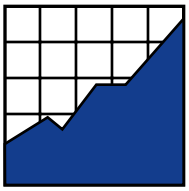
Lessons Learned in Federally Funded Projects that Can Improve the Instruction and Assessment of Low Performing Students with Disabilities

Edited by Martha L. Thurlow, Sheryl S. Lazarus, and Sue Bechard

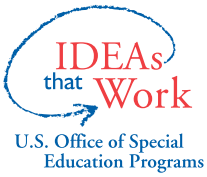
January 2013

Thurlow, M. L., Lazarus, S. S., & Bechard, S. (Eds.). (2013). *Lessons learned in federally funded projects that can improve the instruction and assessment of low performing students with disabilities*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Center is supported through a Cooperative Agreement (#H326G11002) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. This report was funded with partial support from the Multi-state GSEG Toward a Defensible AA-MAS. This project is supported by General Supervision Enhancement Grants (#H373X070021) from the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.



NCEO Core Staff

Martha L. Thurlow, Director
Deb A. Albus
Manuel T. Barrera
Laurene L. Christensen
Linda Goldstone
James Hatten
Christopher J. Johnstone
Jane L. Krentz
Sheryl S. Lazarus
Kristi K. Liu
Ross E. Moen
Michael L. Moore
Rachel F. Quenemoen
Rebekah Rieke
Christopher Rogers
Miong Vang
Yi-Chen Wu

National Center on Educational Outcomes
University of Minnesota • 207 Pattee Hall
150 Pillsbury Dr. SE • Minneapolis, MN 55455
Phone 612/626-1530 • Fax 612/624-0879
http://www.nceo.info

Table of Contents

Chapter 5

Consortium for Modified Alternate Assessment Development and Implementation: Lessons Learned

Stephen N. Elliott
Arizona State University

Michael C. Rodriquez
University of Minnesota

Andrew T. Roach
Arizona State University

Peter A. Beddow
Lipscomb University

Ryan J. Kettler
Rutgers, The State University of New Jersey

Alexander Kurz
Arizona State University

This manuscript was supported, in part, by the U.S. Department of Education Office of Elementary and Secondary Education (Grant No. H373X070026). However, the opinions expressed do not necessarily reflect the position or policy of the U.S. Department of Education and no official endorsement should be inferred.

Introduction

The primary purpose of the *Consortium for Modified Alternate Assessment Development and Implementation* (CMAADI) project was to provide research and technical support in the development and implementation of alternate assessments based on modified academic achievement standards (AA-MAS) to the Arizona (ADE) and Indiana (IDE) departments of education. This project was in many ways a continuation of the CAAVES (Consortium for Alternate Assessment Validity and Experimental Studies; Compton & Elliott, 2006-2009) project, where both ADE and IDE assessment leaders interested in special education students' test performance, worked on ways to improve alternate assessments based on alternate academic achievement standards (AA-AASs) and prepare for the possibility of an AA-MAS test. The results of the CAAVES project were highly influential to the CMAADI project and have been published in refereed journals: Elliott et al., 2010 and Kettler, Rodriguez, Bolt, Elliott, Beddow, & Kurz, 2011.

The CMAADI project started in 2007, just as federal policy for AA-MASs was formally initiated, and it concluded in 2011. During this period, the state of Arizona decided not to move forward with the development of an operational version of an AA-MAS primarily due to financial reasons, while the state of Indiana forged ahead and introduced an AA-MAS in 2010. Regardless of these different decisions regarding an AA-MAS, both state partners were supportive of the CMAADI project and used the resulting research to influence an array of assessment decisions and actions that are likely to have a positive influence on the assessment of students with disabilities, as well as student without disabilities.

The students with disabilities of primary concern to this project were known to have had persistent academic difficulties that resulted in performances in the lowest proficiency level on their state's achievement and accountability test for two or more consecutive years. Many educators in our partner states believed these students were learning, but not at a substantive rate and as a result the general state assessment did not have enough items at the lowest levels of difficulty to detect what these students actually had learned. The bases for these students' low performances, regardless of a specific disability, were thought to be influenced by two common problems: poor reading fluency and limited opportunity to learn the tested content due to a relatively slow pace of instruction and learning. Thus, as we approached the challenge of helping states create tests that could meaningfully assess these students, we focused on ways to reduce the impact of reading fluency problems, and related memory issues that occur with poor readers, and to also increase students' opportunities to learn the content measured by their state tests in reading and mathematics.

A description of the highlights of our research with partners in Arizona and Indiana departments of education follows. Resources listed at the end of this chapter provide readers important additional information, as does the chapter on the MAAPS project later in this same volume.

Overall, this work advanced the understanding and measurement of accessibility of tests and the opportunity to learn the content standards these tests are designed to measure. Key to these advancements in item/test accessibility and opportunity to learn were the development or continued refinement of three tools: the *Test Accessibility & Modification Inventory* (TAMI) (Beddow, Elliott, & Kettler, 2008), *Test Accessibility Rating Matrix* (Beddow, Elliott, & Kettler, 2008), and *Instructional Opportunity Learning Guidance System* (MyiLOGS; Kurz, Elliott, & Shrago, 2009). These tools helped all parties in the technical assistance enterprise communicate about important assessment and instructional concepts that are central to the development of a test-based accountability system that places a high value on valid test score inferences. Now on to the research and the lessons we have learned that might be useful to others who are motivated to enhanced assessments for all students, and in particular students with disabilities who receive most of their instruction in general education classrooms across the country.

Designing Highly Accessible Test Items

Among the challenges facing the project team in developing an assessment for students with special needs was the need to integrate universal design elements into the test and test items while ensuring the new test was still technically sound. A critical objective was to ensure test results yielded scores from which subsequent inferences retained similar validity to the general assessment. The CMAADI project team approached this challenge by using existing test items as source material for the design of the AA-MAS test items, modifying the existing items with a focus on increasing their accessibility for the target population of an AA-MAS. The resulting parallel test forms permitted comparison across experimental conditions. The process provided a focus for item writers while facilitating the alignment of new items with their intended content standards.

To examine the difference in the extent of accessibility between the two assessments, the team required an empirical means of evaluating the degree to which each of the two tests were free from barriers that would decrease a test-taker's ability to demonstrate competence in the manner for which the tests were designed. In terms of the development of AA-MASs for the subset of students identified with disabilities, the test items needed to be modified using a method that precluded, to the degree possible, the need for individualized accommodations. To address this need, the team embarked on the creation of a tool called the *Accessibility Rating Matrix* (Beddow, Elliott, & Kettler, 2009).

Winter, Kopriva, Chen, and Emick (2006) defined *access* as "…the interaction between construct irrelevant item features and person characteristics that either permits or inhibits student response to the target measurement content of the item" (p. 276). In one instance, a test may be maximally accessible to the majority of test-takers, but be inaccessible to the balance of

test-takers who share a common functional impairment, such as blindness. Another test may be maximally accessible to most test-takers, but be largely inaccessible for individuals who are unable to hold a writing instrument or use a computer keyboard. In both of these cases, test developers and users (e.g., test administrators) may increase the accessibility of the test by altering the administration or response conditions of a test to accommodate the needs of test-takers for whom the standard test conditions do not permit complete access. As these examples suggest, some students will continue to need testing accommodations even when items are maximally accessible. While developing items and tests that are perfectly accessible for all test takers is unrealistic, the goal of universal design in assessment and item modification should be to yield tests that are maximally accessible for nearly all test takers. That is the purpose of the Accessibility Rating Matrix (ARM; Beddow, Elliott, & Kettler, 2009), the topic of the next section.

Development of the ARM

Several areas of research influenced the development of the *ARM* (Beddow et al., 2009), which was designed with the purpose of facilitating decision-making when writing or modifying items with a focus on increasing their accessibility for more test-takers, specifically those with special needs. Cognitive load theory (e.g., Chandler & Sweller, 1991; Sweller, 2010), item-writing guidance (Haladyna, Downing, & Rodriguez, 2002; Rodriguez, 2005), and Universal Design for Learning (Rose & Meyer, 2006) were among the dominant areas of prior work that were considered.

Cognitive Load Theory. In his well known "Magical Number Seven, Plus or Minus Two" article, Miller (1956) presented a synthesis of the research on what he termed *channel capacity*—that is, the amount of information a person is able to process about a given stimulus, also known as *working memory*. Across a series of studies investigating participants' channel capacity for several variables including auditory pitch and loudness, taste, and visual identification of size and position, Miller reported the mean channel capacity was approximately seven categories (i.e., number of discriminable pitches or loudnesses, concentrations of saltwater, and object sizes or position, respectively). Across variables, the standard deviation was approximately 3 with an overall range of 3 to 15 categories. Channel capacity was slightly higher when participants were permitted to identify categories on the basis of two or more variables (e.g., saltiness and sweetness for taste, pitch and loudness for audio stimuli, position and size for visuals, hue and saturation for color). Miller was surprised, however, at the minimal degree to which multidimensionality appeared to augment participants' capacity for processing information. (Note contemporary reviews of this classic work suggest that Miller's use of the "magic number seven, plus or minus two" was rhetorical and the results of his work actually suggest the limit of working memory is closer to three or four units [Farrington, 2011]. Actual memory capacity appears to depend on the information being stored and it seems memory span is not a constant. Cowan

[2001] also provided evidence in a number of settings that the limits of cognition is closer to a magical number of four.)

Cognitive load theory (CLT) is a logical and theoretical extension of Miller's (1956) work. Until CLT was applied to assessment (e.g., in the research described in the following section), the theory was singularly used as a model for understanding the demands of learning tasks, grounded in the assumption that the mind has a limited capacity (i.e., in working memory) for processing information. In essence, CLT proponents posit that to properly gain knowledge from instruction, students must: (a) attend to the presented material, (b) mentally organize the material into a coherent structure, and (c) integrate the material with existing knowledge. Thus, the efficiency of instructional tasks depends on the extent to which the cognitive resources needed for this process are minimized.

Accordingly, CLT disaggregates the cognitive demands of learning tasks into three load types: *intrinsic load*, *germane load*, and *extraneous load*. Intrinsic load refers to the amount of mental processing that is required for completing a task. Germane load refers to cognitive demands that are not necessary for gaining essential knowledge but enhance learning by facilitating generalization or automation (e.g., lessons that require learners to extend learned concepts to arenas outside the classroom or apply them to novel situations). Extraneous load refers to the demand for cognitive resources to attend to and integrate nonessential elements that are preliminary to actual learning, but are nonetheless required for a learning task. Proponents of CLT argue that learning tasks should be designed with the goal of minimizing the demand for cognitive resources that are extrinsic to the goals of instruction. The triune model of cognitive load was encapsulated by Paas, Renkl, and Sweller (2003): "Intrinsic, extraneous, and germane cognitive loads are additive in that, together, the total load cannot exceed the working memory resources available if learning is to occur" (p. 2).

Intrinsic load contains all essential elements for understanding a task. The intrinsic load for simple tasks may require a small number of elements that may be understood apart from one another; more complex tasks may require understanding of, and interaction among, several elements. Paas, Renkl, and Sweller (2003) provided the example of learning the assignments of the set of 12 function keys on a typical QWERTY computer keyboard. Each element (i.e., an individual function key) may be understood apart from any other. By contrast, learning how to edit a photo on a computer requires several elements (e.g., changing color tones, darkness, contrast), all of which must be understood interactively to complete the task. The demands on working memory imposed by the intrinsic load of high-complexity learning tasks are greater than those imposed by simpler tasks. Decreasing the intrinsic load of a learning task results in a simpler task.

Based on Miller's (1956) assertion that working memory is an inherent human limitation, assessment tasks with greater intrinsic load may not only require the test-taker to memorize the essential elements of the task, but also to integrate them. Extraneous load, when required by an assessment, is preliminary to (or concurrent with) attending to the test or test item, organizing the material into an existing structure, and integrating the material with existing knowledge. In essence, when extraneous load is included in a test item, it interferes with the test-taker's engagement with the item by demanding the use of working memory for elements that are not essential for demonstrating what he or she knows.

The results of CLT experiments indicated that cognitive load appeared to be lower when essential information disaggregated across two or more sources was integrated (e.g., textual statements describing a diagram were embedded in the diagram itself). Based on lower test scores and longer processing time for learners who were given the "split-source" diagrams, the authors concluded that "presentation techniques frequently result in high levels of extraneous cognitive load that influence the degree to which learning can be facilitated….For this reason…examples that require learners to mentally integrate multiple sources of information are ineffective" (Chandler & Sweller, 1991, p.295). As such, the predominant implications for instructional practice pertained to the integration of graphics and visual representations with corresponding textual concomitants to reduce extraneous load.

Chandler and Sweller (1996) described two negative effects that may result from the improper structuring of multimedia material (e.g., visuals and text). The first is the *split attention effect*, whereby unintegrated split-source information in the presentation of material forces the learner to integrate the information to learn. When one source of information contains all that is necessary to convey the material, the authors suggested the other source of information should be eliminated entirely to prevent the *redundancy effect*, whereby learners are distracted and bogged-down with excessive material.

To the extent the cognitive demands of an assessment are intrinsic to the target constructs, inferences made from test results are likely to represent the person's actual competence on the constructs. Extraneous load demands by an assessment item interfere with the test-taker's capacity to respond (i.e., demonstrate performance on the target construct) and should be eliminated from the assessment process. Further, germane load, while enhancing learning at the instructional level, should be considered for elimination as well: unless an assessment task has the dual purpose of both instruction and assessment, the items on a test should demand only those cognitive resources intrinsic to the target constructs they are intended to measure. Indeed, the addition of germane load to an assessment task may represent an increase in the depth of knowledge of an item if it requires additional elements or interactivity among elements. Depth of knowledge is the cognitive complexity of an item, and is related to the number of steps necessary for its completion. For example, an item may be designed to test recall of a particular concept, but the

item may be written in a manner that requires the test-taker to demonstrate understanding of the context of the question beyond simply recognizing or recalling the intended construct, thus raising the depth of knowledge level of the item beyond its original purpose. Thus, the decision to include or exclude germane load from assessment tasks should be made deliberately.

Clark, Nguyen and Sweller (2006) synthesized the CLT research and generated a set of 29 guidelines for maximizing efficiency in learning. The majority of the recommendations focus on reducing redundancy, eliminating nonessential information from text and visuals, and integrating information from dual sources. There are also a number of cautionary considerations when using audio to supplement instruction. These 29 guidelines have some significant value for individuals writing test items that efficiently focus on a target construct with minimal or no extraneous material.

Item-Writing Research. More than a quarter-century prior to the inception of legislation permitting the AA-MAS for proficiency reporting for students with IEPs, Beattie, Grise, and Algozzine (1982, 1983) conducted experimental work on several test design features, many of which subsequently have been integrated in the majority of current large-scale assessments. For instance, Beattie et al. used format changes including the use of unjustified text for reading comprehension passages, placing passages in shaded boxes to set them apart from other text, including examples at the beginning of each new item section, adding arrow and stop-sign icons to the corners of test pages, and including response bubbles in the test booklet rather than using a separate answer sheet. Results across two studies of students identified with a learning disability ($N = 345$ students in grade 3 and $N = 350$ students in grade 5) indicated the modifications increased students' scores without altering the target construct of the test.

Haladyna, Downing, and Rodriguez (2003) developed a taxonomy of guidelines for writing multiple-choice items based on research and their professional experiences. Of particular relevance in this taxonomy are the various guidelines on writing answer choices (e.g., balancing the *key*, or correct response, with the *distractors*, or incorrect responses). Additionally, the authors recommended avoiding of negative questions (i.e., those using "not" or "except"). Further, based on a meta-analysis of over 80 years of research on item development, Rodriguez (2005) concluded that three answer choices are optimal for multiple-choice items. The author indicated that reducing items from four or five answer choices to three tends to result in nonsignificant or positive effects on the discriminatory power of items, nonsignificant changes in item difficulty, increased reliability of scores and, ultimately, a positive effect on the subsequent validity of inferences from results. As applied to the development or modification of tests with a focus on accessibility, Rodriguez' conclusion suggests best practice is to reduce the number of response options of multiple-choice items to three *when it is feasible to do so*.

Decisions about which responses to eliminate are important ones, and must be made in a planful, consistent way. Rodriguez recommended eliminating nonsensical or absurd responses. When possible, these types of responses should be eliminated first, in favor of responses that are relevant. He also recommended eliminating the least-selected response or the least plausible response. The relevant item-writing guideline is that distractors should be plausible, and equally so if possible. Although it may be intuitive to target the most-selected distractor for elimination, evidence and theory suggests that this dramatically reduces the quality of item functioning and test score reliability. Elimination of the most-selected distractor will reduce the difficulty of the item, but it will also reduce the item discrimination and score reliability. However, the goal of item modification is to increase accessibility and improve measurement.

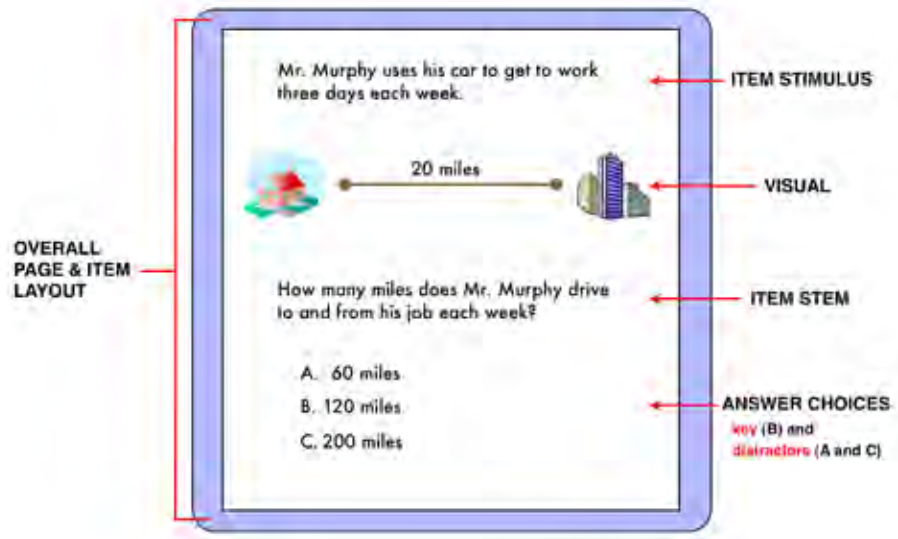
From the experience of the studies described here, the elimination of a distractor is not a simple task. For example, in mathematics items, there is a natural balance that can be achieved through four options: two negative and two positive options, two even and two odd options. Balance in the options is an important goal to achieve so not to provide clues to the correct option. In many cases, it may be easier to edit the remaining distractors to avoid imbalance and cluing. Nevertheless, the research evidence from item writing in many contexts unanimously supports a new standard of three-option items.

Using the ARM to Improve Items

The *Accessibility Rating Matrix* (*ARM*) consists of two scoring rubrics: the Item Analysis and the Overall Analysis. After writing the item number on the *ARM* Record Form, the rater begins by using the Item Analysis rubric to evaluate the accessibility of the item according to five basic elements of a multiple-choice test item (see Figure 1): (a) the item passage or stimulus, (b) the item stem, (c) visuals, (d) answer choices, and (e) the page or item layout. It should be noted that while individual test items may or may not include each of these elements, the *ARM* is adaptable to most current assessment item formats. For the purposes of rating items using the *ARM*, the passage and stimulus are rated separately since it is common for multiple items to be connected to the same passage, with each individual item containing its own stimulus and stem.

Using the Item Analysis rubric, the rater determines the accessibility level of the item on a 4-point scale (see Table 1). For any item element that is rated less than 4 (accessible for nearly all test-takers) the rater selects modifications that are likely to improve the accessibility of the item. After rating the individual item elements, the rater reviews the Item Analysis ratings and uses the Overall Analysis rubric to record overall holistic accessibility rating for the item.

Figure 1. Anatomy of a Multiple-choice Item



Levels of accessibility on the ARM are based on the extent to which an item is determined to be optimally accessible for a given portion of the intended test-taker population, according to accessibility theory and based on universal design and cognitive load theoretical principles (see Table 1). The highest accessibility level refers to an item that poses no access barriers for 95-99% of the test-taker population. Results of validity studies indicate the content of the ARM is valid for the purpose of measuring test item accessibility and that expert raters can be trained to score item accessibility with a high degree of reliability (Elliott et al., 2010; Kettler et al., 2011).

Table 1. Test Item Accessibility Levels

Level	Description	Heuristic
4	*Maximally Accessible for Nearly All Test-Takers*	Optimally accessible for between 95-99% of the population
3	*Maximally Accessible for Most Test-Takers*	Optimally accessible for between 90-95% of the population
2	*Maximally Accessible for Some Test-Takers*	Optimally accessible for between 85-90% of the population
1	*Inaccessible for Many Test-Takers*	Optimally accessible for less than 85% of Test-Takers

To use the *ARM* correctly, several steps are taken to ensure the reliability of the modification and evaluation process. Information about the original items is collected; namely, the item modification team utilizes descriptive and psychometric data for each item, including the target construct, performance indicator, strand, depth of knowledge, difficulty, discrimination, distractor functioning, and response selection frequency. Item modification and evaluation procedures

should mirror the collaborative approach used by item-writing teams across several states, and a rigorous inter-rater agreement procedure should be used to ensure consistent ratings.

Lessons Learned about Designing Accessible Test Items

The combination of the research noted above, the development of the *ARM* (Beddow et al., 2009), and the collaborative process for generating or revising test items for the CMAADI project, yielded a number of recommendations about developing accessible test items. These recommendations are:

Passage/Item Stimulus. The length of text is an essential accessibility factor for the Passage and Item Stimulus elements. Passages and stimuli must contain enough words to communicate the message or present essential information, and should be sufficiently long to provide material for a set of items. If a passage or stimulus is too long, however, readers will be more likely to miss sections, forget details, or skip the element altogether. It is desirable, therefore, that passages and stimuli contain the minimal number of words, written as plainly as possible, to permit the maximum number of test-takers to respond to the item. Accessible passages should not demand additional memory or reading load apart from those required to demonstrate knowledge of the target construct.

One challenge for test developers is the desire to create accessible test items that contain "real-world" application problems. For instance, many passages contain abridged versions of copy written publications that cannot easily be altered to reduce reading load. Likewise, mathematics and science items often require the application of conceptual knowledge to solve problems or demonstrate knowledge. Typically, these items contain more text and a higher degree of complexity than other items. Test developers should be aware that the potential is high for application problems such as these to contain barriers to accessibility due to extraneous cognitive load.

Item Stem. The item stem typically contains the question or directive for an item and should be written as directly as possible to permit test-takers to understand what is required. An unclear item stem may preclude a test-taker from demonstrating what he or she knows even if the person has learned the tested content. To facilitate the identification of the question, item stems should be distinguished from item stimuli.

Visuals. According to the cognitive theory of multimedia learning (e.g., Mayer & Moreno, 2003), visuals can be useful for communicating information in a concise manner, but they also tend to be confusing and, if designed or used improperly, may actually increase the extraneous cognitive demands of learning tasks. The application of this theory to testing suggests test developers should use caution when considering the addition of a visual to an existing item. Ideally, any included visuals are necessary for responding to the item (rather than being included for ancillary reasons such as improving test-taker interest or motivation). Indeed, many items, particularly

in mathematics and science content domains, require visuals to present essential information. From an accessibility standpoint, it is critical all visuals depict the intended image(s) as simply and clearly as possible, with no extraneous text or information.

Answer Choices. Factors that commonly reduce the accessibility of response options are the use of implausible, absurd, or unnecessary distractors, or unbalanced options. For example, if the choices are (a) *Jim,* (b) *Sue,* (c) *Reginald,* (d) *Mary,* and if option C is the correct answer, the other names should be closely matched in terms of their length; likewise choices in Mathematics or Science items should be reviewed to be sure that one answer does not stand apart from the others. As with the other item elements, answer choices should be minimal in length and written as simply as possible.

It is critical that test developers ensure only one option is correct; indeed, if a strong rationale can be made that one of the distractors may be a correct response, then some test-takers who know the tested content may subsequently be marked incorrect for the item. This is an accessibility issue insofar as the item may actually measure the extent to which the test-taker "overthinks" the item, or may test a construct referred to as "test-wiseness," or the degree to which students are able to infer what the test developer intended, as opposed to simply responding based on content knowledge or skills. Psychometric data (e.g., point-biserial statistics) likely will reveal the existence of a set of answer choices with multiple keys, but field test items should be reviewed carefully to avoid this. Raters' alternative rationales may not be evident in all cases, but from an accessibility standpoint, it behooves item writers to attend to these items with this potential issue in mind.

Page/Item Layout. The layout of items on a page, or—if necessary—across pages, is also an important aspect of accessibility. For optimal accessibility, the entire item—including relevant passages, visuals, or stimuli—should be presented on one page. As alluded to previously, this is based on *representational holding,* part of the cognitive theory of multimedia learning whereby a learner must retain a certain amount of information across a page or screen before integrating it with other required information that is necessary for responding (Mayer & Moreno, 2003). To the extent the necessary information for an item is spread across multiple pages, the accessibility of the item is compromised for some test-takers.

Notwithstanding the items used in the CMAADI study were delivered on paper-and-pencil-based test forms, bear in mind the term *page* is used here to refer both to the page containing an item in paper-and-pencil tests, as well as the screen on which an item is presented on computer-based tests. In both cases, representational holding can be an issue. On a computer, scrolling up or down to reveal a portion of text that is hidden can cause the same sorts of accessibility problems as turning a page to find a formula or read the remainder of a passage. While it often is difficult to ensure a passage or common stimulus with its entire item set is presented on a single page,

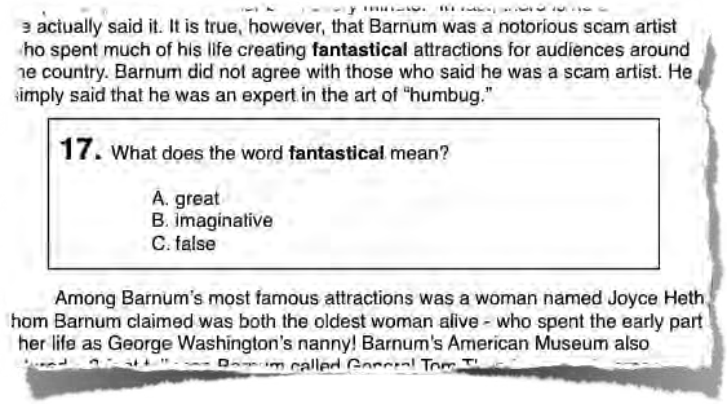
nevertheless, the layout of item and passage sets should be designed with caution to reduce the need for turning the page to respond to an item. For a similar reason, visuals that are necessary for responding should be integrated with the other item elements, rather than placed off to the side.

Computer-based tests hold promise for improving the accessibility of many tests, not only because of the potential for individualization across test-takers in ways that do not reduce the validity of subsequent test score inferences, but also because layouts, fonts, colors, contrast, and other item features can be adjusted with accessibility in mind. A testing interface can be designed, for example, whereby a single click can bring up a test item in a frame or window, so the relevant item elements remain unobscured and available for perusal while responding to the item.

One consistent concern with the item layout noted by the CMAADI item-writing team involved the use of blocks of text contained within bordered boxes. The use of bordered boxes, coupled with the use of borders for the items themselves, caused many items to appear cluttered and likely would distract some test-takers. Similarly, item stems were combined with item stimuli in most cases. It should be noted that for many items, the suggestion to increase white space referred specifically to this issue: the accessibility of many items is likely to improve if item stems are placed beneath item stimuli, with space between the elements. In response to both concerns, the team recommends using consistent item formatting that facilitates responding. Specifically, accessibility may improve if question stems are highlighted by using larger fonts and consistent placement, distinguishing them from other item text.

Figure 2 contains a hypothetical grade 7 language arts item that was modified. The original item followed a long passage, and was presented with several other items related to the passage. The items could not be presented on the same page as the passage due to space limitations, forcing a reader to flip back and forth when referring to the passage. The team suggested the item be placed in proximity to the relevant passage text to facilitate responding. This change also permitted the length of the item text to be shortened (i.e., the directions and excerpted sentence were eliminated). Additionally, bold font was used for the word **fantastical** in the passage to facilitate retrieval of the relevant sentence to ascertain context. The original item contained four response options. Based on Rodriguez (2005), one response option was eliminated (in this case, the least-selected and least-plausible option). Finally, a black border was placed around the item, to separate it from the passage text.

Figure 2. Example Item in Modified Form


e actually said it. It is true, however, that Barnum was a notorious scam artist
ho spent much of his life creating **fantastical** attractions for audiences around
ne country. Barnum did not agree with those who said he was a scam artist. He
imply said that he was an expert in the art of "humbug."

17. What does the word **fantastical** mean?

A. great
B. imaginative
C. false

Among Barnum's most famous attractions was a woman named Joyce Heth
hom Barnum claimed was both the oldest woman alive - who spent the early part
her life as George Washington's nanny! Barnum's American Museum also

Based on theory and research regarding item writing, modification, and accessibility, the *ARM* was created to evaluate current items and provide suggestions for improvements. Such improvements were then evaluated using pilot studies that focused on psychometrics of items and tests in both original and modified conditions, the focus of the next section.

Key Psychometric Indicators for Modified Items

The psychometric indicators for item and test quality apply equally to items in original and modified format. These indicators include the classic psychometric indices of item difficulty and discrimination, as well as test score reliability and the validity of inferences from scores. In addition, an important element useful in the case of item quality is distractor functioning. The review of psychometric considerations for alternate assessments by Rodriguez (2009), which arose from early applications of the TAMI, is useful in this context.

Item Indices

How difficult should the item be? This question is always relative; relative to the intended audience and relative to other items. Psychometricians, however, agree that difficulty should be a function of ability and not other student characteristics like gender, ethnicity, socio-economic, disability, or language status (potential construct-irrelevant sources of variance). We typically evaluate item functioning across groups through measures of DIF (differential item functioning), looking at probability of correct response conditioned on ability. This typically requires sufficient samples in each group to result in stable estimates of DIF.

In the context of item modifications, or experiments in item format (items with or without certain characteristics), it is more typical to monitor item difficulty and discrimination, and in the context of item response theory models (IRT), item location (difficulty), discrimination (in

some models), and item fit. To evaluate the appropriateness of item difficulties, the purpose of the test is the criterion. For tests of minimum competency (mastery), items will generally be easier; whereas for tests of rigorous or high standards, items will generally be more difficult. Test score reliability is generally maximized when items have a difficulty near the mid-point between chance score (e.g., .25 in a 4-option item) and 1.0 (100% correct responses). However, this common principle may be limited in most contexts where distractors are not equally plausible or equally chosen. In any case, the difficulty should be determined by content and cognitive demand, based on a clear definition of the construct given the item blueprint – typically determined through policy-based processes and usually less informed through psychometric criteria.

Unlike the ambiguity regarding appropriate item difficulty levels, item discrimination has a direct relation to test score quality (reliability and validity), and so criteria are easier to set. In any case, it is generally true that the better the item discrimination, the stronger the item and resulting test scores. We see test score reliabilities begin to degrade as item discrimination (item-total correlations) falls below .20 and noticeable improvements in score reliabilities when discrimination rises above .30.

Similarly, we want every distractor (incorrect option) to contribute to the functioning of the item. To be functioning well, each distractor (equally plausible distractors) should be selected by approximately an equal number of test takers and each distractor should be negatively correlated with the total score (ideally less than -.20). This is because the students with lower scores overall should be selecting the distractors. So, for example, if a 3-option item has a difficulty of .64 (64% correct), then 36% responded incorrectly—so the two distractors should be selected by approximately 18% each. The item should have an item-total correlation greater than +.30 and the distractors should each have approximately similar distractor-total correlations less than -.20.

When conducting experimental studies of item formats, where the modified format is intended to improve the functioning of the items, these item-level indicators should improve. In studies of item accessibility, we hope to see improvements in these indicators for all students, but more so for student with disabilities, the audience for which accessibility is generally limited.

Test Score Quality

In large part, the reason we are concerned with item functioning is because it contributes directly to test score reliability and validity. Rodriguez (2009) encouraged test designers concerned with accessibility to think carefully about their intended hypotheses and inferences regarding reliability and validity.

Coefficient alpha, the most commonly reported index of reliability, assumes essentially tau-equivalent measurement, an assumption regarding the nature of item true scores and error scores. These assumptions are rarely tested and rarely met. When item variances differ a great

deal (suggesting items are measured on different scales), when there are fewer items, or when there are multiple response formats, essential tau-equivalence is difficult to achieve. A more appropriate measurement model is generally the congeneric model, where items measure a consistent construct but with different scales and precision. An estimate of reliability based on a measurement model that fits the measure is needed, as reviewed by Graham (2006). Cronbach (2004), to whom the alpha coefficient is often inappropriately attributed, also argued that it is a weak index of score reliability, particularly given its oversensitivity to group variability and number of items. A more generalized approach to estimating measurement errors is found in Generalizability Theory (Brennan, 1992).

Test score validity continues to fuel a healthy debate in the measurement community, not in terms of its importance, but in terms of its conceptualization. The current *Standards for Educational and Psychological Measurement* (AERA, APA, & NCME, 1999) defines validity generally as the extent to which evidence supports the intended inferences and uses of test scores, including multiple sources of evidence. Kane (2002) has refined this substantially to focus on the intended claims from test scores, as well as the inferences and assumptions implied by those claims, and has suggested that evidence be gathered to support the argument inherent in test score interpretation and use.

In the broadest perspective on validity, all evidence that is gathered to inform score interpretation is useful, including item functioning, test score reliability, and other standard forms of validity evidence including content evidence, response processes, internal structure of the measure, and relations to other variables (as suggested by the *Standards*). But the evidence most needed includes the sources most directly addressing the immediate test score inferences and uses. For tests of academic achievement, the primary inference is typically about content-related knowledge, skills, and abilities. Kane (2006) presented a strong model of validation supporting the layers of inferences from an observed score to the target domain, including such sources of evidence as content coverage and sampling, item functioning, response processes, internal structure, and relations to other variables. In future item format studies, the interpretive argument should be clearly delineated to facilitate a productive and useful validity argument and identification of critical validity evidence.

These psychometric principles, item indices and test score quality indices, were used to monitor the quality of CMAADI instruments and impact of item modifications. Some of these results are presented in the summary of item modification studies.

The Arizona Item Modification Studies

Prior to the CMAADI experimental studies of item modifications, lessons learned from the CAAVES study (Kettler et al., 2011) were reviewed to maximize the impact of modifications. Common modifications in the CAAVES study included removal of the least functioning distractor, language simplification, addition of graphics or visual supports, increased white space, and reorganization of item layout. A key modification that set the CMAADI studies apart from other experimental research in this area was the embedding of questions within their connected passages, a change that was intended to reduce the working memory load for students completing the reading tests. Consistent with the TAMI-ARM, these modifications were intended to reduce cognitive load, improve item writing consistently with item writing guidelines, and maximize accessibility. The CMAADI design built upon the lessons learned from specific items, the impact of the modifications, and feedback from students.

CMAADI studies of item accessibility continued where the CAAVES studies left off. CMAADI included items from the Arizona Instrument to Measure Standards (AIMS) that were reviewed and rated with the TAMI-ARM, modified, and experimentally administered to groups of students with disabilities (SWDs) and students without disabilities (SWODs). Two studies were conducted, including 294 students in grades 7 and 10 from four schools in 2009-10 (pilot study) and 240 students grade 7 from 10 schools in two districts in 2010-11 (field-test study).

Pilot Study. Approximately one month following the regular AIMS administration, students from four schools participated in a project involving 15 reading items and 20 mathematics items that were modified from the recent AIMS exams in 7th and 10th grades. The grade 10 sample of students eligible for the AA-MAS was too small for reliable results. Results for the modified forms administered to students in grade 7 were promising, based on 46 SWDs and 106 SWODs.

In mathematics, for SWDs, the items became slightly easier on average by .13 compared to .07 for SWODs. The average item-total correlation (item discrimination) was lower for SWODs but remained the same for SWDs (.22). In reading, again the items became slightly more easy for SWDs (by .17) compared to SWODs (.10); whereas the item discrimination dropped again for SWODs but increased for SWDs (by .11). Items became modestly less difficult (by no more than .17) for all students but much more so for SWDs, supporting the differential boost model (Kettler et al., 2011). And for SWDs, the items became noticeably more discriminating in reading, improving their measurement properties. Table 2 depicts the difficulty and item-total correlations for students across eligibility groups and conditions. The two columns farthest to the right indicate change in difficulty and discrimination, respectively, in enhanced versus original conditions.

Table 2. Pilot P-values and Item-Total Correlations across Groups and Conditions

Group	Original AIMS		Enhanced		Difference	
	Difficulty	Item-Total	Difficulty	Item-Total	Difficulty	Item-Total
Mathematics						
SWODs	.66	.30	.73	.18	.07	-.13
SWDs	.42	.22	.55	.22	.13	.01
Reading						
SWODs	.53	.27	.63	.19	.10	-.08
SWDs	.30	.01	.47	.12	.17	.11

Field Test. The following year, a field-test of modified items for the 7th grade mathematics AIMS test was conducted with a more diverse sample of 183 SWODs and 57 SWDs in 10 schools across two districts. Two forms were administered containing 34 items in original format and 34 items in modified format, with alternating order across forms. The forms yielded coefficient alphas of .85 and .90 for the items in original format and.86 and .92 for corresponding items in modified format.

In this study, more attention was given to the nature of the item modifications. Most commonly, across the 68 items in total, item modifications included an increase in white space within and between items (59%), a simplified stem (41%), an increase in the size of the visual associated with the item (26%), isolation of the stimulus (19%), elimination of the stimulus (16%), and simplification of the stimulus (15%). All items were also reduced from 4- to 3-options, and for 25% of the items this was the only modification. On average, two modifications were made to each item (with as many as four on a given item), in addition to distractor elimination. Unfortunately, no modifications (absence versus presence of a modification) explained changes in item difficulties or discriminations at a significant level. We suspect that this is due to the non-random assignment of modifications. Modifications were made on items on an as-needed basis and were designed to serve the purpose of increasing accessibility as a package of modifications unique to each item (Kettler, 2011).

Because of the smaller numbers of SWDs per form and item set, analyses comparing performance of SWDs versus SWODs are not reported. Instead, we focused on the role of item modification on item performance. The most robust findings indicate that most items became easier in modified format (about 8% more on average selected the correct option), whereas the item-total correlation changed insignificantly (M= +.01). However, the item-total correlations for the distractors became more negatively discriminating (M= -.06), indicating that the distractors (incorrect options) were more likely to be chosen by the poorer performing students, as they should be. Among the functioning distractors (n=64; those with at least 10% selection rate in

original format), 63% improved in their functioning (became more negatively discriminating). Among all 136 distractors (68 items x 2 distractors), 55% improved in functioning.

Lessons Learned Regarding the Psychometric Analyses of Modified Items

This body of work constitutes experimental research on packages of modifications guided by the TAMI to improve item accessibility and more generally to improve item functioning. Several lessons are clear from the item analysis and psychometric work in these studies. Kettler (2011) summarized this work in an apt phrase: *less is often more*. We recognize that improving item and test accessibility is a process of reducing (focusing) the content of an item, as a principle item writing guideline is to base each item on one type of content and cognitive demand (Haladyna, Downing, & Rodriguez, 2002). This typically results in fewer words, less complexity, less cognitive load, less options, and subsequently less construct-irrelevant features.

First, modifications (packages of modifications) suggested by accessibility review of standard MC items lead to higher scores for both SWDs and SWODs, whereas scores for SWDs were improved at a higher level. This results in a reduction in the performance difference between the two groups. This differential boost has been recognized as improving item and test accessibility and thus improving measurement quality of the measure.

Second, in a broad review of results, we have been more successful in improving accessibility and measurement quality of reading test items than mathematics items. In part, there is often more language-related elements of reading items that can be modified (language complexity, excess verbiage, complex options). In mathematics, the structure of the items is often fixed and more difficult to modify without significant rewriting of the item altogether.

In both studies, modifications were identified through the evaluation of item accessibility guided by the TAMI-ARM. In a number of cases, the recommended modifications were not agreeable to the entire research team, including content, measurement, and special education specialists and state assessment personnel. In some cases, the suggested modification was seen as possibly changing the depth of knowledge or the content focus or even making the item more complex. Some modifications were not employed because of fears of changing the item too much. In large part, we believe that the modifications made to some items were not sufficient to improve accessibility where TAMI-ARM evaluation of modified items did not result in significant improvements in accessibility. Some items are so structurally difficult in their original form that no modifications could improve accessibility, and some modifications actually degrade the measurement quality of a few items. In these cases, it may be advisable to simply rewrite the entire item. We believe that the TAMI and the TAMI-ARM should be considered as appropriate tools for item development.

Measuring Access to the General Curriculum: Initial Work

The inclusion of students with disabilities in test-based accountability is intended to provide reliable test scores that permit valid inferences about the extent to which students have progressed in the general curriculum and the extent to which teachers and schools can be held accountable for the students' learning. *Access to the general curriculum* lies at the heart of federal legislation for students with disabilities (Roach et al., 2009) and represents a necessary condition for the validity of these test score inferences. Ideally, students with disabilities access the general curriculum through a teacher's high-quality instruction that offers them the *opportunity to learn* the state's intended curriculum, which is subsequently sampled by the state's large-scale assessment (Kurz & Elliott, 2011). Student achievement of this intended curriculum is subsequently tested via *accessible assessments* purposefully designed to provide students with disabilities optimal access to the measured constructs without introducing construct-irrelevant variance related to extraneous test features. At the time of the test event, students with disabilities should also be provided with appropriate *testing accommodations* that ameliorate the effects of disability-related characteristics, which limit students' access to demonstrate proficiency in the tested domains. 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 empirical evidence (e.g., Elliott, 2009; Kurz, Elliott, Wehby, & Smithson, 2009; Wehby, Symons, & Canale, 1998; Wehmeyer, Lattin, Lapp-Rincker, & Agran, 2003). Figure 3 highlights the various access barriers to the general curriculum for students with disabilities. In fact, the cognitive labs conducted under CMAADI highlighted that numerous students had not been exposed to the content assessed by certain test items. To clarify the extent to which students simply could not remember having been taught the content versus not having had the opportunity to learn the content, we examined students' opportunity to learn the Indiana state content standards via an online teacher log called the *Instructional Learning Opportunities Guidance System* (MyiLOGS; Kurz, Elliott, & Shrago, 2009) in a supplemental study.

Figure 3. Access Barriers to the General Curriculum for Students with Disabilities

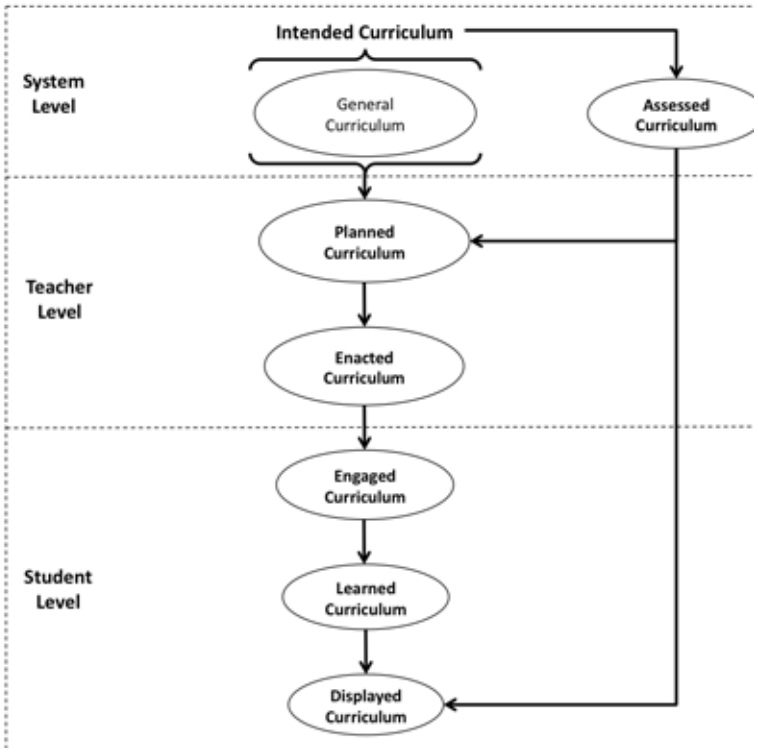
Initial work focused on the relation between access to the general curriculum and alternate assessment performance, resulting in a model based on students' current grade level, teacher reports of students' curricular access, percentage of academic-focused IEP goals, and time spent in general education settings that accounted for 41% of the variance in a latent factor of student performance (Roach & Elliott, 2006). Teacher-reported coverage of general curriculum content was the best predictor in the model (.41) accounting for 23% in the variance in student performance. Kurz et al. (2010) examined students' opportunity to learn the intended curriculum via the *Surveys of Enacted Curriculum* (SEC) (Porter & Smithson, 2001) alignment methodology. The relation between alignment and student achievement averages at the classroom level was examined for general and special education teachers. The authors hereby used the SEC's alignment index (AI) between the enacted and intended curriculum as a proxy for OTL (see Kurz et al., 2010 for further details). The results indicated that the content of instruction delivered by general and special education teachers was not highly aligned with the intended curriculum and did not differ significantly between the two groups. The correlation between AI and (class averages of) student achievement was .64 ($p < .05$). When general and special education teachers were examined separately, the correlation between alignment and achievement remained significant only for the special education group with .77 ($p < .05$).

Alignment, however, represents a very limited proxy for OTL (Kurz, 2011). Expectations for what students should know and be able to do must be articulated across all levels of the educational environment. Academic standards that delineate subject- and grade-specific content and performance objectives are typically used to this end. Collectively, these standards designate the *intended curriculum*, which can be general and applicable to all students or specific to certain subgroups as well as individual students. As such, the intended curriculum represents the normative target for all other curricula. Based on this premise, Kurz (2011) developed a framework delineating key curricula at the system, teacher, and student level. Under the *Intended Curriculum Model* (ICM), all subordinate curricula must be driven by, and reflect, the intended academic standards to the greatest extent possible to ensure consistently aligned educational inputs, processes, and outcomes.

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 Individualized Education Program (IEP) specifies the extent to which the general curriculum (of all students) is applicable to the individual student's intended curriculum. In addition, it may establish educational goals that are not part of the general curriculum. As such, the intended curriculum for students with disabilities is dually determined and, depending on the degree to which their intended curriculum differs from the intended (general) curriculum of all students, different types of assessed curricula may have to be developed to ensure proper alignment (see Figure 4). According to the ICM, 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 (Kurz & Elliott, 2011).

Figure 4. The Intended Curriculum Model for Special Education



[From the *Handbook of Accessible Achievement Tests for All Students: Bridging the Gaps Between Research, Practice, and Policy* (p. 104), by A. Kurz, 2011, New York: Springer. Reprinted with permission.]

Are the concepts of alignment between the enacted and intended curriculum and students' opportunity to learn the intended curriculum interchangeable? An answer to this question depends on the constraints of the alignment method used. First, current alignment methodologies do not account for the IEP 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. Secondly, the only alignment methodology that address teachers' enacted curriculum—the SEC—establishes an alignment index at the class level. Students with disabilities, however, are supposed to receive a differentiated instruction. Classwide indices may thus ignore important instructional differences for individual students with disabilities (Kurz, Elliott, Lemons, et al.

2012). Lastly, 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—an assumption that is not warranted.

Measuring OTL with MyiLOGS

The concept of opportunity-to-learn (OTL) generally refers to schooling inputs and processes necessary for producing important student outcomes (McDonnell, 1995). Standards-based reform has required states to define these important outcomes via rigorous content and performance standards available to all students. As such, a student's *intended curriculum* is largely comprised of state-specific academic standards (Porter, 2006). Empirical associations with student achievement have supported three broad OTL research strands focused on classroom instruction, the so-called *enacted curriculum* (Kurz, 2011). Empirically supported OTL indicators of the enacted curriculum are related to *instructional time* (e.g., extent to which allocated time is used for instruction), *content coverage* (e.g., extent to which instructional content is aligned with academic standards), and *instructional quality* (e.g., extent to which empirically supported instructional practices are implemented). The concept of OTL thus can be operationalized and measured along these three dimensions of the enacted curriculum—time, content, and quality—all of which must occur in conjunction with one another whenever instruction is enacted (Kurz, 2011).

MyiLOGS is an online measure of OTL (www.myilogs.com) originally developed at Vanderbilt University by Alexander Kurz, Stephen Elliott, and Jackie Shrago (2009) as part of a U.S. Department of Education funded Enhanced Assessment Grant called the Modified Alternate Assessment Participation System (MAAPS). Kurz and Elliott have continued further research and development on MyiLOGS at Arizona State University (ASU). MyiLOGS was originally designed as an advancement over currently available OTL measures such as annual teacher surveys (Porter, 2002) and intermittent teacher logs (Rowan & Correnti, 2009). To this end, the authors developed an online software tool that allows special and general education teachers to efficiently record the planning and implementation of standards-based instruction at the class and student level on a daily basis. Teachers use the tool to document their classroom instruction along all three key dimensions of the enacted curriculum: time, content, and quality. The existing software provides teachers with an instructional calendar that features an expandable sidebar, which lists the skills that comprise the intended academic standards as well as Custom objectives (e.g., any instructional objectives not addressed by the standards) and IEP objectives. Teachers can drag and drop planned skills onto the respective calendar days and indicate the number of minutes allocated to each skill. After the lesson, teachers indicate any changes to skills and times at the class level including time not available for instruction (due to announcements, transitions, etc.). Figure 5 provides a screenshot of the calendar used to collect these data. On a subsample of days, teachers are further asked to report on additional time emphases (in minutes)

related to the academic skills listed on the calendar according to cognitive demands (e.g., recall, analyze), instructional groupings (e.g., small group, whole class), and use of evidence-based instructional practices (e.g., direct instruction, reinforcement). This detailed reporting occurs at the class and student level to allow teachers to report on instructional differences for individual students. The information logged by teachers yields key OTL indices related to (a) Instructional Time on Standards (Min/Day and %), (b) Instructional Time on Custom Objectives (Min/Day and %), (c) Non-instructional Time (Min/Day and %), (d) Content Coverage (%), and (e) three scores related to time emphasis of higher-order cognitive processes, evidence-based instructional practices, and individual/small group formats.

Figure 5. Screenshot of the MyiLOGS Instructional Calendar

Detailed information on the training teachers receive to use MyiLOGS and on technical qualities of this OTL measure are provided in detail in the chapter in this volume on the MAAPS project, which overlapped time-wise with the CMAADI study. We now focus on how MyiLOGS was used in Indiana to address fundamental questions about the instructional of student with and without disabilities who received their access to the general curriculum in the same classrooms.

The Indiana OTL Study

Two major research objectives were examined for purposes of the Indiana OTL Study: (a) describe students' opportunity to learn the general curriculum standards across various grades and subjects; and (b) evaluate the extent to which students with disabilities receive a differentiated opportunity to learn the general curriculum standards compared to their class peers.

A total of 45 general and special education teachers of students with disabilities participated in training on the use of MyiLOGS. The training included a series of performance-based assess-

ments that required teachers to log at least two written instructional scenarios via the software with 100% accuracy. During the course of the study, seven teachers dropped out of the study. Four additional teachers had to be removed from analyses due to missing data regarding their allocated class time. Several teachers logged multiple classrooms within or across subjects, which featured some of the same target students. Moreover, three classrooms were co-taught and thus comprised of a general and special education teacher. In the case of co-taught classes, both teachers were asked to confer about their instructional provisions, but the final logging responsibility remained with the general education teacher. In summary, a total of 34 general and special education teachers across multiple districts provided (a) OTL data on 19 Mathematics classes featuring 37 nested target students; and (b) OTL data on 15 English classes featuring 31 nested target students. Table 3 displays the breakdown of teachers by grade and subject area.

Table 3. Breakdown of Teachers by Grade and Subject Area

	Grade 4	Grade 6	Grade 8	Total
MA	7	6	6	**19**
ELA	7	3	5	**15**
Total	**14**	**9**	**11**	**34**

Note. MA = Mathematics; ELA = English/Language Arts.

For purposes of reporting OTL, all participants were asked to log their daily classroom instruction at the calendar level (i.e., instructional time, content coverage) and twice a week in greater detail at the classroom and student levels (i.e., instructional time, content coverage, cognitive expectations, instructional practices, grouping formats, engagement, goal attainment). (Persons interested in more details regarding MyiLOGS and its various scoring indices are referred to in chapter 3.) The procedural fidelity (PF) based on completed calendar days and detailed sample days was monitored on a bi-weekly basis. Missing calendar days or sample days were identified in a follow-up e-mail along with a prompt to complete the missing information before the next PF check. All teachers completed their missing data within the prescribed timeframe. The final instructional data set was 100% complete for all participating teachers.

The lead developer of MyiLOGS trained university personnel in the observation procedures and conducted IOA sessions. For training purposes, the trainer reviewed the MyiLOGS definitions and conventions as well as the observation protocol and subsequently conducted training sessions in actual classrooms. Observers had to obtain an overall agreement percentage of 80% or higher on two consecutive 30-minute sessions. For observation purposes, all classrooms observers (a) prerecorded the skills listed on the MyiLOGS calendar for the given day, (b) started the 1-minute interval with the bell or at the lesson's designated start time, (c) made a tally in both matrices according to the cognitive expectation and instructional practice that occupied the majority of

the time during a 1-minute interval (by skill and grouping format), and (d) kept a frequency count of discreet events such brief praise statements. At the conclusion of the observation, the observer was allowed to make time adjustments to reflect the summative duration of discreet events as well as the MyiLOGS convention of equal emphasis. The latter convention requires teachers to divide instructional minutes equally according to emphasis. For example, a teacher who allowed students to work independently for 10 minutes but concurrently provided students with individual guided feedback throughout the entire time could not log 10 minutes under each practice. Instead, the teacher must divide the instructional minutes accordingly (i.e., 5 minutes per practice). This convention constrains teachers to the allocated class time—the more skills or practices that are addressed, the less instructional time can be dedicated to each one. Accordingly, observers were allowed to make tally adjustment immediately following the observation. All teacher observations are conducted for the entire allocated class time.

For agreement purposes, cell-by-cell agreement was calculated for each matrix based on cell estimates within a 3-minute range or less. That is, two observer estimates of direct instruction at the whole class level of 20 minutes and 23 minutes respectively were counted as an agreement. Likewise, teacher and observer estimates of the Pythagorean Theorem at the Remember level of 4 minutes and 0 minutes respectively were counted as a disagreement. For each matrix, interrater agreement was calculated as the total number of agreements divided by the sum of agreements and disagreements. In addition, a combined interrater agreement percentage was calculated as the total number of agreements across both matrices divided by the sum of agreements and disagreements across both matrices. That latter index was used in establishing the training criterion (at or above 80%) and retraining criterion (below 80%) for observers.

Across sessions in Indiana, overall agreement between two independent observers ranged between 82% and 100% with an average of 98%. Across sessions, agreement between teachers and independent observers for cognitive processes per standard/objective ranged between 64% and 84% with an average of 77%. Across sessions, agreement for instructional practices per grouping format ranged between 86% and 93% with an average of 89%. Overall agreement between teachers and observers across sessions ranged between 79% and 85% with an average of 83%. These observation results indicate that the small number of teachers sampled for validity purposes (N = 4) exhibited comparable to slightly higher agreement percentages as the larger MAAPS study sample mentioned previously. Similar to the MAAPS OTL Study sample, agreement percentages for *Cognitive Processes* were consistently lower than agreement percentages for *Instructional Practices.*

Calendar-based OTL Indices. Calendar-based OTL indices were collected for every school day during the course of the study. These class-level indices included seven OTL indicators related to time and content: (a) *Instructional Time on Standards (Min/Day):* Average amount of instructional minutes spent on state standards per day; (b) *Instructional Time on Standards (%):* Average

percentage of allocated class time used for instruction on state standards; (c) *Instructional Time on Custom (Min/Day):* Average amount of instructional minutes spent on custom objectives per day; (d) *Instructional Time on Custom (%):* Average percentage of allocated class time used for custom objectives; (e) *Non-Instructional Time (Min/Day):* Average amount of non-instructional minutes per day; (f) *Non-Instructional Time (%):* Average percentage of allocated class time not used for instruction; and (g) *Content Coverage of Standards (%):* Percentage of addressed state standards. The results for the calendar-based class OTL indices for 19 math (MA) classes and 15 English Language Arts (ELA) classes are presented in Table 4, while the same indices are presented for class type, that is, General Education (10 classes) or Special Education (24 classes) in Table 5. In addition to these data tables, teachers also received a number of pie charts to compare how instructional time was used in Math and English Language Arts classes, as well as General Education and Special Education classes.

Table 4. Calendar-Based Class OTL Indices By Subject Area

OTL Index	MA (*n* = 19)		ELA (*n* = 15)	
	M	***SD***	***M***	***SD***
Logged School Days	**87**	**(7)**	**85**	**(6)**
Instructional Time on Standards (Min/Day)	27	(9)	36	(20)
Instructional Time on Standards (%)	59	(24)	59	(22)
Instructional Time on Custom (Min/Day)	17	(10)	21	(15)
Instructional Time on Custom (%)	34	(21)	33	(21)
Non-Instructional Time (Min/Day)	4	(7)	5	(7)
Non-Instructional Time (%)	6	(8)	6	(6)
Number of Standards	**56**	**(4)**	**48**	**(1)**
Content Coverage of Standards (%)	38	(23)	49	(27)

Note. MA = Mathematics; ELA = English/Language Arts.

Table 5. Calendar-Based Class OTL Indices By Class Type

OTL Index	GENED (*n* = 10)		SPED (*n* = 24)				
	M	***SD***	***M***	***SD***	***df***	***t***	***ES***
Logged School Days	**87**	**(8)**	**85**	**(6)**	32	0.66	0.28
Instructional Time on Standards (Min/Day)	37	(22)	29	(11)	32	1.40	0.46
Instructional Time on Standards (%)	50	(20)	63	(23)	32	-1.56	0.60
Instructional Time on Custom (Min/Day)	26	(12)	15	(11)	32	2.53*	0.96
Instructional Time on Custom (%)	39	(17)	31	(22)	32	0.99	0.41
Non-Instructional Time (Min/Day)	9	(11)	2	(3)	32	2.87*	0.87
Non-Instructional Time (%)	10	(11)	4	(4)	32	2.23*	0.72
Number of Standards	**54**	**(5)**	**52**	**(6)**	32	0.79	0.36
Content Coverage of Standards (%)	50	(24)	40	(25)	32	1.13	0.41

Note. GENED = General education class; SPED = Special education class; ES = Cohen's *d* effect size.

Sample-Based OTL Indices. Sample-based OTL indices at the class and student were collected in the Indiana classrooms on two random days per week. These class- and student-level indices included five OTL indicators related to instructional quality. These indicators and the method for scoring each one are:

Cognitive Process Score

a. 1.00 indicates an exclusive focus on lower order thinking skills (*Attend, Remember*).
b. 2.00 indicates an exclusive focus on higher order thinking skills (*Understand/ Apply, Analyze/Evaluate, Create*).

Instructional Practice Score

a. 1.00 indicates an exclusive focus on generic instructional practices (*Independent Practice, Other*).
b. 2.00 indicates an exclusive focus on empirically supported practices (*Direct Instruction, Visual Representations, Asked Questions, Think Aloud, Guided Feedback, Reinforcement, Assessment*).

Grouping Format Score

a. 1.00 indicates an exclusive focus on whole class instruction.
b. 2.00 indicates an exclusive focus on individual and small group instruction.

Engagement

a. 4-point scale: Not engaged (0%) = 0; Low % of time (<50%) = 1; Moderate % of time (50%-80%) = 2; High % of time (>80%) = 3.

Goal Attainment/Effort

a. 4-point scale: No effort or product observed (0%) = 0; Low effort or limited portion of work completed (<50%) = 1; Moderate effort or moderate portion of work completed (50%-80%) = 2; High effort or substantial portion of work completed (>80%) = 3.

The results from the Indiana teachers on each of these five indicators are documented in Tables 6 for MA and ELA and in Table 7 for General and Special Education classrooms.

Table 6. Sample-Day Based Class OTL Quality Indices By Subject Area

	MA (*n* = 19)		**ELA (*n* =15)**	
	M	***SD***	***M***	***SD***
Logged Sample Days	29	(3)	28	(3)
Cognitive Process Score	1.66	(0.13)	1.65	(0.20)
Instructional Practice Score	1.75	(0.09)	1.61	(0.13)
Grouping Format Score	1.25	(0.21)	1.36	(0.26)
Engagement	2.45	(0.32)	2.48	(0.33)
Goal Attainment/Effort	2.43	(0.27)	2.47	(0.30)

Note. MA = Mathematics; ELA = English/Language Arts.

Table 7. Differences in Class and Student Key OTL Indices By Classroom Type

	Class		Student				
	M	***SD***	***M***	***SD***	***df***	***t***	***ES***
General Education							
Instructional Time on Standards (Min/Day)	**35**	21	**34**	21	9	1.68	0.05
Instructional Time on Custom (Min/Day)	**26**	13	**25**	10	9	1.00	0.09
Non-Instructional Time (Min/Day)	**11**	14	**14**	16	9	-2.23*	0.20
Content Coverage of Standards (%)	**32**	17	**31**	16	9	1.72	0.06
Cognitive Process Score	**1.57**	0.20	**1.55**	0.21	9	1.31	0.10
Instructional Practice Score	**1.64**	0.13	**1.62**	0.14	9	1.66	0.15
Grouping Format Score	**1.23**	0.16	**1.22**	0.19	9	0.32	0.06
Special Education							
Instructional Time on Standards (Min/Day)	**28**	12	**27**	11	23	1.61	0.09
Instructional Time on Custom (Min/Day)	**15**	11	**15**	11	23	0.46	0.00
Non-Instructional Time (Min/Day)	**3**	4	**3**	7	23	-1.11	0.00
Content Coverage of Standards (%)	**27**	17	**26**	16	23	3.18*	0.06
Cognitive Process Score	**1.69**	0.13	**1.69**	0.14	23	0.09	0.00
Instructional Practice Score	**1.71**	0.13	**1.69**	0.13	23	1.73	0.15
Grouping Format Score	**1.32**	0.26	**1.35**	0.25	23	-1.85	0.12

Note. *p <.05; ES = Effect size measure d; General Education featured 10 students; Special Education featured 24 students.

Lessons Learned from Initial Research on Students Access to the General Curriculum

The findings based on this non-representative sample provided an initial description of students' opportunity to learn the general curriculum standards across various grades and subjects. Key observations included (a) no major differences across several key OTL indices between subject areas (Tables 4 and 5); (b) about 60% of teachers' daily allocated class time was spent on teaching the academic standards, about 30% was spent on custom skills/activities not directly related to the standards (e.g., computer time, games), and about 5% of instructional time was not used for instruction (Table 5); (c) during about 40% of the school year, teachers were able to address about 45% of the academic standards (Table 5) ; and (d) differences in OTL based on classroom type indicated statistically significant differences for *Non-Instructional Time* with a small effect size. For purposes of the second question—the extent to which students with disabilities receive a differentiated opportunity to learn the general curriculum standards compared to their class peers—the findings provided some evidence for OTL as a differentiated opportunity structure. In general education classrooms, students with disabilities received statistically significantly

more *Non-Instructional Time* compared to their classmates with a small effect size. The latter finding is consistent with findings from the MAAPS OTL Study conducted in the states of Arizona, Pennsylvania, and South Carolina (see Kurz et al., 2012).

Summary and Implications of the CMAADI Project

Given the approach to inclusive assessment and the development of an AA-MAS, we defined the term *modification* to refer to 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. When analyses indicated inferences made from the resulting test scores are valid indicators of grade-level achievement, we considered the modifications appropriate. Conversely, if analytic evidence suggested the inferences made from resulting scores were invalid indicators of grade-level achievement, the modifications were determined inappropriate. Thus, just 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 and a potential reduction in testing accommodations.

The CMAADI project was virtually all about item modifications and improving access to statewide achievement tests for all students. As noted through this chapter, we learned some lessons that we believe can help others who wish to enhance their inclusive assessment practices and research.

Inclusive Assessment Practices

To advance inclusive assessment practices for all students, but in particular those students with disabilities that result in persistent academic difficulties and poor test performance on statewide achievement tests, the CMAADI project's findings indicated that:

1. More effort is needed to support teachers in ensuring students have meaningful opportunities to learn the grade-level intended and assessed curricula. Helping teachers monitor their instructional time, content, and quality of instructional actions with a tool like MyiLOGS is a beginning step, but more needs to be done. In many cases, the students of concern who are sitting in general education classrooms will likely need 30 to 40 more days of instruction to actually get the opportunity to learn the content that is measured on state tests.

2. More work with item developers to precisely articulate the target constructs of tests and test items is an important step in facilitating the development of maximally accessible items. The reduction of construct irrelevant variance will generally result in items with fewer extraneous

words, more plausible response options, and overall less text, thus enhancing the readability of the item and reducing cognitive load. Use of tools like the TAMI and TAMI-ARM should help test item developers consistently operationalize the principles of Universal Design and create highly accessible tests that yield reliable scores and valid inferences about students' achievement in language arts, mathematics, and science.

Suggestions for Future Research

Research almost always stimulates more questions than it answers, and in the class of the CMAADI project and test accessibility, a number of issues need more research. In particular, with regard to the development and use of accessible tests so that students with disabilities have a greater chance of demonstrating what they have learned, we suggest conducting more research on (a) the tests, (b) the test takers, and (c) the interaction between the tests and the test takers.

With regard to the tests, research on packages of modifications to improve accessibility has been emphasized recently, in part due to the final regulations of the Elementary and Secondary Education Act (ESEA) (U.S. Department of Education, 2007a, 2007b), which have inspired states to make systematic improvements to their item pools. To date, studies have yielded evidence of small gains in measurement precision that may be tied to a subset of the modifications that have been studied. The results of the movement for more accessible tests are likely to become more positive as ineffective modifications are identified and removed from consideration, yielding packages of modifications that are more effective overall. For example, the current positive results with reading items and tests may be an indicator that embedding items within passages is an effective modification that can help SWODs show what they know and are able to do. However, the designs of the CMAADI pilot test and field studies have not allowed for the isolation of the effects of a single type of modification (e.g., embedding text), and the effects of packages are likely the cumulative effect of both successful and unsuccessful modifications. The majority of work on item modifications outside of the CMAADI project shares this limitation; it is simply very costly to do experimental work on each potential modification individually. Nevertheless, studies on isolated modifications, as well as replications of the current "package studies" on grade levels and content areas that have yet to be addressed, are necessary steps toward developing more accessible tests (Kettler, 2011).

Regarding test-takers, future research should also incorporate information about the eligible population and their experience during testing, both in original and modified conditions (Kettler, 2011). Examples of information that should be collected on the eligible population are indicators of working memory, reading fluency, and freedom from distractibility. The modifications made in the CMAADI studies were aimed at reducing cognitive load, as well as reading load when appropriate. These modifications were made based on the assumption that limitations in

these areas are barriers to success on typical achievement tests. In order to determine whether this assumption is true, samples of eligible students should complete brief tests in these areas.

Finally, future research should address whether modifications are making the testing experience similar for SWDs as compared to the experience of SWODs on original tests. For example, it would be helpful to record the amount of time taken to complete forms in original versus modified conditions (Kettler, 2011). Cognitive labs could also be used to determine whether modifications help students use appropriate strategies. Advances in technology might also allow studies that track eye movements during testing, yielding richer information on the testing experience. All of this research would be helpful to evaluate attempts to make test scores more comparable across groups of students.

Based on the initial findings from the CMAADI IN OTL Study and the MAAPS OTL Study, further research on the instructional provisions for students with disabilities is warranted. We specifically recommend the collection of OTL data in classrooms representative of covering the various intended curricula for students with disabilities via teacher self-report and randomly sampled classroom observations to estimate reliability and fidelity of self-report data, followed by a critical analysis of OTL in terms of instructional time, coverage of intended knowledge and skills, emphasis of cognitive demands, prevalence of evidenced-based practices, and instructional differentiation; shortcomings should be remediated by targeted profession development based on the collected OTL data. The latter point is critical for teachers' instructional improvement efforts. MyiLOGS, for example, provides instructional feedback reports that can be used by teachers to monitor the instructional provisions for their overall class as well as the extent to which they differentiate their instruction for specific students.

Resources

For readers interested in learning more about measuring opportunity to learn, we recommended the following book chapter:

- Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 99-129). New York: Springer.

This chapter reviews the concept of opportunity-to-learn (OTL) and related conceptual methodological challenges.

For readers interested in learning more about technically sound practices behind the writing and assessment of highly accessible test items that are likely to improve the measurement of what all students' know and can do, we suggest the following resources:

- Beddow, P. A., Kurz, A., & Frey, J. R. (2011). Accessibility theory: Guiding the science and practice of test item design with the test taker in mind. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests* (pp. 163-182). New York: Springer.

This chapter consists of a comprehensive discussion of accessibility theory as it applies to the development of accessible test items. The authors focus on the critical importance of addressing cognitive demand in item design and modification, and they use a sample science item three phases of enhancement to provide a detailed example of the item modification process.

- 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.

In this article, the authors examine procedures for developing, modifying, and evaluating items and tests using an evolving modification paradigm, as well as a classic reliability and validity framework. Theoretical influences are discussed, a tool that provides comprehensive guidance is introduced, Cognitive lab methodology is described, and lists of precautions, lessons learned, and questions generated are included.

- http://www.accessibletesting.com

This webpage is dedicated to the purpose of ensuring tests and test items yield scores from which inferences are valid for all students, including students identified with disabilities. A variety of tools and resources are available for download, including the Test Accessibility and Modification Inventory and the Accessibility Rating Matrix.

For readers needing a broader view of issues of test development, item analysis and scoring, we recommend the *Handbook of Test Development*. A comprehensive treatment of techniques in educational measurement and psychometrics is also found in *Educational Measurement* (4th ed.). The references for these books are:

- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Earlbaum.

- Brennan, R. L. (Ed.). (2006). *Educational measurement (4th ed).* New York, NY: American Council on Education, Macmillan.

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.

Beattie, S., Grise, P., & Algozzine, B. (1983). Effects of test modifications on the minimum competency performance of learning disabled students. *Learning Disability Quarterly, 6*, 75-77.

Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test Accessibility and Modification Inventory (TAMI).* Nashville: Vanderbilt University. Available at http://www.accessibletesting.com

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009). *TAMI Accessibility Rating Matrix.* Nashville, TN: Vanderbilt University. Available at http://www.accessibletesting.com

Brennan, R.L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8(4),* 293-332.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*(2), 151-170.

Clark, R., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load.* San Francisco, CA: Pfeiffer.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87-185.

Cronbach, L. J. (2004).My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64,* 391–418.

Elliott, S. N. (2009, November). *Barriers to optimal assessment.* Invited presentation to the Race to the Top Assessment Panel, Atlanta.

Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M., Bolt, D., & Roach, A.T. (2010). Effects of Using modified items to test students with persistent academic difficulties. *Exceptional Children, 76*(4), 475-495.

Farrington, J. (2011) Seven plus or minus two. *Performance Improvement Quarterly, 23*(4), 113–116.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement, 66,* 930–944.

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*(3), 309-344.

Kane, M. T. (2002). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2006). Content-related validity evidence. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 131-154). Mahwah, NJ: Lawrence Erlbaum.

Kettler, R. J. (2011). Effects of modification packages to improve test and item accessibility: Less is more. In S.N. Elliott, R.J. Kettler, P.A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 231-241). New York, NY: Springer.

Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (2011). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education, 24*(3), 210-234.

Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 99-129). New York: Springer.

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., Lemons, C. J., Kettler, R. J., Zigmond, N., & Kloo, A. (2012). *Opportunity-to-learn: A differentiated opportunity structure for students with disabilities in general education classrooms*. Manuscript submitted for publication.

Kurz, A., Elliott, S. N., & Shrago, J. (2009). *My instructional Learning Opportunity Guidance System (MyiLOGS)*. Nashville: Vanderbilt University.

Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2009). 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.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1361.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*, 43-52.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305-322.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63*(2), 81-97.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1-4.

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. (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.

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., & 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., 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.

Rodriguez, M. C. (2009). Psychometric considerations for alternate assessments based on modified academic achievement standards. *Peabody Journal of Education, 84*(4), 595-02.

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. (2006). *A practical reader in universal design for learning*. Boston, MA: Harvard University Press.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the Study of Instructional Improvement. *Educational Researcher, 38*(2), 120-131.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123-138.

U.S. Department of Education (2007a, April 9). Final Rule 34 CFR Parts 200 and 300: Title I —Improving the Academic Achievement of the Disadvantaged: Individuals with Disabilities Education Act (IDEA). Federal Register. 72(67), Washington DC: Author. Retrieved from http://www.ed.gov/admins/lead/account/saa.html#regulations.

U.S. Department of Education. (2007b, July 20). *Modified academic achievement standards: Non-regulatory guidance*. Washington, DC: Office of Elementary and Secondary Education, U.S. Department of Education. Retrieved from http://www.ed.gov/admins/lead/account/saa.html#regulations.

Wehby, J. H., Symons, F. J., & Canale, J. A. (1998). Teaching practices in classrooms for students with emotional and behavioral disorders: Discrepancies between recommendations and observations. *Behavioral Disorders, 24*(1), 51-56.

Wehmeyer, M. L., Lattin, D. L., Lapp-Rincker, G., & Agran, M. (2003). Access to the general curriculum of middle school students with mental retardation: An observational study. *Remedial and Special Education, 24*, 262-272.

Winter, P. C., Kopriva, R. J., Chen, C. S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences, 16*(4), 267-276.