpEd proficient including 2%</i>
Wagner Elementary	Math	191	47	24.61	10.00	18.13
	Reading	191	47	24.61	15.00	23.13
Stargell Elementary	Math	680	81	11.91	30.00	46.79
	Reading	679	80	11.78	40.00	56.98
Clemente Elementary	Math	141	49	34.75	44.00	49.76
	Reading	141	49	34.75	21.00	26.76
Mazeroski Elementary	Math	355	42	11.83	39.00	55.90
	Reading	354	41	11.58	37.00	54.27
Robinson Elementary	Math	467	56	11.99	46.00	62.68
	Reading	468	56	11.97	45.00	61.71

*SpEd= Students receiving special education services

distributed – schools with the greatest resources get a much greater benefit than those with fewer resources.

However, if the new reauthorization of ESEA maintains the current targets of having 100% of students in special education on grade level by 2014 or if this goal is simply moved further out in time, the benefit of the 2% test in terms of AYP will eventually expire for virtually every school. Unfortunately, for schools like Mazeroski Elementary, the most feasible option for meeting the AYP benchmark for the IEP students would be to decrease the number of students with IEPs taking the exam to fewer than 40 – perhaps an IEP team could exit a student or two from special education, the school could encourage the students to transfer to another campus, or the administrators could encourage the family to keep the students home during the testing period. While this last set of circumstances may seem extreme, the pressure of meeting high-stakes assessment goals has led some to do even more questionable things (See Gabriel, 2010; Levitt & Dubner, 2005). Based on this information, we think that the AYP argument is not enough to justify including an AA-MAS in a state accountability system. The benefits are small and apply to only a few, primarily higher-SES, school districts. Another rationale is needed to justify the new assessment.

The 2% test will allow students with disabilities to better demonstrate their understanding of grade-level content. This rationale has a high level of face validity – it makes sense that if there is a group of students who are unable to demonstrate proficiency due to the nature of the assessment, then reasonable accommodations that do not invalidate the assessment should be provided. It is in this vein that states have developed accommodation guidelines for their regular state assessments (Thurlow, Elliot, & Ysseldyke, 1998). Some states, such as Pennsylvania, have different guidelines for accommodations that are allowed for all students and those that are only allowed for students in special education (Pennsylvania Department of Education, 2010).

The notion of creating a “better test” that allows more students to access that content and to demonstrate their knowledge is also supported

by work in universal design, cognitive load theory, and item modification research. An evolving body of research has explored the application of the principles of architectural universal design (Center for Universal Design, 1997) (meant to increase physical environment accessibility) to the educational realm. Universal design in instruction and assessment promote accessibility of curricular or test content so that diverse population of students can accurately demonstrate their knowledge uninhibited by design limitations (Johnstone, 2003). Thompson and colleagues (2002) highlight seven elements of universally designed assessments: (1) inclusive assessment population, (2) precisely defined constructs, (3) accessible, non-biased items, (4) amendable to accommodations, (5) simple, clear, and intuitive instructions and procedures, (6) maximum readability and comprehensibility, and (7) maximum legibility – characteristics that are salient for accountability tests, considering IDEA 2004’s requirements for accessible testing experiences for students with disabilities. Specifically, the law states that “the State educational agency (or, in the case of a district-wide assessment, the local educational agency) shall, to the extent feasible, use universal design principles in developing and administering any assessments” (§ 612(a)(16)(E)). Ultimately, the more universally designed the test, the greater the opportunity for students to demonstrate mastery of test content.

In addition to universal design, cognitive load theory (CLT) has influenced test development and assessment research. CLT research suggests that highly effective instruction is that which maximizes the learner’s ability to manage the mental effort needed to work through tasks without exhausting working memory capacity (Sweller, 1994). Recently, researchers have turned their attention to applying CLT to the assessment of student learning (Kettler, Elliot, & Beddow, 2009). This work is particularly relevant to item development for an AA-MAS intended for a population of significantly struggling learners. Because a student’s ability to demonstrate proficiency or mastery of a task is mitigated by his ability to manage the cognitive load of the

task, research posits that highly effective items are those which minimize the taxing cognitive demands of sorting through extraneous information that is unessential to the tested construct and maximize the amount of mental energy focused on zeroing in on central information to successfully complete the task.

Furthermore, research on item modification suggests that altering items to reduce language load (simplify sentence structure, segment text into manageable parts, etc.) to improve the construction of answer choices (shortening item stem, order choices logically, etc.), to decrease the number of multiple-choice options from four to three, and to enhance the regular test format (add white space, remove unnecessary visuals, etc.) may have significant positive effects on student performance on tests (Elliott et al., 2010; Hess, McDivitt, & Fincher, 2008; Rodriguez, 2005).

However, using the “better test” argument to support the development of a separate, modified test raises the question of whether the aim is to accommodate students or whether it is to modify the content the students are to learn. In the original guidelines, states were told that the AA-MAS could be based on a set of academic achievement standards that reflect a reduced breadth or depth of the grade-level content. In this case, students taking the AA-MAS could be expected to learn a little bit less than their grade-level peers or to do a little bit less with the grade-level content (i.e., cover the same content, but not be held accountable for processing the information at the same level of cognitive complexity). However, as Zigmond and Kloo (2009) highlighted, this notion doesn’t make sense in a system where a grade-level test is given annually. If you only learn a portion of third-grade content, what are you supposed to learn in fourth grade? The rest of third-grade or a limited amount of fourth-grade content? And, if the answer is the latter, how will missing portions of each grade level affect students years down the line? What portions of the academic standards are nonessential pieces that can be removed with little to no effect as students progress through the school system?

But, more recent regulatory guidance has removed the “reduced breadth and depth” language and has reiterated that the AA-MAS is an assessment that covers that same content as the regular test – it can be “a little less difficult” but it must be “a rigorous assessment of grade level knowledge” (Elliott et al., 2008). In addition, the test blueprints for the regular and AA-MAS assessments must be the same. It, therefore, appears that the test is to assess all of the same content, just be more accessible. If this is the intent of the AA-MAS, it is unclear why states are not focusing the efforts instead on making the regular test more accessible – particularly since we know that many students who are not in special education also struggle with accessing the regular assessment and fail to demonstrate proficiency. One argument for the different test may be that the types of accommodations that would be needed to make the test accessible might invalidate the test (Palmer, 2009). For example, reading the text of a reading passage aloud to a student would invalidate an assessment of reading comprehension. But, the same logic would apply to the AA-MAS. If a student has the text of an AA-MAS reading comprehension test read aloud to her or him, the test is no longer a reading comprehension test and it cannot be used to deem a student as proficient on grade-level reading comprehension. (For another viewpoint of why a different test may be the most accessible for students likely to qualify for the AA-MAS, see Kettler, 2011.)

We think that the rationale for developing an AA-MAS solely to enhance accessibility is also not strong enough of a rationale. Instead, we need to be clearer about what we want students taking the AA-MAS to learn and to demonstrate proficiency in. If it is grade-level content, we need to improve the regular assessment; if it is a modified version of grade-level content, we need to determine how to successfully move students through multiple years of reduced expectation in a way that is fair and that leaves students with a solid core of knowledge.

The 2% test will improve instruction for students with IEPs. One primary purpose of the current accountability system is to ensure that

students are receiving high-quality instruction and that this instruction is improving their achievement. One potential reason to develop an AA-MAS is to improve the instruction provided to students in special education. If teachers are provided with a clearer vision of the critical skills to be taught to students taking the 2% test, it is likely that instruction will begin to focus on these skills and higher scores on the assessment that measures the skills will follow. But, this effect on instruction will likely depend on whether the AA-MAS is meant to provide accommodations that allow access to the grade-level test or whether it is aligned with a modified version of academic standards. If the learning goals for students who take the 2% are exactly the same as those students taking the regular exam, it is unclear how this test could improve instruction – the goals for what a student is to learn are the same, only how this learning is measured has changed. On the other hand, if breadth or depth is reduced, teachers may have a greater ability to focus on what would be seen as a set of “essential” skills for the students taking the AA-MAS, a point noted in focus group discussions. This may yield positive results.

Quenemoen (2009) proposed another way in which the 2% may improve instruction for students in special education. Quenemoen argues that it is likely that there are a group of students who struggle to make proficiency on the regular assessment, who were placed into the AA-AAS by their IEP team, but for whom the alternate achievement standards are inappropriately unchallenging. If a test that is more academically challenging than the AA-AAS, but less challenging than the regular assessment, is available, teachers may raise their expectations for this group of students and provide them instruction that better targets their instructional needs. We believe that this is the most likely scenario for the 2% test to actually improve instruction; however, it will really depend on what the test looks like and how academically challenging it appears to be. If the target population for the test is not the type of student proposed by Quenemoen and the test instead appears to be a slightly easier version of the regular assessment, it is unlikely that the instruction provided

to this group of students will significantly change.

The 2% test is the ethical thing to do. This argument for implementing the AA-MAS is perhaps the most legitimate. If there is a student who will not pass the regular test, why not give her or him a shorter, easier exam to make for a more pleasant testing experience? Also, if there are students who cannot reasonably be expected to meet the learning expectations of the regular exam, it may be ethical to acknowledge this fact and to redefine the essential components of the academic standards that we really do expect all students to learn. And, if a student is unable to demonstrate her or his grade-level content knowledge because of the assessment, we should improve that assessment to ensure that the test is not the reason the student is performing poorly. All of these are good reasons to reexamine how we currently include students in special education in the accountability system. However, creating a better testing situation because it appears to be the right thing to do, does not forestall unintended consequences. We discuss these next.

Unintended Consequences

Special education tracking 2010. If schools are given the leniency to redefine which components of the general education curriculum are most important for students in special education, this may enhance teachers’ ability to focus instruction and to see improved student outcomes. However, it may also open up the door for a new type of special education tracking. Once a student starts taking the AA-MAS and is, therefore, only expected to learn the “essential skills,” it is likely the student would continue on the AA-MAS path. And, students who are on the AA-MAS path from fourth grade to graduation will likely have substantially different learning outcomes and experiences from those who are not. Perhaps this is okay and that we should acknowledge that we have a different set of expectations not only for students with the most significant cognitive impairments (i.e., students who take the AA-

AAS), but also for another group of students who are unlikely to be able to obtain mastery of the same amount of grade-level academic content as their typically developing peers.

Nonetheless, we feel that this may be one of the largest areas of confusion surrounding the 2% assessment. The students most likely to learn the largest portion of grade-level content are the students who are already performing close to proficient. This group, however, doesn't appear to be a logical target audience for the AA-MAS. Instead, the lowest-performing group of students in special education appears to be the better target group. But, feedback from educators in the field suggests that if this is the target group, we need to consider what we can realistically expect in terms of academic performance. In a sense, it may be helpful to consider whether the 2% test is more like the 1% or more like a more accessible regular assessment. If the logic of the 1% applies, the strong focus on grade-level academic content may need to be reconsidered and instead a redefined set of academic goals may be needed. Perhaps we should be a little more honest about which skills this group of students with a history of academic failure needs to obtain independence, employment, and happiness once they finish school. If, on the other hand, the logic is much more aligned with enhancing accessibility of the regular exam, perhaps we should redirect efforts to improving that assessment instead of creating a separate but supposedly equal alternate.

How much will the 2% test cost and is this the best use of funds? Another consideration that does not appear to have been discussed is the fiscal cost of implementing the AA-MAS. Palmer (2009) reported that one state estimated that the cost for implementing the AA-MAs would be over \$6 million. While we do not purport to be experts on school finance, it does seem fair to raise the question of whether the real-world benefits of the AA-MAS are worth the financial costs – particularly in a time in which many state departments of education and schools are facing budget crises. If the test does not meet the apparent goal of moving more school districts toward meeting AYP, perhaps these funds

would be better directed toward other avenues. For example, the same funds and efforts could be directed toward improving the accessibility of the regular exam – an effort that would benefit many students, including those in general education. Or, the funds could be focused on providing additional professional development and training to teachers aimed at enhancing the instruction provided to students in special education who are not making adequate growth. Further, states could even hire additional staff to provide more intensive intervention for students.

But, are we not just pretending? While it is oft forgotten, the current accountability system is aimed at answering one very simple question – at the time of the assessment, what percentage of students in a given grade have learned a sufficient amount of grade-level content to be called proficient? Schools can demonstrate improvement by doing better (i.e., having more students meet proficiency) with each year's new group of students in a given grade. The system is not concerned with individual growth over time or whether a higher percentage of students in a given cohort pass the exam each year. Thus, the measure is whether schools can do better with this year's group of third graders than they did with last year's; not whether more of the third graders did better when they got to fourth grade. One important factor in considering whether or not to adopt an AA-MAS is to think about the question for which an answer is sought. If we are asking whether more students in this year's group met grade-level expectations (and are, therefore, proficient) the test has to measure grade-level academic content.

If, on the other hand, we develop a substantially easier assessment and set the expectation for meeting proficiency on this assessment at a level that schools can be assured that at least 2% of their population can achieve proficiency, it seems that we are simply pretending. And, pretending doesn't answer the original AYP question – in fact, it muddies the waters and makes answering the question even more challenging. And, as states have been given the leeway to select different target populations (i.e., the “almost proficient” versus the “nowhere near proficient;” see

Zigmond & Kloo, 2009) and develop tests that are potentially of varying levels of difficulty, making comparisons of the proficiency of students across states becomes even more difficult.

Concluding Thoughts

We believe that pretending that students in special education have met grade-level expectations, when they have not actually done so, is not the honest solution. Instead, it is what we refer to as the “Unicorn syndrome” – the academic equivalent of gluing a horn to the forehead of a pony and telling her and others that she can fly. It is a disservice to the students, their parents, and their teachers to pretend they have achieved something they have not. The consequences of this wishful thinking will likely have the opposite effect of what is intended. It seems quite possible that allowing schools to falsely deem students as proficient lets schools off the hook, in a sense, in terms of what society expects them to accomplish (see Zigmond & Kloo, 2009). A more honest solution would be to acknowledge that special education is designed to meet the individual needs of a group of students who are in special education primarily because they have *not* been able to meet grade-level expectations. In this sense, perhaps, the accountability system should be asking a different set of questions for this group of learners – namely, how much academic growth occurred this year and how much closer to achieving meaningful academic and social goals is each student?

In this light, maybe, we also need to gain clarity on the purpose of special education. Is the aim to assist students with disabilities in meeting all of the goals of their general education peers? Or, is the purpose to provide an individualized program aimed at helping students meet a set of goals that may look quite *different* from those of their general education peers? Or, maybe, the purpose is both – depending on the student? Fuchs, Fuchs, and Stecker (2010) argue that the student-level focus, which is the historic root of special education, needs to be brought back into

focus. The authors contend that special education should go “back to the future” and embrace the model of experimental teaching in which special educators are expert instructors who are able to connect their instruction to each student’s individual needs and to ensure that each student is making adequate progress toward her or his educational goals. This approach does not involve any pretending – instead, it honors both the teacher and the student who are working to meet the challenges posed by disability and allows the progress toward individualized goals to be rewarded. If special education as a field can refocus and move away from pretending, we will be able to reconnect with the “individualized” and “special” components of our profession and better prepare our students for success in the real world that they are about to enter.

References

- Carnine, D. W. (1994). Introduction to the mini-series: Diverse learners and prevailing, emerging, and research-based educational approaches and their tools. *School Psychology Review*, 23(3), 341–350.
- Center for Universal Design. (1997). *What is universal design?* Center for Universal Design, North Carolina State University. Retrieved January, June 9, 2010, from <http://www.design.ncsu.edu>
- Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurtz, A. (2011). Creating access to instruction and test of achievement: Challenges and solutions. In S. N. Elliott, R. J. Kettler, P. A. Beddow & A. Kurtz (Eds.), *Handbook of accessible achievement tests for all students* (pp. 1–16). New York: Springer.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., et al. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children*, 76(4), 475–495.
- Elliott, S. N., Kettler, R. J., & Roach, A. T. (2008). Alternate assessments of modified achievement standards: More accessible and less difficult tests to advance assessment practices? *Journal of Disability Policy Studies*, 19(3), 140–152.
- Fuchs, D., Fuchs, L. S., & Stecker, P. M. (2010). The “blurring” of special education in a new continuum of general education placements and services. *Exceptional Children*, 76, 301–323.
- Fuchs, L. S., & Fuchs, D. (1998). General educators’ instructional adaptation for students with learn-

- ing disabilities. *Learning Disability Quarterly*, 21(1), 23–33.
- Gabriel, T. (2010, June 10). Under pressure, teachers tamper with tests. *The New York Times*. Retrieved June 22, 2010, from <http://www.nytimes.com>
- Hess, K., McDivitt, P., & Fincher, M. (2008, June). *Who are the 2% students and how do we design items and assessments that provide greater access for them? Results from a pilot study with Georgia students*. Paper presented at the CCSSO National Conference on Student Achievement, Orlando, FL. Retrieved June 16, 2010, from http://www.nciea.org/publications/CCSSO_KHPMMF08.pdf
- Individuals with Disabilities Education Act Amendments. (1997). Pub. L. No. 105-7 (1997) Retrieved January 21, 2010, from http://www.cec.sped.org/law_res/doc/law/index.php
- Individuals with Disabilities Education Improvement Act, Pub. L. No. 108-446. (2004). Retrieved January 21, 2010, from <http://www.ed.gov/policy/speced/guid/idea/idea2004.htm>
- Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 13, 2010, from <http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm>
- Kame'enui, E. J., Carnine, D. W., Dixon, R. C., Simmons, D. C., & Coyne, M. D. (2002). *Effective teaching strategies that accommodate diverse learners* (2nd Ed.). Upper Saddle River, NJ: Merrill Prentice Hall.
- Kettler, R. J. (2011). Holding modified assessments accountable: Applying a unified reliability and validity framework to the development and evaluation of AA-MASs. In M. Russell (Ed.), *Assessing Students in the Margins: Challenges, Strategies, and Techniques* (pp. 311–333). Charlotte, NC: Information Age Publishing.
- Kettler, R. J., & Elliott, S. N. (2009). Introduction to the special issue on alternate assessment based on modified academic achievement standards: New policy, new practices, and persistent challenges. *Peabody Journal of Education*, 84, 467–477.
- Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, 84, 529–551.
- Lazarus, S. S., & Thurlow, M. L. (2009). The changing landscape of alternate assessments based on modified academic achievement standards: An analysis of early adapters of AA-MASs. *Peabody Journal of Education*, 84, 496–510.
- Levitt, S. D., & Dubner, S. J. (2005). *Freakonomics: A rogue economist explores the hidden side of everything*. New York: HarperCollins.
- Marion, S. (2007). *A technical design and documentation workbook for assessments based on modified achievement standards*. Minneapolis, MN: National Center on Educational Outcomes. Retrieved April 2, 2010, from <http://www.cehd.umn.edu/nceo/Teleconferences/AAMASTeleconferences/AAMASworkbook.pdf>
- No Child Left Behind Act, Pub. L. No. 107-110. (2001). Retrieved January 21, 2010, from <http://www.ed.gov/policy/elsec/leg/eseas02/index/html>
- O'Sullivan, P. J., Ysseldyke, J. E., Christenson, S. L., & Thurlow, M. L. (1990). Mildly handicapped elementary students' opportunity to learn during reading instruction in mainstream and special education settings. *Reading Research Quarterly*, 25(2), 131–146.
- Palmer, P. W. (2009). State perspectives on implementing, or choosing not to implement, an alternate assessment based on modified academic achievement standards. *Peabody Journal of Education*, 84, 578–584.
- Pennsylvania Department of Education. (2010). *PSSA & PSSA-M Accommodations Guidelines for Students with IEPs and Students with 504 Plans* (Revised 1/11/2010). Retrieved June 17, 2010, from http://www.portal.state.pa.us/portal/server.pt/gateway/PTARGS_0_123031_744146_0_0_18/PSSA_Accommodations_Guidelines_2010.pdf
- Quenemoen, R. (2009, July). Identifying students and considering why and whether to assess them with an alternate assessment based on modified achievement standards. In M. Perie (Ed.), *Considerations for the alternate assessment based on modified achievement standards (AA-MAS)* (Chapter 2, pp. 17–50). Albany, NY: New York Comprehensive Center. Retrieved June 17, 2010, from http://nycomprehensivecenter.org/docs/AA_MAS.pdf
- Rodriguez, M. C. (2005). Three options are optimal for multiple choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24, 3–13.
- Samuels, C. (2007). Spec. ed. advocates wary of relaxing testing rules. *Education Week*, 26(43), 24–28.
- Schiller, E., Sanford, C., & Blackorby, J. (2008). *A national profile of the classroom experiences and academic performance of students with LD: A special topic report from the Special Education Elementary Longitudinal*. Menlo Park, CA: SRI International. Accessed on June 20, 2010. Available at http://www.seels.net/info_reports/national_profile_students_learning_disabilities.htm
- Skiba, R. J., Simmons, A. B., Ritter, S., Gbb, A. C., Rausch, M. K., Cuadrado, J., et al. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children*, 74, 264–288.
- Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction*, 4, 295–312.
- Title I—Improving the Academic Achievement of the Disadvantaged; Individuals with Disabilities Education (IDEA) Act; Assistance to States for the Education of Children with Disabilities. (2005). *Proposed Rule*, 70(Fed. Reg.), 74624–74638.
- Title I—Improving the Academic Achievement of the Disadvantaged; Individuals with Disabilities

- Education (IDEA) Act. (2007). *Final Rule*, 72(Fed. Reg.), 17748–17751 (to be codified at 34 C.F.R. pts. 200 and 300).
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 9, 2010, from <http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html>
- Thurlow, M. L. (2008). Assessment and instructional implications of the alternate assessment based on modified academic achievement standards (AA-MAS). *Journal of Disability Policy Studies*, 19, 132–139.
- Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (1998). *Testing students with disability: Practical strategies for complying with state and district requirements*. Thousand Oaks, CA: Corwin Press.
- US Department of Education. (2007). *Modified academic achievement standards: Non-regulatory guidance*. Washington, DC: Author. Retrieved February 14, 2010, from <http://www.ed.gov/policy/speced/guid/nclb/twopercent.doc>
- Wagner, M., Newman, L., Cameto, R., Levine, P., & Garza, N. (2006). *An overview of findings from wave 2 of the national longitudinal transition study-2 (NLTS2)*. Menlo Park, CA: SRI International. Retrieved June 20, 2010. Available at www.nlts2.org/reports/2006_08/nlts2_report_2006_08_complete.pdf
- Zigmond, N., & Kloo, A. (2009). The “two percent students:” Considerations and consequences of eligibility decisions. *Peabody Journal of Education*, 84(4), 478–495.
- Zigmond, N., & Matta, D. (2004). Value added of the special education teacher in secondary school cotaught classes. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Secondary interventions: Advances in learning and behavioral disabilities* (Vol. 17, pp. 57–78). Oxford, UK: Elsevier Science/JAL.

Accessible Tests of Student Achievement: Access and Innovations for Excellence

18

Stephen N. Elliott, Ryan J. Kettler, Peter A. Beddow,
and Alexander Kurz

Access means more than participation in general education classes and end-of-year achievement tests. Real access involves the opportunity for all students to learn the intended curriculum and demonstrate what they have achieved as a result of their learning. In an era of standards-based accountability, these goals of access are at the heart of high-quality instruction and testing practices.

As articulated by the authors of the 17 preceding chapters, access may start with policy and federal regulations, but to realize the benefits of educational access for students with disabilities many people must coordinate their efforts. Policy makers, educational leaders, Individualized Education Program (IEP) team members, individual teachers, test developers, and students themselves are all involved in advancing educational practices toward optimal access. Researchers also need to continue their press to develop tools that support teachers' efforts to facilitate all children's access to quality instruction and valid assessment of the knowledge and skills designated by the intended curriculum. Although promising methods and practices are emerging that can improve access for all students, more work is needed on a much larger scale, if it is to have a meaningful impact.

S.N. Elliott (✉)
Learning Sciences Institute, Arizona State University,
Tempe, AZ 85287, USA
e-mail: steve_elliott@asu.edu

We must move beyond the rhetoric of universal design principles and delve into the details of evidence-based teaching and testing practices, if educators and students are to realize the benefits of optimal access. Thus, in this concluding chapter of the *Handbook of Accessible Achievement Tests*, we take stock of what has been learned and where we are headed with regard to educational testing and access for students with disabilities.

What We Know

In many ways, this book was stimulated by amendments to the NCLB Act in 2007 that gave states the option of creating an alternative assessment based on modified academic achievement standards (AA-MAS) for eligible students with disabilities. The central theme of meaningful access to achievement tests, however, goes beyond students with disabilities and a particular testing option for accountability purposes. If you have read the majority of the chapters in this book, you know there are a number of issues or barriers to achieving meaningful access to both instruction and tests. You have also learned about some research insights and innovations related to classroom instruction and test design that have the potential to overcome critical barriers to optimal access. We briefly recount key issues and innovations relative to policy and regulations, classroom practices, and test design with the goal of providing an integrative summary that can focus our thinking about needed future actions.

Policy and Regulations About Access

Maximizing students' opportunity to learn (OTL) valued knowledge and skills for a population of students with a diverse range of achievement and ability levels represents one of the main goals of educational policies and regulations about access. Improving access to valid assessment is the typical strategy for pursuing this goal, because (a) improved OTL should result in achievement gains, (b) established measures of achievement gains are more available than are established measures of OTL, (c) assessment results can be used to identify areas where instruction needs to be improved, and (d) the promise of assessment results from which valid inferences can be drawn, coupled with consequences for failure to meet proficiency standards, increases the motivation to successfully teach the knowledge and skills that compose the intended curriculum. In [Chapter 4](#), Zigmond, Kloo, and Lemons provided a sequential model of this relationship, detailing the steps from the participation of students with disabilities in high-stakes assessments, as mandated by the Individuals with Disabilities Education Act (IDEA, 1997), to improved academic outcomes for this group. The authors warned that the system could be undermined by options that set lower standards for students who are capable of achieving at grade level. In [Chapter 2](#), Weigert also identified the delicate balance that alternate assessment policies have had to maintain, between holding students with disabilities to such high proficiency standards that the resulting instruction cannot be meaningfully accessed, and having such low expectations that valid assessment is no longer a lever for increasing the quality of academic instruction. Of course, while IEP teams must make categorical decisions about which students are eligible for each assessment option (general assessment with or without accommodations, modified assessment with or without accommodations, alternate assessment of alternate academic achievement standards [AA-AAS]), in reality the best option is determined based on each individual's achievement along a continuum and her or his unique profile of

strengths and weaknesses; these considerations taken together do not nicely divide students into categories aligned with assessment options. While students who should complete the general assessment and those who should complete the AA-AAS are non-overlapping groups, the lines between who should complete an AA-MAS and either the general assessment or the AA-AAS are much fuzzier. The situation gets even more complicated when considering the possibility of using accommodations deemed appropriate on an individual basis.

In [Chapter 3](#), Phillips outlined a set of guidelines, based on legal requirements and psychometric principles, for states to develop nonstandard test administration policies for students with disabilities and English language learners (ELLs). The guiding psychometric principle of the author's recommendations was preservation of the construct being targeted by the test. Phillips suggested defining accommodations as nonstandard test administrations that do not change the target construct, and defining modifications as nonstandard test administrations that do change the target construct. Consistent with this suggestion, the author recommends that scores from tests that have only been altered by the use of accommodations be aggregated with and interpreted alongside scores from a general assessment, but that scores from a modified administration be reported separately unless the tests have been statistically linked. Under such a system, AA-MASs that are not linked to the general assessment would provide students with disabilities (SWDs) greater opportunity to experience success, but would not necessarily provide greater access. Zigmond and colleagues addressed the selection of the appropriate assessment on the individual level, emphasizing the requirement that IEP teams select the most appropriate option at the time of the evaluation. Given that the most appropriate assessment option is the one that provides the score from which the most valid inferences can be drawn, this suggestion is in agreement with the psychometric commitment to preserve the construct being measured.

In [Chapter 17](#), Lemons, Kloo, and Zigmond drew similar conclusions to those of Phillips

about the issue of considering scores derived from AA-MASs as comparable to, and combinable with, scores from a general assessment. The authors dismiss the notions that the AA-MAS policy will lead to better measurement of the knowledge and skills of students with disabilities, or that the test will help many schools and districts meet annual yearly progress (AYP) goals. Their grounds for questioning improved measurement are that reducing breadth and depth would change the construct, and that changes which strictly affect access could be applied to the general test. Lemons and colleagues did not address the possibility that an easier set of items might provide more precise measurement for a group of students that is certain to achieve at a lower level (i.e., students eligible for an AA-MAS). They acknowledged that the policy might be a more ethical solution and might lead to improved instruction for a subset of students who would otherwise take the AA-AAS, but ultimately contended that developing an AA-MAS is not worth the risk when one considers the cost of production and possibility of unintended consequences (e.g., instructional tracking). Echoing Phillips' concerns, Lemons and colleagues indicated that comparing AA-MAS scores with scores from the general assessment to determine proficiency is just "pretending," and suggested that alterations to tests that solely address access be applied to the general examination, and that special education move away from the danger of uniform expectations that are too high to meaningfully inform instruction.

Over the course of 40 years of the US assessment policy, we are slowly approaching a system that maximizes access to measurement of knowledge and skills for all students. In [Chapter 5](#), Davies and Dempsey presented an interesting comparison with Australian policy on achievement accountability, demonstrating that laws there are less clear about the inclusion of students with disabilities. While policy in Australia indicates that all students should be included in national achievement assessment, a large number of these students are still exempted or may be absent during testing due to disabilities. Until Australia adopts practical solutions analogous to

those used to include students with disabilities in the United States, those students will not benefit from the improvement in instruction that is associated with appropriate assessment.

The purpose of improving education through better measurement is only served when we prioritize a goal of access, which in turn leads to better measurement precision among the population for which each test is intended; much work in this area, evaluating the reliability and validity of tests modified to increase access, remains to be done. The success of any large-scale accountability system will be ultimately measured based on its consequences for improving classroom instruction and increasing access to the general curriculum for all students – the topic of the next section.

Classroom Instruction and Access to the General Curriculum

The assessed curriculum of any large-scale accountability system clearly exerts a strong influence on the content of classroom instruction and the extent to which students receive access to the general curriculum (Cizek, 2001; Ysseldyke et al., 2004). In [Chapter 6](#), Kurz discussed a comprehensive framework that delineates the connections between the various curricula of the educational environment at the system, teacher, and student level. The framework posits the intended curriculum (i.e., the content designated by the academic standards of the general curriculum) as the normatively desirable target of all other (subordinate) curricula. For students with disabilities, each Individualized Education Program delineates the extent to which the general curriculum is part of the student's intended curriculum. That is, the IEP curriculum may designate the entire general grade-level curriculum as the student's intended curriculum or specify some modified or alternate objectives as well as additional social, behavioral, or functional goals. Thus, Kurz presented the intended curriculum for students with disabilities as dually determined by both the general curriculum and the student's IEP curriculum.

The framework developed by Kurz indicates that the intended curriculum directly informs the content of the assessed curriculum used for accountability purposes. This model is consistent with the current accountability system, which mandates that the tested content of large-scale assessment be used to sample exclusively across the content domains of the intended curriculum. Only on the basis of such curricular alignment are stakeholders able to draw valid inferences about the extent to which students have been exposed to, and subsequently achieved, the intended curriculum. Moreover, stakeholders who wish to draw valid inferences about student achievement that has occurred as a function of classroom instruction must also account for students' opportunity to learn the intended curriculum. The chapters by Kurz (Chapter 6), Ketterlin-Geller and Jamgochian (Chapter 7), and Beddow, Kurz, and Frey (Chapter 9) specifically addressed two major *access points* related to the intended curriculum: (a) the enacted curriculum and (b) the assessed curriculum. With regard to the enacted curriculum, the content of classroom instruction presents the primary means through which students access the intended curriculum. With regard to the assessed curriculum, accessible tests are necessary for students to fully demonstrate achievement on the tested constructs of the intended curriculum. All authors emphasized that without addressing *both* barriers, it is virtually impossible to make valid test score interpretations. To this end, Ketterlin-Geller and Jamgochian provided two clear examples. In the first example, students who were given the *opportunity to learn* the intended curriculum were not able to fully demonstrate what they knew and were able to do, because the assessments were inaccessible. Poor test performance could thus be an artifact of inaccessible tests rather than a lack of progress in learning the intended curriculum. In the second example, poor test performance obtained through *accessible tests* was difficult to interpret, because students were not provided with the opportunity to learn the intended curriculum in the first place.

OTL, of course, is not a dichotomous reality at the enacted curriculum level. Kurz described OTL as a matter of degree related to three instructional dimensions: time of instruction,

content of instruction, and quality of instruction. These key dimensions of classroom instruction have a strong and long-standing research base as correlates of student achievement. Kurz discussed the challenges and considerations related to measuring this operational conceptualization of OTL in detail in Chapter 6. As a future direction of research on OTL, he advocated considering the formative benefits for teachers' classroom instruction that could result from integrating OTL measurement into ongoing instructional processes.

Ketterlin-Geller and Jamgochian focused on the content dimension of classroom instruction by examining the instructional adaptations that may be necessary for students with disabilities to fully access the intended curriculum through the teachers' classroom instruction. The authors hereby differentiated between *instructional accommodations* and *instructional modifications*. Instructional accommodations are adaptations to the design or delivery of instruction (including materials) that do not change the breadth of content coverage and depth of knowledge of the general grade-level curriculum. Examples include large fonts or Braille for students with visual impairments (i.e., presentation accommodation). Instructional modifications are adaptations to the design and delivery of instruction (including materials) that result in a reduction of the breadth of content coverage and/or depth of knowledge of the general grade-level curriculum. Examples include extended time or frequent breaks (on an otherwise timed task) for students with attention and/or processing deficits (i.e., timing/scheduling modification). The authors emphasized that an important consequence of instructional modifications is a reduction of students' opportunity to learn grade-level content standards and/or the level of proficiency expected of the general population. Due to the impact of instructional modifications on OTL, Ketterlin-Geller and Jamgochian argued that their use should be limited in favor of instructional accommodations, which maintain access to the general curriculum at grade level.

Many of the instructional accommodations suggested by Ketterlin-Geller and Jamgochian

provide students with access skills that increase students' benefit from an opportunity to learn the intended instructional constructs (e.g., highlighting of key words to focus attention on critical information). Kettler, Braden, and Beddow discussed *test-wiseness* as an access skill in the context of achievement testing. Test-wiseness can affect the extent to which students are capable of utilizing test-taking skills or strategies in a test-taking situation to receive a higher score. Typically, test-wiseness does not represent an intended construct of achievement tests. Yet, many test-preparation programs include test-taking skills or strategies as part of their curriculum. Kettler et al., however, indicated that research does not support the extended use of classroom time and instruction for purposes of teaching test-taking skills. Instead, teachers should focus their instructional time on the content of the intended curriculum, which is subsequently sampled by properly aligned achievement tests.

As noted by the authors of chapters in the *Classroom Connections* section, considerations for accessible achievement testing should not be restricted to the test event only. Accessible achievement testing and accessible instruction represent two sides of the same coin. That is, all students should receive the opportunity to learn the intended curriculum through classroom instruction that is well managed in terms of time, content, and quality, as well as maximally accessible through judicious use of differentiated instruction and/or instructional adaptations. The resulting test scores derived from accessible achievement tests thus should permit more valid inferences about what students know and are able to do as a function of classroom instruction. Available strategies for purposeful test design to allow test-takers the opportunity to demonstrate their knowledge to the greatest extent possible are discussed next.

Test Design That Supports Access

Commensurate with the importance of access to the general curriculum, as addressed by Kurz and others, is the need for ensuring that assessments

of this curriculum yield data that reflect actual student achievement. Beddow, Kurz, and Frey (Chapter 9) offered a theoretical model (i.e., accessibility theory) that defines the variety of potential sources of construct-irrelevant variance based on test accessibility. In essence, the authors contended that the accessibility of an assessment is a function of test features as well as individual test-taker characteristics. Specifically, Beddow et al. operationalized accessibility as the sum of interactions between these variables. An optimally accessible test, therefore, yields scores from which subsequent inferences reflect only the interaction between the test and the target construct as demonstrated by the test-taker, controlling the influences of access skills. The authors present an example of a high school science item in three stages of modification to illustrate several aspects of test items that may be considered in the effort to reduce access barriers and increase the validity of subsequent inferences for more test-takers. Beddow et al. recommended test developers subject tests to accessibility reviews prior to administering them with actual test-takers for their intended purposes. They argued reviewers should ground these reviews in empirical data, disaggregated by high and low performers, disability status, and other group categories, whenever possible.

In Chapter 10, Tindal and Anderson discussed the types of evidence needed to ensure test score inferential validity, particularly for assessments designed for students with special needs. The authors reviewed several studies on response processes and concluded extended time and read aloud can reduce potential access barriers and increase validity. They also noted a number of shortcomings in many efforts to provide validity evidence and made several suggestions for test developers to improve research and practice. These included increasing the precision with which constructs are defined, operationalizing test design and implementation procedures, providing deliberate training for test administrators/teachers on the science and practice of assessment, attending to the practical organization and implementation of tests to ensure external validity of research, and

examining test consequences programmatically to support the iterative increase of scientific evidence.

Rodriguez added to the test design discussion by surveying the extant knowledge on test item–writing practices. In [Chapter 11](#), Rodriguez reviewed a broad base of research and theory spanning more than a century, providing several considerations for designing tests and test items to maximize measurement utility. He examined the range of common item formats and evaluated the potential benefits and limitations of a number of innovative formats that are made possible by increases in computer technology. Drawing on decades of test development guidelines, Rodriguez offered several recommendations for selecting the proper item format for the intended purpose, and for writing good test items. These include content and formatting concerns, such as keeping vocabulary and grammatical construction as simple as possible, maintaining singularity in target constructs, and formatting items consistently. He also described strategies that should be avoided by item writers, including using negative stems, writing trick items, and including overly-specific or trivial content. Rodriguez focused specifically on the options for multiple-choice items, contending that it is in this area that empirical evidence is most conclusive. Based on this evidence, he asserted three choices are optimal for multiple-choice items. Rodriguez concluded the chapter by discussing the relation between item development and accessibility and its relevance to inclusive assessment.

Another critical aspect of test accessibility is the degree to which language issues may influence the validity of test score inferences. In [Chapter 12](#), Abedi addressed linguistic and cultural issues and the problem of reducing access barriers for students with a broad range of abilities and needs. He specifically focused on the performance gaps between English language learners and non-ELL students. Abedi noted these gaps are observed not only in reading, but in science and mathematics as well. In content areas in which language and culture are not the target constructs, he argued, the concept of linguistic and cultural accessibility is essential. He

proposed employing teams of experts in assessing ELLs to address the linguistic complexity of assessments prior to using them to make decisions on behalf of students. To determine the linguistic complexity of an assessment and its resultant accessibility for ELLs, Abedi contended these experts should attend to issues such as word frequency, familiarity, and length; grammatical structure; and concrete versus abstract narrative.

Kettler, in [Chapter 13](#), reviewed the limited extant data on the effects of using packages of item modifications to increase test accessibility. The author drew several conclusions about this emerging field of research. Kettler indicated that while examining packages of modifications in validation studies necessarily limits the degree to which these studies yield empirical knowledge about the effects of individual item enhancements, modification typically is provided in packages in practice. The author further contended that the measurement properties of each original assessment should be a “gold standard” by which the measurement properties of modified assessments with the intended populations are evaluated. Indeed, Kettler argued while it is expected item difficulty typically decreases following modification, reliability is primary among indicators of the effects of test changes on the measurement of the target construct. The author suggested access barriers likely diminish the consistency with which a test measures the target skills across the test-taker population; thus, if reliability is not maintained or increased, little else can support the use of a modified test over the general assessment as a better measure of what students know and can do. Further, Kettler suggested decreased reading load likely is helpful for the portion of the test-taker population for whom reading may present access barriers. This may be manifest in shorter stems, fewer words in visuals, and fewer options for multiple-choice items. Lastly, Kettler noted that while data indicate reliability can be maintained in reading across grade levels, and in mathematics and science at the elementary level, reliability may decrease when modifications are applied to mathematics and science items at the middle and high school grade bands; the cause of this pattern is yet unknown.

In [Chapter 14](#), Roach and Beddow discussed how student input can be incorporated in the design of assessments for students with a broad range of abilities and needs. These authors examined students' drawings, focus group interviews, think-aloud cognitive labs, and survey data to understand how student input may influence the development of accessible assessments. The authors contended that soliciting student opinions and perspectives during the design phase is important, but students often may not necessarily comprehend the degree to which item or test features may impact their performance on the test. Notwithstanding, ample evidence suggests (a) students often experience anxiety when presented with an assessment task, (b) students report assessments often are more complex than they would like, (c) students, particularly those identified with disabilities, often feel they have not had sufficient opportunity to learn the tested content prior to being required to demonstrate their proficiency, and (d) students tend to prefer test items that have been modified to improve their accessibility over test items in their original form.

Russell in [Chapter 15](#) addressed the opportunities computers offer for enhancing test accessibility across the range of test-takers. He invoked Bennett's (1999) prediction that the inevitable ubiquity of computer-based testing would permit its increased integration with instruction. Russell observed, however, that in the decade since Bennett's conjecture, computers primarily have been used to increase the efficiency of testing as opposed to its integration with instruction, thus missing the mark insofar as the potential of using computer assessments to maximize learning across the educational environment. While computers can be used solely for the purpose of enhancing the efficiency of data collection, they also can be used to reduce barriers for students for whom current assessments may not be optimally accessible. Russell surveyed the current state of computer-based assessment technology, specifically in terms of addressing the individual access needs of test-takers through integrated accommodations and access tools. He described innovative means of presenting audio content for poor

readers or visually impaired test-takers, signed avatars for deaf test-takers, adapted response modes for test-takers for whom the traditional method of responding is inappropriate, and alternate representations for test-takers whose access needs require them (such as tactile representations of visuals). Russell concluded with a call for researchers to examine ways of helping teachers and test administrators decide which tools should be used for individual test-takers to ensure all students are given access to demonstrate what they know and can do.

In [Chapter 16](#), Egan, Schneider, and Ferrara delved deep into standard setting, a critical step in the assessment development process, and ultimately in the interpretation of achievement test results. Given the goal of accessible assessment to ensure the measurement of achievement is equally accurate and consistent across the range of test-takers, the process used by test developers to set proficiency levels must be equally applicable to all test-takers as well. The authors described a standard-setting framework in six phases: Define, Describe, Design, Deploy, Deliver, and Deconstruct. Mirroring the clarion call in other chapters for precisely defined target constructs for tests and test items, Egan et al. asserted the establishment of clear, precise definitions of achievement-level descriptors is integral to test development, cut-score recommendation, and test score interpretation. The authors contend validity evidence should be collected throughout the standard-setting process, particularly when developing tests that are intended for use with students with special needs.

Where Are We Going

Two years have passed since we conceptualized this book on accessible achievement tests. Some new research has been published on item modifications; several federally funded grants are underway to examine innovative OTL measurement and explore attributes of more accessible tests; some test developers are using accessibility tools to guide the development of new items; and at least seven more states have opted to implement

AA-MASs. Yet at the same time, testing practices in the United States and in some foreign countries, such as Australia, continue to be influenced by dated standards and limited conceptualizations of item modifications and universal design principles for tests. Policies, practices, and procedures around instruction and testing seem to change too slowly for the thousands of student test-takers who continue to experience difficulty gaining access to both the intended and assessed curricula. There are, of course, challenges confronting those of us interested in improving instruction and testing for all students. Several of the central challenges are discussed next, followed closely by needed innovations that have the potential to advance accessible instructional and testing practices.

Challenges to Access

Barriers that limit access to the intended and assessed curricula are real. This assertion is not new and has been part of the vision of education policy makers, test developers, and many teachers for more than a decade. Based on the collective examination of the scholars and leaders contributing to this book, we believe there are five central challenges that educators and test developers must address if they are to improve access to valued content and knowledge for all students:

1. Ensuring students have meaningful opportunities to learn the grade-level intended and assessed curricula.
2. Precisely articulating target constructs of tests and test items to facilitate the development of maximally accessible items.
3. Creating highly accessible tests that yield reliable scores and valid inferences about students' achievement in language arts, mathematics, and science.
4. Determining the consequences of participation in any of a number of assessment options for a diverse population of learners.
5. Moving beyond the existing proficiency testing paradigm to one that privileges access and progress to grade-level academic content.

These challenges are not unique to services for students with disabilities, although they may be more salient in their educational lives and that of their teachers and parents. Access to high-quality instruction and assessment is a commitment extended to all students and must mean more than participation, testing accommodations, and tests refined with universal design principles in mind.

Needed Innovations to Improve Access

The words *access* and *excellence* often are used in the same sentence by educational leaders as if they are two different goals. In fact, we contend, as do the majority of our co-authors in this book, that access is a prerequisite for achieving excellence in teaching, learning, and testing. Indeed, excellence depends on access! As documented in this book, the knowledge base concerning educational access has grown substantially over the past several years and in the process new ways of thinking and tools for taking action have emerged for both researchers and practitioners. Collectively, the knowledge and the related tools examined in this book provide us with several innovations that can be used to address the central challenges to access and help educators progress toward the goal of excellence for all learners.

To ensure students have meaningful opportunities to learn intended and assessed curricula at grade level, we believe teachers need more support and feedback concerning their state's content standards. The work by Kurz and colleagues on OTL and the development of practical, teacher-oriented tools such as MyiLOGS have the potential to advance access by providing teachers a framework for organizing instructional content, documenting instructional time, and stimulating evidence-based instructional decisions for entire classes and individual students. Such data-based tools must take little time, yet provide teachers with powerful and ongoing feedback about their instructional actions. The information obtained about OTL from these tools helps to contextualize achievement data, potentially resulting

in instructional changes that increase access to excellence for all students. Professional development for teachers that features instructional content and resources for activating key concepts or procedures associated with the content, coupled with tools for managing instructional time and student outcomes, will certainly improve students' opportunity to learn the intended and assessed curriculum.

Precisely defining the target constructs of tests and test items is a long-standing challenge confronting test developers and researchers interested in test score validity. Precision in measurement of student achievement requires attention to many details; often these details are influenced by item writers and sometimes by educators responsible for defining the content to be assessed. No innovation discussed in this book will fully meet this challenge; however, meaningful improvement is possible. We believe that by providing item writers who have deep knowledge of a content domain with explicit training in the use of item-writing tools like the Test Accessibility and Modification Inventory (TAMI) and its diagnostic complement, the Accessibility Rating Matrix (ARM), that much of the extraneous content of items can be reduced. Extraneous item content often leads to construct-irrelevant variance. Thus, supporting item development that values accessibility can be achieved today. Coupled with this type of training, states and other organizations that oversee achievement tests are encouraged to implement accessibility review panels. Such panels should have representatives knowledgeable of students with disabilities and also content experts. These panels could function much like bias review panels and focus on eliminating item attributes and content that are extraneous and irrelevant to the constructs being measured.

The challenge of creating highly accessible tests that yield reliable scores and valid inferences about students' achievement begins with the development of well-written items without extraneous content. For some students, accessible tests that yield reliable and valid scores, however, require more than accessible items. Many students with disabilities or English

language learners need accommodations to be able to meaningfully access items and respond to questions. Testing accommodations provide more complete access to questions and answers and reduce irrelevant variance in responses caused by a disability or language difference. The provision of needed and appropriate testing accommodations with integrity for all qualified students is achievable and will make a difference in the accessibility of tests, and, for many students, will result in higher test scores. When tests are delivered via computers, the potential for delivering needed accommodations with high integrity and consistency is even greater. Continued innovation of computerized item management and delivery of tests promises to improve access to needed accommodations for more students.

Tests have instructional consequences. Determining the consequences of participating in an assessment for students is important to understand and has been challenging to document. Ideally, assessment results are used to refine instruction for each student. Unfortunately, when results of tests are provided two to three months after the test is completed, the results are very unlikely to be used with the group of students who took the test. This disconnection between assessment and instruction can no longer be accepted if we want to improve the consequences of assessment. Computerized testing and more frequent short tests with nearly immediate feedback offer a solution that can lead to improved access to appropriate instruction. Tests also have motivational consequences. That is, repeated testing experiences where a student is confronted with test content that has not been taught or is not readily accessible can serve to decrease one's motivation to engage in testing, because it seems futile and unfair. More accessible – accommodated and modified – tests hold promise for improving the emotional or attitudinal side effects of testing and also yield scores with more validity.

Moving beyond the existing proficiency testing paradigm to one that privileges access and progress to grade-level academic content has been challenging for most states and test

developers. As documented in this book, a number of instructional and testing innovations are available to improve access for all students. Some of these innovations are being taken to scale in five or six states, but a more pervasive and sustained effort is needed nationwide to have a meaningful impact. In addition, more state assessment systems would benefit from the use of tests that allow for achievement growth to be recorded. Collectively, we know how to build more accessible tests and assessment systems that yield academic growth information. For a number of reasons we have not accomplished all that is possible when it comes to accessible tests and accessible testing practices.

Conclusion

It is time to stop admiring the challenges to accessible tests and think about innovative solutions that will improve access and lead to excellence in education for all students. We have written this handbook to help educators, test developers, and researchers understand the importance of access and its role in providing excellent instruction and testing practices. We hope this volume stimulates expedient actions that lead to the invention of more solutions, because today too many students sit in classrooms where access to the intended and assessed curriculum is the exception and not the norm.

Subject Index

A

- AA-AAS, *see* Alternate assessment based on alternate achievement standards (AA-AAS)
- AA-MAS, *see* Alternate assessment based on modified academic achievement standards (AA-MAS)
- Academic Competence Evaluation Scales (ACES), 190
- Academic grading system, 83
- Acceptability, 37, 239, 246–247, 253
- Access, 44
- alternate language, 266
 - audio, 260–261
 - barriers/challenges, 5, 147–148, 165, 172, 326
 - and cut score setting, 275–276
 - definition, 3
 - to general curriculum, 321–323
 - innovations to improve, 326–328
 - to instruction and tests of achievement, 1–2, 131–132, 137–139, 158, 319
 - to intended curriculum, 100
 - linguistic, 217–218
 - pathway, 179–180
 - policy and regulations, 320–321
 - as policy tools, 202
 - signed, 263–266
 - tactile, 266–267
 - test-taking skills and, 153
 - and test-wiseness, 147–148
 - through OTL, 5–8
 - through testing accommodations, 8–9
 - through well-designed test items, 10–11, 166–171, 231, 323–325
 - v. success, 44–45
- Accessibility
- across educational environment, 179–180
 - and construct preservation/score comparability, 38–39
 - definition, 2–3
 - item modifications for, 210
 - levels, 168
 - limited, 201–202
 - linguistic, concept of, 217–220, 223–224
 - means for increasing, 279
 - modification packages on test and item, effects of, 231
 - proof paradox, 177–178
 - test-taking skills and, 147–148
- Accessibility Principles for Reading Assessments*, 213
- Accessibility Rating Matrix (ARM), TAMI, 11, 168–169, 171–176, 179, 327
- Accessibility theory, 2–3, 10–11, 163–165
- accessibility proof paradox, 177–178
 - accessibility rating, 174–177
 - accessible test items, 166–171
 - characteristics, 169
 - TAMI/TAMI ARM, 166, 168
- CLT
- categories, 166
 - long-term/working memory, concept of, 166
 - principles of, 166–168
- and educational environment, 179–180
- test-takers, 164–165, 190
- ARC, 165
 - barriers, 165
 - interaction domains/categories, 164
 - UD, 163–164
- Accommodation(s)
- access through testing, 8–9
 - categories, 258–260
 - definition, 4, 38
 - for ELL, 60, 219, 223–224
 - instructional, 13, 131–136, 138–142, 144, 180, 257, 322–323
 - read aloud, 140, 186–187, 189, 195–196, 256–261, 263
 - “standard”, 23
 - test, 8, 64, 185–186, 193, 211, 257–258, 260, 269–271
 - test-taking skills and, 153–154
- Accommodations Survey (Terra Nova 18), 190
- Accountability systems, 5, 12, 24, 59, 62, 69–71, 80, 107–108, 113, 153, 243, 295, 297, 311–315, 321–322
- Accountability testing challenges, U.S. legal issues, 59–63
- California case, 61–62
 - construct shift, 61
 - ELL accommodations/modifications, 60
 - majority/minority ELL, 60–61
 - NCLB Act, 59–60
 - NCLB cases, 62–63
 - Coachella Valley, 62–63
 - Reading School District, 62

- ACES, *see* Academic Competence Evaluation Scales (ACES)
- Achievement
 alternate, standards, 27–28
 IEA, 110
 instruction and tests of, 1–2
 levels, 20, 30, 64, 71, 93, 193, 275–276, 278, 280–282, 284, 286–287, 289–290, 325
 modified, standards/tests, 29–30, 71, 73, 77–79, 106, 197, 232–235, 239–240, 295–296
See also Alternate assessment based on modified academic achievement standards (AA-MAS)
 student, 5–8, 12, 27, 88–89, 94, 99–100, 103, 107–113, 115–116, 153, 278, 282, 296, 319, 322–323, 327
- ADA, *see* Americans with Disabilities Act (ADA)
- Adapted interactions, 258
- Adapted response modes, 258, 325
- Adelaide Declaration, 87–88
- Adequate Yearly Progress (AYP), 26, 29, 59–60, 137, 183, 295–297, 306–311, 314, 321
- Advanced Placement exam, 207
- Advocates for Special Kids v. Oregon Dep't of Educ.*, 45
- AERA, *see* American Educational Research Association (AERA)
- Alignment index (AI), 102, 110, 121
- Alignment methodologies, 7
 constraints, 107
 SEC method, 102, 111
 Webb method, 102
- All-of-the-above (AOTA) items, 205, 208–209
- Alternate achievement standards, 27–28, 70–71, 73–74, 78, 80, 94, 313
- Alternate assessment based on alternate achievement standards (AA-AAS), 27–28, 71, 73–75, 79, 94, 105, 137, 210–211, 231–232, 275–278, 280–282, 284–285, 287, 290, 296–297, 309, 313, 320–321
- Alternate assessment based on modified academic achievement standards (AA-MAS), 3–4, 29–30, 71, 73–75, 77–79, 93, 105, 137, 183, 185–187, 210–211, 231–236, 238–240, 244, 275–282, 284–285, 287, 290, 295–298, 302–303, 305–308–314, 319–321, 326
- Alternate Assessment Standards for Students with the most Significant Cognitive Disabilities, 73
- Alternate-choice format, 203, 205
- Alternate contrasts accommodation, 259, 268–269
- Alternate scoring methods, 214
- Alternative assessment, 23–24, 93–94, 319
- Ambach case, 36, 52–53
- American Educational Research Association (AERA), 3–4, 37, 183–185, 226, 232
- American Psychological Association (APA), 3–4, 33, 37, 103, 183–185, 187, 210, 226, 232, 245–246, 253
- American Sign Language (ASL), 263, 266
- Americans with Disabilities Act (ADA), 23, 35–36, 100, 164
- Amplification devices, 139–141
- Analytic rubrics, 168, 174, 220
- Ancillary requisite constructs (ARC), 165, 172, 175, 178
- Anderson v. Banks*, 36
- Annual accountability assessment, primary participation decisions, 76, 78
- APA, *see* American Psychological Association (APA)
- Applied Measurement in Education*, 204
- Approaching Expectations, 281–282
- Approximation techniques, 39
- ARC, *see* Ancillary requisite constructs (ARC)
- Arizona Instrument to Measure Standards, 238
- ARM, *see* Accessibility Rating Matrix (ARM), TAMI
- ARM Record Form, 174
- ARM rubrics, 168, 174
- Assessed curriculum, 12, 99–100, 102–106, 110, 113, 121, 179–180, 321–322, 327–328
- Assessed skill levels, 118–119
- Assessment
 accommodations, 23, 153–154, 158
 instruments, 10, 25, 164, 166, 177, 227, 244
 policies, 1960s and 1970s, 19–20
 modifications, 153
- Assistive listening devices, 91
- Assistive technologies, 2, 25, 27–28, 141, 213, 258
- Assistive Technology Act, 2, 164
- Attention deficit hyperactivity disorder, 85
- Auditory accommodations, 139–142, 257–258, 260–263, 270–271
- Australia, 83–94
See also specific entries
- Australian Curriculum, Assessment and Reporting Authority (ACARA), 88, 94
- Australian Federal Court, 84–85, 87
- Australian Government Productivity Commission, 84–85
- Australian Human Rights Commission, 85–87
- Australian policies, inclusive assessments and, 83–84
 compared with U.S. policies, 87, 93–94
 DDA legislation, 84–87
 historical development of, 85–87
 educational system, 83–84
 enrolment options for SWD, 83–84
 two tiered, 83
- NAPLAN, 88–90
 comprehensive assessment process, 88–89
 exemptions/absences/withdrawals and related issues, 90–93
 participation and eligibility, 90
 performance scores/statements, 89–90
 special provisions/considerations and related issues, 91–94
 test administration authority, role of, 89
 types of test questions/formats, 89
 national goals for schooling, 87–88
- Australian public education systems, 85
- Australian student exemptions, absences/withdrawals, 92
- “Authentic” assessments, 22
- Automated scoring, 204

- Avatars, signing, 165, 264–266, 325
 AYP benefits, 309–310
- B**
Bd. of Educ. of Northport v. Ambach, 36, 52
 Beddow model, 164–165
 Biases, 194–195, 218–219, 225–226, 228, 233–234, 252, 288, 327
 Binet-Simon Scale, 201
 Bloom's taxonomy, 134
 Braille accommodations, 8, 23, 35, 41, 44, 48, 54, 91, 139–140, 256–257, 259, 266–267, 322
Brookhart v. Illinois State Bd. of Educ., 35–36, 51–52
Brown v. Board of Education, 19, 295
 Bureau of Accountability and Assessment (BAA; Pennsylvania), 305
 Bureau of Special Education (BSE; Pennsylvania), 305
- C**
 CAAVES, *see* Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES)
 Calculators, 7, 28, 33, 39–40, 42–43, 45–46, 53, 55, 65, 112, 136, 139, 142, 144–145, 212, 233
 California case
 accountability testing challenges, 61–62
 graduation testing challenges in SWD, 54–56
 alternate assessments, 55
 graduation test, 54–55
 legislative intervention and settlement, 56
 waiver policy, 55–56
 California High School Exit Examination (CAH SEE), 54, 58, 256
 California Supreme Court, 58
 Carroll model, 7, 108–109, 112
 CBM, *see* Curriculum-based measurement (CBM)
 CEC, *see* Council for Exceptional Children (CEC)
 CEC testimony, 22
 Chain gameplay, 156
Chapman (Kidd) v. State Bd. of Educ., 53–54, 256
 Chicago Public Schools, 250
 Child x Instruction interactions, 118
 Civil Rights Act, 19
 Classical measurement theory, 218, 225
 Classical test theory, 225, 238
 Classroom Assessment Scoring System (CLASS), 118–119
 Classroom-based assessment practices, 137–138, 142, 257
 Classroom instruction
 accommodation/modification adaptations, 131–145
 OTL and ICM, 99–125
 test-taking skills, 147–158
 “Clickers”, *see* Electronic input devices
 Closed-product formats, 206
 Closed scoring keys, 206
 CLT, *see* Cognitive load theory (CLT)
 CMAADI, *see* Consortium for Modified Alternate Assessment Development and Implementation (CMAADI)
- Coachella Valley v. California*, 62–63
 Coaching programs, 150, 152
 Coefficient alphas, 203, 232–239
 Cognitive complexity, 134–135, 234, 277, 284, 312
 Cognitive continuum, 206
 Cognitive demands, types of, 7, 10–11, 102–103, 110, 112, 115, 120–122, 166–167, 173–175, 177, 251, 312
 Cognitive load theory (CLT), 5, 10–11, 202, 211, 213, 252, 311
 categories, 166
 item stimulus, 171
 long-term/working memory, concept of, 166
 principles of, 166–168, 178–179
 Cognitive overload, 143, 155, 167, 174
 Cognitive process dimensions, 123–124
 College Board, 201
 Committee on Goals 2000, 93
 Commonwealth and State and Territory governments, 85–86
 Communication devices, 91, 139, 142, 258, 271
 Comparable scores, 33–34, 37–38, 40, 44–46, 55, 64, 66
 Complex multiple-choice (Type K) format, 203
 Computer-adaptive assessments, 25, 212
 Computer-based tests (CBT), 147
 access tools, recommended, 155
 advantages/disadvantages, traditional methods, 154–155
 suggestions for developers, 155–156
 video games and, 156–157
 Computer-enabled innovations, 211–212
 Computerized tests and individual needs, 255–257, 327
 accommodation representations/categories, 258–260
 alternate language based, 266
 audio based, 260–263
 sign based, 263–266
 tactile based, 266–267
 test validity, 269–270
 accommodations, rethinking test, 257–258
 adapted presentations, 267–269
 alternate contrasts, 268
 magnification, 267–268
 masking, 268–269
 prospects, 270–272
 Computer spell-check, 45
 Consensus-building activities, 286
 Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES), 166, 233, 235–238, 240, 250–251
 Consortium for Modified Alternate Assessment Development and Implementation (CMAADI), 166, 233, 236–239
 Constructed response (CR) items, 148, 168, 197, 202–204, 206–208, 212, 233, 235, 248, 284–285
 Construct-irrelevant variance, 8, 147, 157–158, 179, 211–212, 218–219, 226, 228, 323, 327
 Construct preservation, 34, 38–39, 66

- Construct-relevant factors, 30, 37, 40, 56–57, 60–61, 64, 147, 155, 214
- Construct validity, 37, 119–120, 147–148, 185, 194–196, 246
- Content
 - coverage, 134–137, 139, 142–143, 179, 322
 - emphasis, 102, 113
 - exposure, 112–113
 - map, SEC, 121–122
 - overlap, 7, 108, 110–111, 121
 - and quality, 7–8, 113–114, 118, 121, 323
 - related evidence, 34, 183, 202
 - validity, 34, 43, 47, 235
- Context-dependent item set format, 203
- Conventional MC format, 202, 205
- Co-teaching, 117, 302
- Council for Exceptional Children (CEC), 22, 56, 247
- Council of Australian Governments (COAG), 88, 92
- Criterion-referenced tests, 23, 203
- Criterion-related validity, 122, 226
- Cronbach alpha, 177
- Cue-using strategy, 150–151
- Curricular validity, 50–51, 53
- Curriculum
 - general, 1–2, 5–6, 22, 24, 27, 73, 75, 100–102, 104–108, 111, 113, 118, 121–123, 155, 214, 321–323
 - intended, 1, 5–7, 11, 99–108, 110–118, 120–122, 124–125, 179–180, 319–323
 - 1980s and 1990s: IEP as, 20–23
- Curriculum-based measurement (CBM), 24, 28, 77, 190, 195, 301
- Custom skills/activities, instructional time allocation to, 122–123
- Cut-score recommendations, 279, 325
 - methods, 285–286
 - technical report, 288
 - workshop, 276, 280, 284, 286–288, 290
- Cut score setting, *see* Standard setting process
- D**
- DDA, *see* Disability Discrimination Act (DDA)
- Debra P. v. Turlington*, 49–51, 53, 108
- Deductive reasoning strategy, 150–151
- Department of Education Organization Act (PL 96-88), 20
- Depth of knowledge (DOK), 11, 30, 131–132, 134–137, 139, 142–144, 177, 179, 322
- Design
 - access, supporting, 323–325
 - extended time studies, 190–192
 - language issues in, 217–228
 - student voices in, including, 243–253
 - test item, 163–180
 - UD, 2–3, 5, 10, 70, 156, 163–164, 178, 197, 202, 206, 211, 213, 243–247, 256–258, 271, 303, 307, 311, 319, 326
- Developing and Validating Multiple-Choice Test Items*, 204
- 6D Framework
 - deconstruct phase, 288
 - define phase
 - AA-MAS/AA-AAS and unmodified grade-level assessment, relationship between, 277–278
 - identifying desired ALD rigor, 277
 - identifying examinee population, 277
 - identifying intended uses of ALD, 279
 - identifying means for increased accessibility, 279
 - stakeholder involvement, 279
 - validity evidence, 279
 - deliver phase, 287
 - using reporting ALD, 287
 - validity evidence, 288
 - deploy phase, 286
 - refining/reporting target ALD, 286–287
 - validity evidence, 287
 - describe phase, 280
 - ALD, types of, 280–281
 - ALD, writing range and target, 281–284
 - proficient definition, 281–282
 - validity evidence, 284
 - design phase
 - AA-MAS/AA-AAS test designs, 284–285
 - methodologies, 285–286
 - validity evidence, 286
- Differential boost, 3, 8–9, 186–187, 237–238
- Differential item functioning (DIF), 187–188, 213, 244, 246
- Digital magnifying glass, 267–268
- Digital technologies, *see* Computerized tests and individual needs
- Disabilities
 - Australian policies, 83–94
 - SWD, *see* Students with disabilities (SWD)
 - SWOD, *see* Students without disabilities (SWOD)
 - U.S. policies, 19–30
- Disability Discrimination Act (DDA), 84–87, 91–92, 94
- Disallowed accommodations, 91
- Discovery Education Assessment, 250
- Discrepancy model, 28, 193
- E**
- EAHCA, *see* Education for All Handicapped Children Act (EAHCA)
- Educational environment, 100, 108, 131, 138, 144, 163, 179–180, 321, 325
- Educational evaluation, 20
- Educational measurement, 166, 204, 206
- Educational Measurement* (book), 201, 204
- Educational productivity model, 112
- Educational Researcher*, 115
- Educational systems, 26, 35, 52, 63
- Educational testing, 3, 20, 33–66, 87, 147, 246, 253, 255, 319
- Educational Testing Service (ETS), 207

- Education for All Handicapped Children Act (EAHCA),
20, 36
- Electronic input devices, 142
- Elementary School Journal*, 115
- Element interactivity, 171, 173–174
- Elimination testing, 214
- ELL, *see* English language learners (ELL)
- Embedded mnemonic devices, 142
- Enacted curriculum, 7–8, 100, 102–103, 106–110,
112–116, 120–125, 179–180, 322
- English language arts (ELA), 53–55, 63, 65, 121, 225,
233–235, 248, 283
- English language learners (ELL), 33–34, 43, 49, 56–66,
99, 111, 217–220, 222–228, 253, 308, 320,
324, 327
- Equitable OTL, 108
- Error-avoidance strategy, 150–151
- Essay formats, 203–204
- Evidence-based instructional practices, 113–114, 122,
326
- Evidence-based interventions, 24, 28–29
- Experimental studies, 184, 232–233, 235–239
- Extended MC items, 212
- External validity, 185, 194–196, 323–324
- Eyeglasses, 39, 48, 140, 149
- F**
- Factor analysis model, 119, 226
- Fair assessment practices, 24, 26
- Fair Housing Amendments Act, 164
- Fairness in testing, 12, 23, 63, 103, 149, 163, 201–202,
207, 211–212, 245
- Federal District Court (California), 256
- Federal guidance documents, 74–75
- Federal legislation, U.S. legal issues in SWD, 35–37
AA-AAS, 137
AA-MAS, 137
ADA, 36, 164
IDEA, 2, 36–37, 134–135
NCLB modified tests, 2, 46–47, 134–135
Rehabilitation Act, 2, 35–36
- Federally funded research projects, 202
- Federal Register, 70, 77, 297, 305
- Florida Comprehensive Achievement Test, 190
- Formal assessments, 201
- Formative assessments, 28, 88, 226–227
- Frame of reference, 183
- Free appropriate public education (FAPE), 19–20, 23,
104
- G**
- “Gap kids, the”, *see* Students with disabilities (SWD)
- General achievement test, 231–232, 236–237, 239
- General curriculum, 1–2, 5–6, 22, 24–25, 27, 73, 75,
100–102, 104–108, 111, 113, 118, 121–123,
155, 214, 321–323
- General Supervision Enhancement Grant (GSEG), 236,
297–300, 303, 305–308
- GI Forum v. Texas Educ. Agency*, 49–51, 57–58
- Goals 2000 Educate America Act (PL 103-227), 22
- “Gold standard”, 114, 232, 324
- Government/legal policies supporting inclusive
assessment practices
Australian policies, 83–94
IEP team decision-making, 69–80
U.S. legal issues, 33–66
U.S. policies, 19–30
- Grade-level proficiency, 71, 74–75, 144, 185, 232, 281,
295–297, 306
- Graduation testing challenges, U.S. legal issues
claiming racial/ethnic discrimination, 49–51
Debra P. and GI Forum cases, 49–50
notice and curricular validity, 50–51
retests and remediation, 50–51
involving ELL, 56–59
California case, 58–59
Texas case, 57–58
involving SWD, 53–56
Ambach case, 52–53
Brookhart case, 51–52
California case, 54–56
Indiana case, 53–54
- GRE innovations, 212–213
- Grid-in items, 204
- GSEG, *see* General Supervision Enhancement Grant
(GSEG)
- Guessing strategy, 150–151
- 2005/2007 Guidance, U.S. DOE, 73
- Guidelines for Constructed-Response and Other
Performance Assessments*, 207
- H**
- Handbook of Test Development*, 204
- Hierarchical linear modeling, 110–111, 116
- Hills Grammar School v. HREOC*, 84
- Human interpreters, 155, 256, 264–265
- Human-like digital figures, *see* Avatars, signing
- I**
- IDEA, *see* Individuals with Disabilities Education Act
(IDEA)
- IEP, *see* Individualized education program (IEP)
- IEP team decision-making
annual accountability assessment, participation in, 76
historic role of, 72
inclusive assessments, 69–72
1994 ESEA, full participation requirement, 69
flexibility of alternate assessments for SWD,
70–71
IDEA amendments, 69–70
modified/alternate achievement standards, 70–71
NCLB accountability provisions, 69
participation assumptions, 70
recommendations, 74–80
AA-AAS, participate through, 75
annual IEP meetings, decisions be made at, 74
considerations, 75, 79
general assessment, default decision, 74, 76

- state guidelines, 73–75
- statewide assessment participation, 72–74
 - decision-making, 72–73
- Improving America's Schools Act, 69
- Inclusion of Students with Disabilities for the National Research Council, 93
- Inclusive assessment, 2, 10, 19–20, 30, 69–72, 83–84, 164, 243–244, 246–248, 311, 324
- Indiana case, 53–54
- Individualized education program (IEP), 36, 93, 133, 301
- Individualized intended curricula, 105
- Individualizing Student Instruction Observation and Coding System (ISI), 118–119
- Individuals with Disabilities Education Act (IDEA), 2, 21, 24–26, 28–29, 36–37, 133, 153, 164, 243, 296
- Individuals with Disabilities Education Improvement Act (IDEA), 28, 72, 77, 100, 105, 134–135, 164, 255–256
- Initial Content Standard III, 247
- Innovative practices, *see* Test design and innovative practices
- Inspiration[®], 142
- Institute for Research on Learning Disabilities (IRLD), 184
- Instruction
 - access to, 1–12, 131–132, 137–138
 - classroom, 321–323
 - content on, 110–112
 - differentiated, 132–133
 - quality of, 112–113
 - unfolding of, 113–114
- Instructional accessibility, 131–132, 137–138, 180, 298, 323, 326
- Instructional accommodations, 13, 131–136, 138–142, 144, 180, 257, 322–323
- Instructional activities, 100, 118, 120, 135, 143, 258
- Instructional adaptations, 131–132, 139
 - accommodations, 139–142
 - location/condition, changes to, 141
 - and modifications, distinguishing between, 133–134
 - presentation, 139–141
 - response format, 142
 - timing and scheduling, 141–142
 - alignment with grade-level content standards, 134–135
 - case example, 145
 - consequences, 136–137
 - differentiated instruction, 132–133
 - integration based on students' needs, 144–145
 - interdependence, 137–138
 - modifications, 142–144
 - and accommodations, distinguishing between, 133–134
 - location/condition, changes to, 143
 - presentation, 142–143
 - response format, 144
 - timing and scheduling, 143–144
 - performance expectations and general education consistency, 135–136
- Instructional delivery, 113, 142–143
- Instructional dimensions, 108–121, 123–124, 322
- Instructional groupings, 7, 112–113, 122
- Instructional task demands, 10–11
- Instruction and achievement test access, 1–2
 - affecting SWD, 1
 - barriers, overcoming, 5
 - as central issue in instruction, 1
 - improvement actions and innovations, 11–12
 - access to achievement tests, 12
 - access to instruction, 11–12
 - legislative context and key concepts, 2–4
 - AA-MAS, 3
 - accessibility, 2–3
 - accommodation, 4
 - IDEA, 2
 - modification, 3–4
 - NCLB, 2–3
 - purpose, 3–4
 - Rehabilitation Act, 2
 - UD principles rationale, 2
 - via* OTL, 5–8
 - content/time/quality of instruction, 6–7
 - definition, 6
 - factors affecting, 6
 - measurement of, 6–7
 - SEC method, 7
 - via* testing accommodations, 8–9
 - differential boost, 8
 - effect sizes of, 9
 - implementation challenges, 9
 - interaction hypothesis, 8
 - research review, 8–9
 - typical changes in, 8
 - via* well-designed test items, 10–11
 - CLT, 10–11
 - development stages, 11
 - primary options and goals, 11
 - recommendations, 10
 - TAMI, 11
 - test accessibility, 10
- Integrated instructional supports, 145
- Intended constructs, 34, 37, 39, 41, 43, 64, 148, 158, 262, 269, 323, 140, 142
- Intended curriculum, 1, 5–7, 11, 99–108, 110–115, 117, 120–122, 124–125, 179–180, 319–323
- Intended curriculum model (ICM)
 - access to, 100
 - framework, 100–101
 - for general education, 101–104
 - failure to distinguish, 104
 - intended/assessed/planned/enacted curriculum alignment, 102–103
 - 2001 NCLB Act, mandate, 101–102
 - and OTL, 99–100

- for special education, 104–106
 - assessment options, 105
 - IEP/general curricula overlap, 105–106
 - student specific, 105
 - Intent consideration strategy, 150–151
 - Interaction hypothesis, 8–9
 - Internal validity, 184, 193–195
 - International Association for the Evaluation of Educational Achievement (IEA) studies, 110
 - Internet, 255
 - See also* Computerized tests and individual needs
 - Interpanelist consistency, 286, 288, 290
 - Interpretative/qualitative methods, 244–245
 - Interrater agreement, 115, 248
 - Interrater reliability, 119–120
 - Intraclass correlation coefficient (ICC), 111
 - Intrapanelist consistency, 286, 288, 290
 - Introduction to the Theory of Mental and Social Measurements*, 201
 - Iowa Test of Basic Skills (ITBS), 40, 189–192
 - IQ-performance discrepancy model, 28
 - Irlen Syndrome, 268
 - IRT, *see* Item response theory (IRT)
 - IRT framework, 237
 - ITBS, *see* Iowa Test of Basic Skills (ITBS)
 - Item
 - based OTL measures, 6–7, 110–111
 - development, 5, 10, 166–171, 176, 178–179, 197, 213, 227, 244, 251, 275, 280, 311, 324, 327
 - difficulty, 177–179, 208, 212, 232, 250, 324
 - discrimination, 177–179, 209
 - total correlations, 177–178, 203, 225
 - Item response theory (IRT), 214, 237
 - Item-writing, practice and evidence
 - access/assessment policy tools, 202
 - CR item
 - formats, 203–204
 - guidelines/taxonomy of, 206–208
 - format choice, 208
 - guidelines, empirically studied, 208–209
 - historical view, 201–202
 - MC item
 - formats, 202–203
 - guidelines/taxonomy of, 204–206
 - modifications for accessibility, research on, 210–214
 - AA-AAS, 210–211
 - AA-MAS, 211
 - alternative scoring methods, 213–214
 - assistive devices and technologies, 211–213
 - research, 201–202
 - three-option items, optimal, 209–210
 - tools
 - ARM, 327
 - TAMI, 327
- J**
- 2007 Joint Title I IDEA Regulations, 29–30
- K**
- Kansas Assessments of Multiple Measures (KAMM), 233
- Kansas General Assessments, 233–234
- K-3 educational system, 118–119
- K-12 educational system, 26, 104, 121, 187, 201, 209, 276, 284–286
- Kidspiration[®], 142
- Knowing What Students Know*, 245
- Knowledge, measurement of, 232, 321, 327
- Kuder-Richardson Formula 20, 235
- L**
- LAA 2, *see* Louisiana Educational Assessment Program Alternate Assessment, Level 2 (LAA 2)
- Language-based accommodation, *see* Language issues in item designing
- Language issues in item designing, 217
 - affecting performance, 219–220
 - complexity, unnecessary, 219
 - impact, 218–219
 - improvement steps, 226–228
 - linguistic accessibility, concept of, 217–218
 - modification procedures, 220–223
 - effectiveness, 224–225
 - measurement error, sources of, 225
 - validity, impact on, 226
 - research/methodological issues, 223–224
- Laptops, 91, 255
- Large-print test materials, 35, 48–49, 91, 134, 139–140, 154, 180, 191, 196, 267, 269
 - See also* Magnification methods
- Large-scale assessment, 6, 10, 13, 99, 102, 137–138, 157, 196, 210, 227, 243, 247, 322
- Large-scale tests/testing, 2, 22, 93, 183, 185, 195–196, 198, 202, 244, 247–250, 252, 257, 270–271
- Laws, 19, 22–24, 26, 35–38, 46, 48, 51, 56–58, 61–63, 69–70, 72–73, 84, 87, 105, 279, 289, 296, 311, 321
 - See also specific acts*
- Learning disabilities (LD), 23, 25–26, 29, 33, 39–43, 45–46, 48, 51, 76, 85, 91, 112, 149, 184, 186, 189, 193, 298, 301, 306
 - labels, 187, 189–191, 193, 252
- “Leveling the playing field”, 44–45, 149, 295
- “Life-skills” for independent living, 27
- Likert scale, 110, 168, 220–221, 227
- Listening comprehension, 38–39, 43–44, 142
- Local education agencies (LEA), 21, 26, 71, 78, 80, 275, 277, 279, 297, 311
- Louisiana Educational Assessment Program Alternate Assessment, Level 2 (LAA 2), 233–235
- Louvain University, 201
- M**
- Magnification methods, 49, 139–140, 154, 256, 259, 267–269, 271
- Manipulatives, use of, 8, 133, 139, 141–142, 285
- Masking methods, 155, 258–259, 268–269, 271

- Massachusetts Department of Education, 248
- Matching format, 43, 203, 205, 285
- MAZE measure, 238–239
- MCEECDYA, Early Childhood Development and Youth Affairs (MCEECDYA), Ministerial Council for Education
- MCEETYA 4 Year Plan, 88, 92
- Measurement errors, 155, 218–219, 224–225
- Mehrens-Kaminski continuum, 152
- Melbourne Declaration on Educational Goals for Young Australians, 87–88, 92
- Mental load analyses, 252
- Meta-analysis, 7, 9, 110, 150, 173, 204, 208–209
- Microsoft PowerPoint®, 142
- Ministerial Council for Education, Early Childhood Development and Youth Affairs (MCEECDYA), 87, 90, 94
- Modification(s)
- CAHSEE, 54–56
 - definition, 3–4, 38
 - ELL, 60, 62
 - goals of, 11
 - instructional, 131–138, 142–144, 322
 - item, 3, 11, 30, 46–47, 172, 175–177, 179–180, 208, 210–211, 231–240, 246–247, 250–251, 279, 311–312, 324–326
 - linguistic, 220–228
- Modification packages, test and item accessibility improvement, 231
- AA-MAS policy, 231–232
 - characteristics, 232–233
 - coefficient alphas, 232, 234
 - measurement for proficiency, 232
 - new policy criteria, 232
 - experimental studies, 235–236
 - CAAVES, 236–238
 - CMAADI, 238
 - OASIS, 238–239
 - findings, 239–240
 - prospects, 240
 - state-modified achievement tests, 232–233
 - KAMM, 233
 - LAA 2, 233–235
 - TAKS-M, 235
- Modified achievement tests, 106, 232–235, 239, 295–296
- AA-MAS rationale, 308–313
 - consequences, 313–315
 - federal regulations, history of, 296–297
 - GSEG project, 297–298
 - focus groups, 302–303
 - PSSA performance trends analysis, 303–305
 - recommendations and guidance, 305–308
 - survey, 298–302
- Multi-level analysis, 110–111, 116
- Multi-level models, 116
- Multi-level OTL studies, 112, 116
- Multimedia learning theory, 166
- Multiple choice (MC) items, 45, 54, 89, 201–203, 239, 261
- formats, 202–204
 - guidelines/taxonomy of, 204–206
 - key characteristics, 168–169
- Multiple regression, 110
- Multiple true-false format, 202, 205
- MyiLOGS, *see* My Instructional Learning Opportunities Guidance System (MyiLOGS)
- My Instructional Learning Opportunities Guidance System (MyiLOGS), 120–124, 326
- N**
- NAEP, *see* National Assessment of Educational Progress (NAEP)
- NAPLAN, *see* National Assessment Program for Literacy and Numeracy (NAPLAN)
- National Accessible Reading Assessment Projects (NARAP), 213
- National achievement scales, 89
- National Assessment of Educational Progress (NAEP), 21, 60, 131, 201, 207, 222, 224, 284, 287
- National Assessment Program for Literacy and Numeracy (NAPLAN), 83, 88–94
- National Association of School Psychologists, 77, 247
- National Center for Special Education Research, 213
- National Center on Educational Outcomes (NCEO), 21–22, 74, 185–186, 243, 250
- National Commission on Excellence in Education, 21
- National Council on Measurement in Education (NCME), 3–4, 33, 37, 103, 184–185, 197, 210, 226, 232, 245–246, 253
- National Instructional Materials Accessibility Standards (NIMAS), 256
- National Longitudinal Transition Study, 21
- National Research Council, 8, 93, 245
- NCEO, *see* National Center on Educational Outcomes (NCEO)
- NCLB, *see* No Child Left Behind Act (NCLB)
- NCME, *see* National Council on Measurement in Education (NCME)
- “New generation” assessment, *see* “Race to the Top” (RTT) assessment initiatives
- New Hampshire Department of Education, 271
- New York Supreme Court, 36
- No Child Left Behind Act (NCLB), 2–3, 26–28, 46–47, 55, 59–60, 62–65, 69–70, 72, 80, 93, 100–106, 108–109, 134–135, 153, 185, 187, 202, 210, 231, 240, 243–244, 256, 276, 279, 295–296, 306, 308–309, 319
- Non-biased items, 10, 164, 213, 311
- Non-comparable scores, 37–38, 43–47, 53–54
- Non-English language learners (non-ELL), 43, 57, 60–61, 63–64, 217–218, 224–226, 228, 324
- None-of-the-above (NOTA) items, 205, 208–209
- Non-essential visuals, 172
- Non-LD labels, 187, 189–190, 193
- Non-Regulatory Guidance*, U.S. DOE, 244, 297

- Non-standard test administrations, 33–36, 38, 40–41, 44–46, 49, 53, 64–66, 320
 eligibility, 47–48
 labeling, 39
 Norm group, 40
 Norm-referenced tests, 22–23, 40–41
 Notice validity, 50–51
 Numbers/operations, instructional time to, 122–123
- O**
- OAASIS, *see* Operationalizing Alternate Assessment for Science Inquiry Skills (OAASIS)
 Office for Civil Rights (OCR), 54
 Office of Special Education Programs (OSEP), 21, 24, 26, 28, 213, 297
 Online teacher tools, *see* My Instructional Learning Opportunities Guidance System (MyiLOGS)
 Operationalizing Alternate Assessment for Science Inquiry Skills (OAASIS), 233, 236, 238–240
 Opportunity to learn (OTL), 2, 99–100
 alignment methodologies
 constraints, 107
 SEC method, 102
 Webb method, 102
 dimensions, 108–113
 content of instruction, 110–112
 quality of instruction, 112–113
 synthesis of, 113–114
 time on instruction, 108–109
 documentation, 107–108
 ICM
 framework, 100–101
 for general education, 101–104
 for special education, 104–106
 student access to, 100
 research, 106–108
 conceptual and substantive relevance, 107–108
 measurement of, 114–124
 prospects of, 124–125
 Optical mark recognition software, 89
 Option items, three/four/five-, 209–210, 212, 214, 233
 Oral presentations, 8, 54, 120, 134, 139–140, 193, 201, 203, 259, 263
 Oregon case, 45–46
 Oregon Statewide Test, 192
 OSEP, *see* Office of Special Education Programs (OSEP)
 Otherwise qualified individuals, 20, 35–36, 44, 55–56
 OTL, *see* Opportunity to learn (OTL)
- P**
- Paper-based tests, 39, 42–43, 55, 148, 154–155, 225, 248, 257–260, 266, 271
 “Pathway to accessibility”, 169
Peabody Journal of Education, 204, 211
 Peer-assisted learning strategies, 139–140, 145
 Peer-reviewed alternate assessments, 27, 60, 63, 74
 Pennsylvania GSEG survey, 298–302
 Pennsylvania System of School Assessment (PSSA), 298, 301, 303–305, 307–308
- Performance expectations, 6, 131–133, 135–136, 138, 141–143
 Performance levels, 93, 135, 184, 276, 302, 307
 Performance standards, 22–23, 63, 70, 101–102, 104, 135, 296
 See also Performance expectations
 Peripheral devices, 266, 271
 Personal decisions, 69–80
 Policies, *see* Government/legal policies supporting inclusive assessment practices
- Practice(s)
 classroom-based assessment, 137–138, 142, 257
 ethics and standards for professional, 246–247
 evidence-based instructional, 113–114, 122, 326
 fair assessment, 24, 26
 government/legal policies supporting inclusive assessment
 Australian policies, 83–94
 IEP team decision-making, 69–80
 U.S. legal issues, 33–66
 U.S. policies, 19–30
 innovative, *see* Test design and innovative practices
 validity evidence, 185–188
- Praise statements, usage of, 7, 112
 Predictive validity, 119–120, 122
 Prerequisite skills, 10, 50
- Presentation
 accommodations, 139–141
 auditory, 140–141
 tactile, 141
 visual, 140
 modifications, 142–143
- President’s Commission on Excellence in Special Education, 26
- Principals, role of school, 72, 89–92, 286
Principles for Professional Ethics, 247
 Professional judgment, 27, 37, 56, 90
 Proficient performance, 79, 275–291
 Programme for International Student Assessment, 83
 Prompts, 54, 78, 99, 101, 139, 141–142, 189, 213, 248–249, 251, 260, 268–269, 285, 303
 Psychometrics, 4, 33–35, 37–39, 46, 61–66, 118, 122, 153–154, 202, 206, 214, 228, 244–245, 277–278, 320
 See also Rasch model
- Q**
- QSA, *see* Queensland Studies Authority (QSA)
 Qualified individual with disability, 36
 See also Otherwise qualified individuals
 Quality indicators, 7, 112–113
 Quality of instruction, 6–8, 107–108, 112–113, 115, 125, 322
 Quantitative methods, 244
 See also Statistical analyses
 Quasi-experimental studies, 9, 184–185, 187, 193, 197–198
 Queensland Special Provisions, 91

- Queensland students afforded special consideration and exemptions, 92
- Queensland Studies Authority (QSA), 90–92
- R**
- “Race to the Top” (RTT) assessment initiatives, 30
- Racial/ethnic discrimination, 49–51, 87, 212, 250
- Range ALD, 280–284, 288–291
- Rasch model, 238
- Read aloud accommodations, 41–42, 139–141, 145, 148, 154, 185–189, 191–193, 195–198, 236, 239, 256–263, 266, 268, 271, 312, 323
- Reading comprehension, 37–39, 43–44, 54, 142, 172, 188–193, 219, 251, 312
- Reading School District case (Pennsylvania), 62
- Reasonable accommodations, 19, 25, 36, 38–39, 41, 60, 311
- “Reasonable adjustment”, 86, 91
- Reasoning competency, types of, 206
- Reauthorization, Elementary and Secondary Education Act (ESEA)
- 1993, 21–22
 - 1994, 26, 69
- Reauthorization, Individuals with Disabilities Education Act (IDEA)
- 1990, 21
 - 1991, 35
 - 1997, 2, 22, 24–27, 69–70, 72, 105, 296, 320
 - 2004, 28, 72, 77, 100, 105, 134–135, 164, 255–256
- Rehabilitation Act, 2, 19–20, 35–36, 39, 52
- Reliability, 11, 33–34, 111, 119–120, 124, 147–148, 165, 177–178, 184, 194, 196, 202–203, 209, 211, 214, 223–225, 228, 233–237, 239–240, 244, 321, 324
- Remedial instruction, 21, 24, 29, 50–53, 55–56, 59, 63, 103, 284, 306
- Remote control, use of, 264
- Rene v. Reed*, 53
- Reporting ALD, 281, 286–291
- Response modes
- accommodations, 142
 - adapted, 258, 325
 - modifications, 144
- Response theory, 214
- Response to intervention (RTI), 28–29, 77, 193, 256
- Retests and remediation, 50–51
- Rhode Island Department of Education study, 257
- Rhode Island Performance Assessment, 190
- Rodriguez meta-analysis, 209
- RTI, *see* Response to intervention (RTI)
- RTT, *see* “Race to the Top” (RTT) assessment initiatives
- Rule violations, item writing, 209
- S**
- San Francisco Unified School District (SFUSD), 61–62
- SAT-9, Stanford Achievement Test Ninth Edition (SAT-9)
- Scaffolded instructional prompts, 142
- Scale scores, 89, 279
- School learning model, 7, 108–109, 112
- School psychologists, 247, 302
- Sciences, 37, 169–170, 172–176, 201–202, 224, 259, 323–324
- Scores
- comparable, 33–34, 37–40, 44–46, 55, 64, 66, 224
 - cut, 275–291
 - non-comparable, 37–38, 43–47, 53–54
 - reliability, 202–203, 209, 211–212, 326–327
 - validity, 2–3, 34, 138, 269, 272, 327
- Screen readers, 139–141
- Scribes, 91, 139, 141–142, 155, 180, 256, 258
- SEC, *see* Surveys of the Enacted Curriculum (SEC)
- Settings (location/condition changes)
- accommodations, 141
 - modifications, 143
- Short answer items, 168, 201, 203–204
- Sign based accommodation, 263–266
- Signed English, 263, 265–266
- Signing avatars, 165, 264–266, 325
- Sign language, 8, 54, 165, 256, 259, 263
- SLD, *see* Slow Learning Disability (SLD)
- Slow Learning Disability (SLD), 26–29
- Socioeconomic status (SES), 109–110, 116, 310–311
- Southeastern Community College v. Davis*, 20, 23, 35
- Spearman-Brown formula, 234–236
- Special education tracking, 306, 313
- Special populations, 23, 33–66
- See also* Students with disabilities (SWD)
- Spencer Foundation, 245
- Standardized tests, 23, 25, 53, 69, 264
- Standards-based accountability system, 23, 69, 80, 100, 253
- Standards-based assessment/testing, 30, 183–184, 188, 250, 276–277
- Standards-based reforms, 2, 6, 22, 26, 104, 110, 306
- Standard setting process, 275–276
- ALD development prior to, 290
 - definition, 276
 - design and development steps, overview, 290
 - implementation, 289–290
 - panelists/committee members, 280
 - process planning/evaluations, 279–280
 - technical report validity, 288–289
 - terminology, 276
 - validity, 276–277
- See also* 6D Framework
- Standards for Educational and Psychological Testing*, 3–4, 33–34, 37–39, 43–45, 47, 50, 56, 61, 63, 183–185, 197, 210, 226, 232, 245–246, 253
- Standard testing conditions, 36, 39–40
- Stanford Achievement Test Ninth Edition (SAT-9), 21, 61–62
- State, Territory, and Commonwealth Ministers of Education, 87
- State assessment participation guidelines, 74–75
- State Board of Education (SBE), 54–55, 61, 276
- State Department of Education, 74, 76

- State education agencies (SEA), 26, 275–276, 279, 281–282, 289
- Statistical analyses, 194, 244, 251
- Statistical conclusion validity, 184–185, 194–196
- Students
- achievement, 5–8, 12, 27, 88–89, 94, 99–100, 103, 107–113, 116, 153, 278, 282, 296, 319, 322–323, 327
 - characteristics, 188–189, 193
 - ELL, 226–228
 - exemptions, 69, 90–92
 - needs, 8, 77, 88, 132, 144, 156, 262, 266
 - See also* Instructional adaptations
 - OTL, *see* Opportunity to learn (OTL)
 - participation, 90
 - PSSA performance trends analysis, 303–305
 - response data, 245–246, 248–252
 - test-taking skills, *see* Test-taking skills and impact on accessibility for students
 - withdrawals, 90–93
- Students with disabilities (SWD), 1–9, 42, 44–46, 69–73, 76–78, 88, 90–91, 93–94, 100–102, 104–108, 112, 117, 149, 153, 157–158, 179–180, 183–185, 187, 190, 192, 194, 196–197, 213, 219, 231–232, 236–239, 243–244, 247, 251–253, 256, 261, 270–271, 275, 295–297, 299, 302–303, 305–308, 311, 315, 319–322, 326–327
- See also* Instructional adaptations
- Students without disabilities (SWOD), 2, 4, 8, 41, 72, 91, 102, 105–107, 113, 149, 158, 236–239, 251
- Student voices in inclusive assessments, 243–244
- epistemological and methodological frameworks, 244–245
 - extant research integrating students, 247–252
 - using drawings, 248–249
 - using interviews, 249–252
 - professional practice, ethics/standards for, 246–247
 - prospects, 252–253
 - response data, uses of, 245–246
- Study of Instructional Improvement (SII) logs, 118, 120–121
- Surveys of the Enacted Curriculum (SEC), 7, 102, 110–111, 120–122, 125
- Sydney Morning Herald*, 92
- Symbolic communication systems, 27, 41–42, 91, 259, 263
- See also* Sign language
- T**
- Tactile accommodations, 139–141, 155, 164, 259, 266–267, 325
- TAKS-M, *see* Texas Assessment of Knowledge and Skills – Modified (TAKS-M)
- Talking Tactile Tablet (TTT), 266–267
- TAMI, *see* Test Accessibility and Modification Inventory (TAMI)
- TAMI Accessibility Rating Matrix (ARM), 11, 168–169, 171–176, 179, 327
- Target ALD, 280–284, 286–291
- Target constructs, 2–3, 8, 10–11, 103–104, 147, 157, 163–167, 169–170, 172, 174, 177–180, 212, 226, 251, 320, 323–327
- Taxonomies, content, 6–7, 110–111, 117, 134, 203–206, 208
- Teacher-child interactions, 115, 118
- Teacher information, 298–299
- Teacher logs, 115–116, 120
- Teacher-recommended cut scores, 284, 286
- Teachers' instruction, 6, 100, 136
- Technology Assisted Reading Assessment (TARA) project, 213
- Telecommunications Act, 164
- Terra Nova 18, 190
- Terra Nova Multiple Assessment Battery, 192
- Test Accessibility and Modification Inventory (TAMI), 11, 166, 168–170, 176, 179, 211, 213, 250–251, 327
- Test accessibility theory, 165
- Test accommodations, 8, 64, 185–186, 193, 211, 257–258, 260, 269–271
- Test Anxiety Inventory for Children and Adolescents (TAICA), 252
- Test delivery systems, 155, 165, 168, 258–260, 263, 267, 271
- Test design and innovative practices
 - accessibility theory for test-takers, 163–180
 - accommodated/modified large-scale tests, 183–198
 - CBT, 255–272
 - 6D Framework, 275–291
 - item-writing, 201–214
 - linguistic complexity in, 217–228
 - modification packages, 231–240
 - student voices, inclusion of, 243–253
- Test developers/development, 2, 4, 20, 23, 37, 65, 147, 153–155, 158, 163, 165, 171, 178–179, 188, 202, 206, 213, 223, 239, 244–246, 253, 258, 260, 262, 270, 275, 279–280–281, 289–291, 295–296, 311, 319, 323–328
- Tested construct, 33–34, 43, 61, 178, 269, 312, 322
- Testing accommodations, 1–5, 8–10, 12–13, 23, 33–34, 39, 74, 93, 138, 155, 164, 179–180, 237, 246–247, 251, 257, 326–327
- See also* Non-standard test administrations
- Testing Individuals of Diverse Linguistic Backgrounds, 56–57
- Testing protocols, 4, 89, 94
- Test item design, 10–11, 163–180, 217–228, 231–240, 267–269
- Test Preparation Handbook*, QSA, 90–91
- Tests of achievement, 1–19, 295–315, 319–328
- Test Standards, *see* *Standards for Educational and Psychological Testing*
- Test-takers/taking, 2–3, 8, 10–11, 103, 148–150, 152–155, 157–158, 163–180, 197, 211, 236, 249, 271, 323–326

- Test-taking skills and impact on accessibility for students, 147
- interactions with other improvement methods, 153–157
- accommodations, 153–154
- CBT, 154–157
- modifications, 154–155
- practical implications, 157–158
- test-wiseness
- access and, 147–148
- definition, 148
- frameworks and findings, 150–153
- threshold hypothesis, 148–150
- Test-taking strategies, 149, 157
- Test-wiseness, 147–153
- Texas Assessment of Knowledge and Skills – Modified (TAKS-M), 233, 235
- Texas case, graduation testing challenges involving ELL, 49–51, 57–58
- Texas Education Agency, 235
- Texas Education Code, 51, 57
- Texas Reading Assessment of Knowledge and Skills (TAKS), 191
- Texas Student Assessment Program, 232, 235
- Text-based information, 259–261, 263, 266–267
- Third International Math and Science Study (TIMSS), 224
- Third-party observations, 114
- Threshold hypothesis, 148–150
- Time
- and content dimension, 11, 113
- and quality dimension, 8, 11, 113–114
- using strategy, 150–151
- Timers, 139, 142
- Timing and scheduling
- accommodations, 139, 141–142
- modifications, 143–144
- 2002–2003 Title I ESEA regulations, 27–28
- TOEFL innovations, 212
- True-false format, 147, 202–203, 205
- Two-tier education system, Australian, 83
- Type-K formats, 205, 208–209
- U**
- Understand/apply, cognitive process dimension, 123–124
- “Unicorn syndrome”, 315
- United Nations Convention on the Rights of the Child (UNCRC) Article 12, 247
- Universal design (UD), 2–3, 5, 10, 70, 156, 163–164, 178, 197, 202, 206, 211, 213, 243–247, 256–258, 271, 303, 307, 311, 319, 326
- See also specific designs*
- Universal Design for Learning (UDL), 256, 307
- “Universal proficiency” goal, 29
- University of Bologna, 201
- University of Kansas, 140–141
- University of Minnesota, 184
- Unmodified grade-level assessments, 275–278, 282, 284–285, 290
- U.S. Department of Education (U.S. DOE), 2–3, 22–24, 26, 59–60, 64, 93, 153, 185, 231–232, 235–236, 240, 244, 271, 295, 297
- U.S. Government Accountability Office, 217
- U.S. inclusive assessment policies for SWD, 19–30
- 1960s and 1970s, inclusion and equal access, 19–20
- EAHCA, 20
- ESEA amendments of 1974, 20
- FAPE, 19–20
- Rehabilitation Act, 19–20
- Title VI of the Civil Rights Act of 1964, 19
- 1980s and 1990s, IEP curriculum, 20–23
- ADA, 23
- CEC testimony, 22
- criterion-referenced tests, 23
- Department of Education Organization Act (PL 96-88), 20–21
- education and assessment gap, 22
- ESEA reauthorized (1993) attempts and reform approaches, 22
- IDEA reauthorized (PL 101-476; 1990) focus and funded studies, 21
- National Commission on Excellence in Education, 21
- norm-referenced tests, 22–23
- OSEP, 21
- participation guidelines, 23
- Rehabilitation Act regulations, 20
- sensory impairments v. cognitive disabilities, debate, 23
- 1997 IDEA and alternate assessment options, 24–26
- CBM measures, 24
- classes of SWD, 24–25
- communication, mode of, 25
- computer-adaptive assessments, 26
- elements, new, 25
- endorsement, 24
- IEP as tool, 26
- initiated by Hoppe, 25
- non-native English speakers, inclusion of, 24
- 1999 Part B regulations mandate, 26
- SLD students, 26
- 2001 NCLB Act, 26–27
- annual assessments, 26
- Paige and Spellings emphasis on, 26–27
- President’s Commission on Excellence in Special Education, 27
- “wait-to-fail” model, 27
- 2002–2003 Title I Regulations on alternate assessments, 27–28
- CBM-based assessments, 28
- “life-skills” for independent living, 27
- 2004 IDEA and RTI assessment, 28–29
- 2007 Joint Title I IDEA Regulations, 29–30
- RTT assessment initiatives, 30
- U.S. legal issues in educational testing for SWD, 33
- accessibility and construct preservation/score comparability, 38–39
- accountability testing challenges, 59–63

- California case, 61–62
 - construct shift, 61
 - ELL accommodations/modifications, 60
 - majority/minority ELL, 60–61
 - NCLB Act, 59–60
 - NCLB cases, 62–63
 - construct fragmentation and shift, 43–44
 - ELL v. non-ELL, 43
 - mathematical assistance to SWD, 43
 - federal legislation, 35–37
 - ADA, 36
 - IDEA, 36–37
 - NCLB modified tests, 46–47
 - Section 504 of Rehabilitation Act, 35–36
 - graduation testing challenges
 - claiming racial/ethnic discrimination, 49–51
 - involving ELL, 56–59
 - involving SWD, 53–56
 - labeling non-standard test administrations, 39
 - “leveling the playing field”, access v. success, 44–45
 - non-standard test administrations, 33–35
 - access, 33
 - burdens, undue, 47–48
 - content validity evidence/judgments, 34
 - definition, 33
 - eligibility for, 47–48
 - tested construct, 33–34
 - testing accommodation, 33
 - Oregon case settlement, 45–46
 - professional standards, 37–38
 - public policy exceptions, 44
 - recommendations, 63–66
 - skill substitution, 40–43
 - calculator use, 42–43
 - extended time alteration, 40–41
 - readers, 41–42
 - U.S. policies, *see* U.S. inclusive assessment policies for SWD
 - U.S. Supreme Court, 20, 23, 35–36, 48, 58
- V**
- Valenzuela v. O’Connell*, 57–58, 63
 - Validity evidence, 183–185
 - constructs, 197
 - definition, 184
 - extended time research
 - elements for, 190
 - student characteristics, 187–189
 - task demands, 187–188
 - operationalization, test design, 197–198
 - outcomes, validation of, 198
 - read aloud research
 - elements for, 191–192
 - student characteristics, 188, 193
 - task demands, 188–193
 - research designs and quality
 - extended time, 194–195
 - read aloud accommodations, 195–196
 - response processes, current practices, 185–188
 - systems organization, 198
 - teacher training, 198
 - types of, 184–185
 - See also* 6D Framework
 - Variance decomposition, 111, 116
 - VHS/DVD/CD/MP3 players, 141, 264
 - Video game technologies, 156–157
 - Visual accommodations, 139–142, 155, 166–167, 172–173, 211, 213, 251, 324–325
 - Visual acuity difficulties, 149
 - Voice-over technology, 233, 236, 239
- W**
- “Wait-to-fail” model, 27
 - Waivers, 44, 53, 55–57, 62
 - Walberg model, 112
 - Washington Administrative Code, 193
 - Washington statewide achievement test (WASL), 191
 - Webb method, 102
 - Webb’s taxonomy, 134
 - What Every Special Educator Must Know: Ethics, Standards, and Guidelines*, 247
 - Withdrawals, student, 90–93
 - Within-method consistency, 286, 290
 - Writing Test Items to Evaluate Higher Order Thinking*, 204