Modifying Achievement Test Items: A Theory-Guided and Data-Based Approach for Better Measurement of What Students With Disabilities Know

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

Vanderbilt University

Federal regulations allow up to 2% of the student population of a state to achieve proficiency for adequate yearly progress by taking an alternate assessment based on modified academic achievement standards (AA-MAS). Such tests are likely to be easier, but as long as a test is considered a valid measure of grade level content, it is allowable as an AA-MAS (U.S. Department of Education, 2007b). In this article, we 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, such as principles of universal design, cognitive load theory, and item development research, are discussed. The Test Accessibility and Modification Inventory, a tool that provides systematic and comprehensive guidance to help educators modify grade-level tests, is introduced. Cognitive lab methods and experimental field tests are then described, along with examples and key findings from each, relevant to AA-MASs. The article concludes with a discussion of precautions, lessons learned, and questions generated about the methods used to improve both access and test score validity for the students who are eligible for this new alternate assessment.

This article describes a systematic process of item modification, founded in theory, individual item analysis through statistics and a modification tool, and pilot testing in a cognitive lab. Recent changes in federal policy allow a small percentage of students with disabilities to take an alternate assessment based on modified academic achievement standards (AA-MAS) to facilitate greater access to grade-level content standards. Such tests are expected to yield more accurate measures of these students’ achievement. The intended results of increased access to the standards are twofold. The first expectation is that students are able to demonstrate what they know and are able to do on a test designed to measure their achievement. The second expectation is an increased emphasis on high-quality instruction of the measured content. Although the AA-MAS policy permits a reduction in difficulty of the test content and standards following modification, making the test easier is not the primary intent of the policy. The primary intent of the AA-MAS policy is to improve measurement of abilities on grade-level content for a group of students for whom the general state assessment may not yield scores from which valid inferences can be made. The

We thank Dr. Thomas R. Kratochwill for his helpful review of this manuscript.

Correspondence should be sent to Ryan J. Kettler, 410E Wyatt Center, Peabody #59, 230 Appleton Place, Nashville, TN 37203-5721. E-mail: r.j.kettler@vanderbilt.edu
policy permits states to administer the AA-MAS with or without individualized accommodations as needed.

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

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

The purpose of both testing accommodations and modifications is to increase individuals’ access to tests. The definitions of these two access-enabling strategies, however, have been the subject of debate, in part because of their inconsistent use in the *Standards for Educational and Psychological Testing* (Standards; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) and some states’ testing guidelines. Recently, the inclusion of the term *modified* in the AA-MAS has added to the debate about these terms. An examination of the *Standards* finds that the term *modification* is cited nearly a dozen times in the index. A number of the standards refer to a modification as an accommodation. For example, in the *Standards*’ section titled “Testing Individuals with Disabilities,” it is stated,

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

At this point in time, the implementation of testing accommodations for students with disabilities is a universally endorsed policy in all states. The modification of test content, however, is inconsistent with the definition of a testing accommodation in the majority of state testing accommodation guidelines (Lazarus, Thurlow, Lail, Eisenbraun, & Kato, 2006). Accommodations are widely recognized in state testing guidelines as changes to the setting, scheduling, presentation format, or response format of an assessment (Kettler & Elliott, in press). Accommodations are made to increase the validity of inferences that can be made from a student’s scores, so that those scores can be meaningfully compared to scores of students for whom testing accommodations are not needed.

The AA-MAS policy extends the notion of access and the spirit of individualized accommodations to changes made to item content. Such changes are defined as modifications by most test developers and administrators. When item and test alterations are made (e.g., by changing the layout, reducing the length of the reading passage, adding graphic support), it is not always clear without test results whether the changes affect only access to the test or, in fact, also affect
the construct being measured and the subsequent inferences that are drawn from scores. If the content of an item or test has been changed and the scores are no longer comparable to the original construct to be measured, it has been customary to consider the alteration a modification (Koretz & Hamilton, 2006; Phillips & Camara, 2006). To ensure the modifications are acceptable under AA-MAS policy, research is needed to confirm the modified test measures the same construct(s) as the original test. Indeed, the policy assumes AA-MAS for eligible students measure the same grade-level content standards and performance objectives (constructs) as general assessments for students without disabilities. It should be noted that some modifications may result in an AA-MAS that yields scores that are not comparable to scores obtained from the original test. These modifications could still be permissible for an AA-MAS, assuming the content is commensurate with the intended grade-level of the test (Kettler, Russell, et al., 2009).

In summary, the application of the term modification to the context of these recent legislative regulations is not acceptable to all parties, and admittedly the use of these terms is subject to some controversy. There does not, however, exist a consensus on an alternative term for alterations made to item or test content that has been shown to result in tests of unchanged constructs that yield scores from which comparable inferences can be made. Given the approach we recommend for developing an AA-MAS, we have used the term modification to refer to a process by which the test developer starts with a pool of existing test items with known psychometric properties, and makes changes to the items, creating a new test with enhanced accessibility for the target population. When analyses indicate inferences made from the resulting test scores are valid indicators of grade-level achievement, the modifications are considered appropriate. Conversely, if analytic evidence suggests the inferences made from resulting scores are invalid indicators of grade-level achievement, the modifications are inappropriate. Thus, like individualized testing accommodations, modifications must be studied to determine their appropriateness. Unlike accommodations, modifications as conceptualized within AA-MAS policy are intended to afford an entire group of students access and better measurement of their achieved knowledge.

AN EVOLVING MODIFICATION PARADIGM

Over the past 2 years, we have developed a paradigm for the development and evaluation of an AA-MAS, beginning at the level of a single test item. This model has evolved out of two federally funded projects—the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES1; Elliott & Compton, 2006–2009) and the Consortium for Modified Alternate Assessment Development and Implementation2 (Elliott, Rodriguez, Roach, & Kettler, 2007–2010).

This emerging Modification Paradigm offers a conservative and systematic approach to improving the accessibility of an item, without changing the difficulty to the point that it would

---

1CAAVES was a U.S. Department of Education Enhanced Assessment grant co-directed by Elizabeth Compton and Stephen N. Elliott. Several studies on item modification were conducted within this multi-state project during 2007–2009. For more information, visit http://peabody.vanderbilt.edu/LSI_Projects/CAAVES_Project_Home.xml.

2The Consortium for Modified Alternate Assessment Development and Implementation is a U.S. Department of Education General Supervision Enhancement grant co-directed by Stephen N. Elliott, Michael C. Rodriguez, Andrew T. Roach, and Ryan J. Kettler. Several studies on item modification are being conducted within this multi-state project during 2008-2010. For more information, visit http://peabody.vanderbilt.edu/x8314.xml.
no longer be an indicator of a grade-level content objective. The first step is to evaluate the original item for accessibility. This step is accomplished by appraising the constituent elements of the item, attending specifically to barriers that could affect some of the target population. In the case of a multiple-choice item from an eighth-grade mathematics test, such as the one featured in Figure 1, the elements (or anatomy) of the item that could be modified include the item passage or stimulus, any visuals or graphics, the item stem or question, and the response options. An evaluation tool for appraising the accessibility of an item across these elements, the Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, & Elliott, 2008) is discussed later. Along with the structural inspection of the item, characteristics related to its reading load and complexity should be documented. For example, the item in Figure 1 has 42 words and the Flesh-Kincaid readability index is 4.7. The depth of knowledge of the item, an indicator of its complexity, is 2: Skills and Concepts. This depth of knowledge indicates that the item involves the application of what the test-taker knows, but that the item has only one correct answer (A). Last, item characteristics based on response data should be considered. One type of item characteristic that should be considered is the difficulty, operationalized as the percentage of persons who responded with the correct answer, which in this case was .69. Rodriguez (2009) also suggested considering the proportion of students who selected each of the incorrect answer choices, as well as the point biserial correlation between each of the response options and the total score to determine how well each response option functions.

The second step in our modification paradigm is to reduce sources of construct-irrelevant variance. Unnecessary language is removed (e.g., “working part-time in a music store” in the
4. Pete earns $100 per month. Based on the graph, how much can he spend on clothes and entertainment each month?

![Pete's Budget](image)

- A. $50.00
- B. $30.00
- C. $20.00

**FIGURE 2** A multiple-choice item in modified form.

example from Figure 1). Irregular names are changed to more easily pronounced names (e.g., “Joan” to “Pete”). In the example in Figure 1, the visual became larger and easier to read; the portion of the chart that referred to “Savings” was changed to look more like 50%, the proportion to which it corresponded. Unnecessary and potentially distracting color was removed from the visual. Among the answer choices, white space was increased to help with tracking. Answer choice “B” was eliminated on the basis of being the least selected and least plausible distractor (i.e., incorrect answer choice). The resulting item is presented in Figure 2. The item is streamlined and more accessible.

The third step in the modification paradigm is to document changes to the item. Completing the TAMI on the modified item is one way to document changes to the item. Another helpful documentation tool is a log of all of the changes that are made to each item. Also, any changes in the aforementioned item characteristics should be mentioned. In our example, the number of words was reduced to 20, the Flesh-Kincaid readability index was reduced to 4.4, and the depth of knowledge remained 2. Although various stakeholders may disagree on the extent to which the word count, readability, and depth of knowledge of an item can be changed and still remain on grade level, documentation of these changes is an important step in the modification paradigm.

The fourth step involves piloting items with students and soliciting their input about the items using both cognitive labs and postassessment interviews. Such a step is strongly encouraged by the Standards (i.e., Standard 10.3, “Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications,” p. 106).
Gathering response data during pilot testing allows test developers to identify items with features students perceive as confusing. The Standards further suggest information about student response processes and test-taking behaviors can provide evidence to support the construct validity of an assessment. In the case of AA-MAS items, student response data can provide important information about the reasons for observed differences in performance across item types (original vs. modified) and student groups (students with and without disabilities; AA-MAS eligible vs. noneligible students). The use of concurrent think-aloud protocols and follow-up questioning may allow researchers to “unpack” unexpected results. This type of evidence has been central in the development of an AA-MAS in Arizona, where test items generally include features that are intended to reduce or eliminate construct-irrelevant influences on student outcomes.

The final step in the modification paradigm is to field test the new version of the item, ideally on a sample of test takers with and without disabilities that also complete the original version of the item, and to examine its performance. Item difficulty, discrimination, and reliability are all important characteristics to measure. Again, the degree to which these indices are expected to change through the modification process may be debatable, but the direction of changes is not. We expect modified items to be less difficult, to discriminate ability better, and to have better reliability than the original versions of the items. The proportion of test takers selecting each response option, as well as the point biserial correlation for each option, should also be calculated (Rodriguez, 2009). Documenting changes on these psychometric features is essential. This five-step modification paradigm, applied at the item level, is a precursor to a reliability and validity argument at the test score level.

Modifications and Reliability

The reliability of a set of scores is affected by variance based on the construct being measured, or true-score variance, as well as by variance that is based on other factors, or error-score variance. Error-score variance can be further subdivided into that which emanates from systematic sources (e.g., test-wiseness, concentration) and that which emanates from random sources (e.g., inconsistency in item writing or administration). Appropriate test modifications for students eligible for the AA-MAS should increase test score reliability by decreasing error-score variance from random sources. For example, a group of students who are slow readers could take a mathematics test primarily consisting of long story problems that were written and administered inconsistently (e.g., sometimes using the active tense and sometimes using the passive tense, using implausible or absurd answer choices on some items, sometimes arranging vertically and sometimes horizontally, using visuals that are irrelevant on some items), and their scores would reflect a combination of mathematics ability (based on true-score variance), reading ability (based on systematic error-score variance), and variance that cannot be explained (random error-score variance). Modifying these items to be more uniform, and to only contain necessary visuals and plausible answer choices, would increase the reliability of the test scores by removing some of the random error-score variance. Modifying these items to reduce the length and complexity of the reading passages likely would not increase the reliability of the test scores. These modifications would improve the validity of the scores, the focus of the next section, by reducing systematic error-score variance. All of the modifications in this example...
would likely make eligible students’ scores more comparable to the scores of students for whom the general assessment is appropriate and, as such, would be in the spirit of the AA-MAS policy.

Two methods for estimating reliability, internal consistency and test–retest stability, are often used to evaluate achievement tests. Internal consistency, a measure of how well items in a set fit together, is commonly estimated using coefficient alpha. Coefficient alpha is related to interitem correlations, and it tends to be a high estimate of reliability. Coefficient alpha is not sensitive to systematic error based on maturation or practice between administrations, because it is calculated based on a single administration. Test–retest stability is another indicator of the reliability of a test and is typically calculated by having students complete the same or equivalent forms of a test at different points in time, and then calculating the correlation between the two sets of test scores. A key difference between test–retest stability and internal consistency is that test–retest stability is based on comparisons between test scores, rather than an average of comparisons among items and a test score. Test–retest stability is sensitive to maturation between the two times of measurement, and correlations typically are expected to be lower when intervals between testing events are longer. Test–retest designs that utilize the exact same test on both occasions are more sensitive to practice effects, whereas designs that utilize parallel forms are less sensitive to practice effects but are sensitive to any differences that may exist between the two forms.

Modifications made to items and tests for the AA-MAS are intended to preserve or increase reliability for students who would be eligible. Such improvements should be observable whether reliability is characterized as coefficient alpha or test–retest stability. Measurement of reliability is a critical first step toward establishing a validity argument for both unmodified and modified tests.

Modifications and Validity

The validity evidence for inferences drawn from AA-MAS scores is central to determining whether the instrument has value for its intended purpose (Kettler, Russell, et al., 2009). Validity is the degree to which scores from an instrument reflect the construct that they are intended to measure. Since the 1985 Test Standards (APA, 1985), a unified view of validity has prevailed, replacing a previous emphasis on construct, content, and criterion validity. This view of validity has been expanded to embrace an evidence evaluation orientation in the most recent version of the Standards (AERA, APA, & NCME, 1999). This approach to validity values evidence about the test content, response processes, relationships with other variables, internal structure, and consequences of testing.

Evidence based on content validity is the degree to which items or tests appear to represent the content that they are designed to measure. Content validity can be prioritized during the modification process by having experts involved and can be evaluated by having experts rate items for importance, making comparisons with the grade-level content standards that are supposed to be measured. Content experts can also evaluate items and tests in their original conditions, and make comparisons with the modified conditions, to determine if any change in content validity is caused by the modification process.
Validity evidence based on response processes refers to whether the students responding to an item or test interpret it in the intended way. One promising method for measuring and documenting validity evidence based on responses has been through cognitive lab methodology. Cognitive labs (also often called cognitive interviews, verbal protocols, or think-aloud studies) involve participants verbalizing their thoughts and strategies while completing tests. The application of cognitive lab methods to item and test modifications is discussed in detail later. Other methods for measuring validity based on responses include item analysis protocols, which would prompt students to list their specific considerations evoked by an item, and surveys about the testing experience. Both of these methods would involve prompting students after they had completed items, rather than during item completion. Qualitative analysis would then be used to determine whether students taking modified tests reported equal or less cognitive effort, compared to original conditions.

Validity evidence based on internal structure is the degree to which the different parts of an assessment instrument fit together as intended. This process often involves evaluating items and their relationships with each other. Factor analysis is performed to evaluate internal structure validity, indicating whether the data-based pattern of relationships among items matches the intended theoretical model of scales and subscales. Item response theory is also used to determine the match between items and theory, including whether the items are distributed in terms of difficulty in the way that the developers or modifiers intended. Both factor analysis and item response theory can be used to evaluate a finished product, or as part of an iterative development or modification process. Research comparing original with modified conditions of items and tests would determine whether the factor structure remained unchanged, and whether the items’ difficulties remained on grade level.

Validity evidence based on relationships with other variables is the degree to which scores from the newly developed or modified test converge with scores from measures of similar constructs but diverge with scores from dissimilar constructs. New or modified tests can be evaluated by measuring the correlation of scores from the test being validated with scores from a test with established properties. Validity based on relationships with other variables can also be measured by evaluating the performance of known groups (e.g., eligible students and noneligible students), to determine whether the groups differ in the expected direction. In this case, group membership becomes the other variable, against which scores from the modified items or tests can be evaluated. A special case of validity evidence based on relationships with other variables, the interaction paradigm, has been proposed as a framework for evaluating improved access based on accommodations or modifications (Kettler, Rodriguez, et al., 2009; Phillips, 1994). The interaction paradigm indicates that the gap between scores from eligible students and scores from noneligible students should be reduced through the modification process, because students who would be eligible should gain greater access as a result of the modification process.

All of these types of validity evidence can be used to evaluate a modified test as well as to compare a modified test with the original test. A comprehensive argument for the validity of an AA-MAS should be based on a combination of evidence from these sources. The goal is that measurement of achievement for grade-level content for students eligible for an AA-MAS should become more reliable, valid, and comparable to measurement of the same content for students who are not eligible and who complete the general assessment. The challenges and consequences of pursuing this goal are discussed in the next section.
Consequences of AA-MAS

A fifth form of validity evidence mentioned in the Standards, consequential validity, is relevant only to the relationship between test scores and the constructs that they are designed to reflect, to the extent that error-score variance can lead to negative consequences from test scores. Other consequences of testing (positive and negative) are also important to evaluate but are not directly related to validity. Evidence about the consequences of an AA-MAS and of other types of testing is difficult to evaluate in the short term and depends on gathering longitudinal data on a wide variety of variables. The U.S. Department of Education (2007a) identified intended consequences of implementation of the AA-MAS, including (a) providing students with disabilities access to grade-level content, (b) ensuring that the performance and progress of students with disabilities are accurately assessed, and (c) allowing teachers and schools to get credit for this progress and performance.

Elliott, Kettler, and Roach (2008) identified several potential unintended consequences, including (a) that there may be time and money for test modification and professional development, (b) many accommodations typically used with these students may not be necessary on the modified tests, and (c) lowered achievement standards may not lead to improvements in motivation or learning. This latter consequence is particularly complicated given the emphasis in the AA-MAS policy on allowing standards to be easier, so long as they are not below grade level. The term grade level is not operationally defined, notwithstanding the legitimate concern that excessively lowering achievement standards could undermine students’ motivation to learn. Another consequence of the AA-MAS policy is that modifications may yield scores that are not comparable to those from the regular assessment. This could occur even if both assessments measure the same grade-level content, because the proficiency level descriptors on which cut scores for the AA-MAS are set may not match the descriptors used for the general assessment (see Egan et al., 2009/this issue, for a thorough discussion of this topic). One final unintended consequence is the mathematical likelihood that percentages of students who achieve proficiency on the general assessments will increase, because students who consistently achieved below proficient are removed from the population completing these tests.

These intended and unintended consequences are representative of the manifold issues and concerns that can only be evaluated by tracking trends in the data over the course of several years, both preceding and following the implementation of AA-MAS. Until substantial research and evaluation data can be collected to answer these questions, we will rely more on theory to guide our practice of item and test modification.

TEST ACCESSIBILITY: THEORY AND RESEARCH

Test accessibility is the extent to which a test (and its constituent item set) permits the entirety of the test-taker population to demonstrate proficiency on the target construct(s) of the test. To the extent a test demands physical, material, or cognitive resources in excess of those required to measure the intended construct, inferences made from the scores on the test likely reflect in some part the test-taker’s incomplete access to the test.
Current federal legislation requires the application of universal design principles to the development of all statewide and district-wide achievement tests. Universal design, as defined in the Assistive Technology Act (P.L. 105-394, 1998), is

a concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly usable (without requiring assistive technologies) and products and services that are made usable with assistive technologies. (§3(17))

Although the term accessibility is not used in this definition, universal design principles as applied to assessment technology clearly are intended to address issues of access while responding to the concern raised in the Standards that the use of individualized accommodations may increase measurement error. The legislation provides the rationale for the use of universal design principles as follows:

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

In anticipation of aforementioned regulations under No Child Left Behind, permitting states to develop AA-MAS, the CAAVES project team conducted a series of studies to examine the differential effects of item modifications on the performance of students across disability and eligibility groups. To assist in the item modification process with the goal of developing more accessible tests, Beddow et al. (2008) developed the TAMI. The TAMI was designed to assist with the process by which general assessment items are modified in response to the demand for universally designed assessments to be used for the AA-MAS. Thus, the instrument has two purposes: The first is to meaningfully quantify the construct of accessibility with regard to tests and items; the second is for use as a training tool to guide item writers in the development and modification of tests and items that are highly accessible. To wit, the development of the TAMI was influenced by (a) universal design principles for learning and assessment, (b) cognitive load theory, (c) research on test and item development, and (d) guidance on web and computer accessibility.

Universal Design for Learning and Assessment

The Center for Universal Design (1997) lists seven primary aspects of universal design: (a) equitable use, (b) flexibility in use, (c) simple and intuitive use, (d) perceptible information, (e) tolerance for error, (f) low physical effort, and (g) size and space for approach and use. The last decade has seen a shift in focus toward applying universal design principles across the educational arena, called universal design for learning (e.g., see Rose & Meyer, 2006) and universal design for
Several documents from the National Center on Educational Outcomes informed the development of the TAMI, specifically with respect to universal design for assessment principles (e.g., Johnstone, Thurlow, Moore, & Altman, 2006; Thompson, Johnstone, & Thurlow, 2002).

**Cognitive Load Theory**

As conceptualized by Sweller and his colleagues, cognitive load theory (CLT; Chandler & Sweller, 1991) is a logical extension of Miller’s (1956) work on processing limitations and working memory, exemplified in Miller’s now famous “Magical Number Seven, Plus or Minus Two” article. Based on an information-processing framework, CLT has heretofore been conceived as a model for understanding the demands of learning tasks and is grounded in three primary assumptions about how the mind works: (a) the mind processes information through two separate channels: the auditory/verbal channel and the visual channel, (b) each of these channels has a limited capacity for processing information, and (c) both channels are required for most learning tasks. Proponents of CLT 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 element interactivity and the requisite cognitive resources needed to process essential information 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 requisite 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).

Notwithstanding the broad overlap between instruction and testing, CLT heretofore has had little research application to school-age students with or without special needs or to the assessment of student learning. Considering the numerous similarities between instructional tasks and the variety of tasks required in many forms of tests, the development of the TAMI (Beddow et al., 2008) focused explicitly on the degree to which cognitive load demands may impact a test-taker’s ability to demonstrate performance on assessments. Particular attention was paid to how CLT has been used to understand the cognitive demands of multimedia learning. To the extent the cognitive demands of an assessment are intrinsic to the target constructs of the assessment, 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 demonstrate performance on the target construct and should be eliminated from the assessment process.
Further, germane load, although 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. The reduction of the germane load to an assessment task may represent a decrease in the depth of knowledge of an item, if it requires additional elements or interactivity among elements. It should be noted that the ability of the test taker to integrate elements to complete a task depends not only on the materials themselves but also on his or her expertise. Thus, the decision to include or exclude germane load from assessment tasks should be made deliberately and with knowledge of the population being tested.

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. Where applicable, these guidelines were incorporated into the TAMI.

Research on Test and Item Development

The TAMI also was influenced by the collective expertise of test and item development research from a number of scholars. Specifically, the instrument includes adaptations of several recommendations from a taxonomy for writing multiple-choice items by Haladyna, Downing, and Rodriguez (2002) based on a comprehensive review of research. The authors recommend avoiding overspecific and overgeneral content in items, avoiding trick items, minimizing the use of negative stems (e.g., using the words not and except), and keeping vocabulary as simple as possible given the population being tested. Further, the authors suggest the reading load of items should be limited to that which is necessary to measure the given construct.

A large number of Haladyna et al.’s (2002) guidelines deal with the construction of the answer choices (i.e., the key and distractors). Specifically, to reduce the potential for cueing, they recommend placing choices in logical or numeric order, keeping choices homogeneous in terms of content and/or grammatical structure, keeping the length of choices about equal, ensuring all distractors are plausible (avoiding specific determiners such as always and never, and excluding absurd options), and balancing the answer choices in terms of content.

Based on a meta-analysis of more than 80 years of research on item development, Rodriguez (2005) concluded that three answer choices are optimal for multiple-choice items. Increasing the number of plausible distractors beyond two does little or nothing to improve item characteristics (i.e., difficulty, discrimination, and reliability) and adds to the cumulative reading load of a test. Accordingly, the Answer Choices category descriptors on the TAMI implicitly support the practice of reducing the number of response options from 4 to 3. It should be noted that when reducing the number of distractors from 4 to 3, Rodriguez recommends eliminating the least-selected distractor unless an implausible distractor would remain. Although eliminating a more plausible distractor likely would decrease the difficulty of an item, the effect may be due to cueing of the correct response by an implausible distractor, theoretically reducing the validity of a subsequent inference about the meaning of a correct response as it regards proficiency on the target construct.
TABLE 1

Theory-Based and Research-Supported Modifications

<table>
<thead>
<tr>
<th>Modifications to reduce unnecessary language load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rewrite to replace pronouns with proper nouns</td>
</tr>
<tr>
<td>Simplify sentence and text structure with an emphasis on clarity</td>
</tr>
<tr>
<td>Reduce vocabulary load and non-construct subject area language</td>
</tr>
<tr>
<td>Chunk and segment the text into manageable pieces</td>
</tr>
<tr>
<td>Base the item on the content it is written to measure by removing any trivial content</td>
</tr>
<tr>
<td>Minimize the amount of reading necessary by reducing excess text</td>
</tr>
<tr>
<td>Replace negatives (e.g., NOT or EXCEPT) with positive wording</td>
</tr>
<tr>
<td>Edit the items for errors in grammar, punctuation, capitalization, and spelling</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modifications to answer choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eliminate any implausible distractors until as few as three choices remain</td>
</tr>
<tr>
<td>Move a central idea that is in the item choices to the item stem</td>
</tr>
<tr>
<td>Avoid cuing for a correct or incorrect answer</td>
</tr>
<tr>
<td>Place answer choices in a logical order and make them structurally homogenous</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other general modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make items more factual, rather than opinion-based</td>
</tr>
<tr>
<td>Add white space to make tracking easier</td>
</tr>
<tr>
<td>Remove visuals that are not necessary or helpful</td>
</tr>
<tr>
<td>Format items to be read vertically</td>
</tr>
<tr>
<td>Use <strong>bold</strong> text for important words</td>
</tr>
</tbody>
</table>

Although Haladyna et al.’s (2002) and Rodriguez’s (2005) research preceded the AA-MAS policy, many of their findings are relevant to the modification process. More recently, researchers have inventoried theory-based and research-supported modifications, as helpful guides to the modification process (Hess, McDivitt, & Fincher, 2008; Kettler, Russell, et al., 2009). Table 1 contains a list of these practices to consider when modifying items.

Guidance on Web and Computer Accessibility

In Bennett’s (2001) “How the Internet Will Make Large-Scale Assessment Reinvent Itself,” he argued the advancing pervasiveness of computer technology into all areas of modern life would lead, for better or worse, to the inevitable subsumption of standardized testing. Indeed, the ever-increasing use of online testing across the range of student assessment types supports his hypothesis. Commensurate with this apparent trend, the TAMI (Beddow et al., 2008) includes a Computer-Based Test Analysis section to be used by developers of online tests to ensure adequate attention is paid to the accessibility of these tests for as many students as possible. In addition, the authors reviewed documents by the Web Accessibility Initiative (W3; 2008), which has proposed a set of guidelines to ensure that Web content is accessible to all users, including those with disabilities. According to W3, the Web Content Accessibility Guidelines are expected to become a Web standard recommendation by 2009. The guidelines focus on four key principles: *perceivable, operable, understandable*, and *robust* (for an example of a universally designed test delivery system that integrates many of the recommendations included in the TAMI, see NimbleTools at http://www.nimbletools.com).
The existing version of the TAMI is divided into two sections: Item Analysis and Computer-Based Test Analysis. The Item Analysis section is divided into categories based on the anatomy of an item. The Computer-Based Test Analysis section is divided into categories based on the key elements of a multimedia test system. Each section of the TAMI corresponds to an Accessibility worksheet for documenting ratings and suggested changes. The rater reads the descriptors in each category and records accessibility ratings for the categories on the worksheet using a 4-point scale, ranging (0 = not accessible, 1 = minimally accessible, 2 = moderately accessible, 3 = maximally accessible), yielding a summative total accessibility rating. Research on the utility of the TAMI is ongoing, but early field studies have indicated test developers and educators find the instrument useful to guide the development of accessible tests. Some of this research has involved the use of the TAMI to modify items for use in cognitive labs, where students provide feedback relevant to universal design principles, CLT, and item development.

Using Cognitive Labs to Guide Item and Test Modifications

The value of information regarding student responses and perceptions in supporting the development of assessments is addressed at multiple points in the Standards (AERA, APA, & NCME, 1999). Standard 10.3 indicates, “Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications” (p. 106). The Standards also suggest information regarding student response processes can provide evidence to support the construct validity of an assessment. “Questioning test takers about their performance strategies can yield evidence that enriches the definition of a construct . . .” (AERA, APA, & NCME, 1999, p. 12). The Standards further note the potential contribution of student response data: “Process studies involving examinees from different subgroups can assist in determining the extent to which capabilities irrelevant or ancillary to the construct may be differentially influencing (student) performance” (p. 12). We have found that student response data can provide important information about the reasons for observed differences in performance across item types (original vs. modified) and student groups (students with and without disabilities; AA-MAS eligible vs. noneligible students). The use of concurrent think-aloud protocols and follow-up questioning allows researchers to “unpack” unexpected results. For example, differential item functioning may indicate a particular item was difficult for students with identified disabilities in comparison to their peers. Recording students’ concurrent verbalizations while solving the item in question, as well as questioning students following completion of the task, may illuminate item features that contribute to the observed results. This type of evidence is central in the development of an AA-MAS, where test items generally include features that are intended to reduce or eliminate construct-irrelevant influences on student outcomes. However, with the exception of a few studies on testing accommodations (e.g., Elliott & Marquart, 2004; Fulk & Smith, 1995; Koscielek & Ysseldyke, 2000; Lang, Elliott, Bolt, & Kratochwill, 2008; McKeivit & Elliott, 2003), the use of student response data for the purpose of test construction or test validity is virtually absent in the research literature.

In their seminal book Protocol Analysis: Verbal Reports as Data, Ericsson and Simon (1993) describe the rationale for the development of think-aloud methods for obtaining concurrent and retrospective verbal reports. Drawing on recent research on information processing, Ericsson
and Simon developed an approach to collecting concurrent and retrospective verbal reports that demonstrated minimal influence on participants’ problem solving and cognitions. Because of their desire to create “hard data” about individuals’ cognitive processes, Ericsson and Simon’s approach is somewhat restrictive. For example, the experimenter provides limited prompting or encouragement, and often sits behind the participant to discourage interaction. Moreover, “it is important that subjects verbalizing their thoughts while performing a task do not describe or explain what they are doing—they simply verbalize the information they attend to while generating the answer” (p. xiii). Ericsson and Simon (1993) suggested that think-aloud verbalization during assessment tasks reflects the cognitions’ simultaneous happening in participants’ short-term memory. Students may find it difficult to verbalize their problem solving on test items that are too simple and involve skills and concepts at a level of automaticity (i.e., stored in long-term memory). Conversely, items that are too difficult and complex may result in confusion and frustration on the part of student respondents.

Components of a Cognitive Lab: The Arizona Example. In our efforts to develop AA-MAS, we collected data from students (with and without disabilities) taking short tests as part of a think-aloud lab. The work of Johnstone, Bottsford-Miller, and Thompson (2006) influenced the design of our think-aloud method. This method has become a common part of our test development procedures and is useful for evaluating salient characteristics of test items and testing procedures. The think-aloud procedure typically involves two steps:

Step 1. During video- and audio-recorded sessions, data are collected “in real time” by asking students to think-aloud as they complete or solve assessment tasks. Researchers verbally prompt students as infrequently as possible to avoid distracting them during problem-solving activities (Ericsson & Simon, 1993). If students are silent for several seconds, researchers can use neutral prompts, such as encouraging the student to “keep talking” (Johnstone, Bottsford-Miller, et al., 2006).

Step 2. Once the think-aloud process is complete, researchers may ask follow-up questions about the process. Answers to these questions can supplement any unclear data derived from the students’ verbal responses (e.g., which item modifications did students find most useful). These follow-up questions may also be useful for probing the understanding of students who, as a result of skill deficits or disability, are unable to meet the cognitive demands of thinking aloud while problem solving (Johnstone, Bottsford-Miller, et al., 2006).

Cognitive labs generally are conducted with rather small numbers of participants to allow for in-depth analyses of how individuals process and use information. Although no comprehensive review articles have been published on cognitive labs, the studies that have been completed generally feature fewer than a dozen participants and use direct observations, follow-up questions, and post hoc reviews of permanent products (Johnstone, Bottsford-Miller, et al., 2006; Roach, Beddow, Kurz, Kettler, & Elliott, 2009). Cognitive labs with small, representative samples of participants provide some advantages for in-depth examinations of responses; however, small samples of participants also create limitations for data analyses and generalization of the results. Cognitive labs can play an important role in the early phases of the item and test development process when developers are exploring the function and utility of new materials or item formats.
In our collaborative work with the Arizona Department of Education to build an AA-MAS, we used a cognitive lab to address a number of questions about test items and the students potentially eligible to take such an assessment. Specifically, we asked:

1. Do students eligible for an AA-MAS perform better on modified grade-level test items than on original test items? If so, is this improvement greater than that experienced by noneligible students?
2. How do students view the modified versus original grade-level test items? What test/item features, if any, did they prefer or perceive as helpful?
3. What is the relationship between classroom instruction on an objective and student performance on an item designed to measure that objective?
4. Are there significant qualitative differences between the test-taking behaviors of eligible students versus noneligible students?

The main materials used in this Arizona study with students were grade-specific test booklets composed of 12 items (three original and three modified in both reading and mathematics). The original and modified items were clustered in subsets and the order of the original and modified item subsets was counterbalanced across Forms A and B at each grade level.

At least one of the teachers for each participating student was asked to complete a Curricular Experiences Survey (see Figure 3). This survey listed the content standard objective for each item tested and asked teachers to report the amount of instructional coverage provided for each objective. Thus, the purpose of the survey was to document the degree to which students had been provided an opportunity to learn the content assessed by test items.

Testing of students occurred at school sites in each of the school districts. Testing sessions ranged from approximately 35 min for noneligible students to more than 60 min for many of the eligible students. Investigators worked in pairs with each student; one of the investigators prompted the student during the think-aloud process and administered the follow-up survey, whereas the other investigator ensured the integrity of the video- and audio-tapings. Following the directions suggested by Johnstone, Bottsford-Miller, et al. (2006), an investigator explained the think-aloud procedures and modeled the process of verbalizing while thinking. The script used to explain the process is presented as Figure 4.

After explaining instructions and providing a short demonstration of how to verbalize, the investigator asked the student to engage in a couple of sample items to practice verbalizing his or her thoughts. Following this, the testing started with every student completing reading items and then mathematics items. Following Johnstone, Bottsford-Miller, et al.’s (2006) recommendation, investigators prompted students only when they were silent for 10 consecutive sec. If students verbalized infrequently while working on test items, they were reminded to “keep thinking aloud” or “keep talking.” Other than these prompts, the investigators remained silent while students were thinking aloud to avoid disrupting their thought patterns.

After each test (reading and mathematics) was completed, the student was asked to rate the perceived difficulty of each item he or she had completed (see Figure 5). Student responses to these questions provided one form of “cognitive effort” or conversely “cognitive ease” data for calculating efficiency metrics (as developed by Clark et al., 2006). We also recorded the amount of time students spent completing each test item as another measure of “cognitive effort.” Both measures of “cognitive effort” ultimately were combined with data on student performance to
determine item modification effects on efficiency as defined by Clark et al. At the end of each content area subtest, students were asked follow-up questions to (a) clarify any inconsistencies or confusion regarding their think-aloud responses, and (b) gather additional information about their perceptions of the original and modified test items. Student responses to these follow-up questions also were recorded. After completing the cognitive lab session, the investigators administered a curriculum-based, grade-level oral reading fluency prompt to each student.

In our Arizona Cognitive Lab Study, we expected students without disabilities would correctly answer a higher percentage of all items than their classmates with disabilities. In particular, however, we expected both groups of students would (a) perform better on the modified items than the original Arizona’s Instrument to Measure Standards items, (b) indicate the modified items required less effort to answer than their original sibling items, and (c) perceive the modifications to items to be helpful or make items easier to answer. We collected evidence about the items (e.g., depth of knowledge, readability level, difficulty) and students’ performances, including percentage of items correct, perceptions of effort needed to answer an item, and reading fluency. Collectively, the evidence was interpreted to support the first two expectations. With regard to the third expectation, the trends in the evidence indicated that modified items were generally
“We are interested in how students solve items on tests, so we want to ask you and other students to answer some test problems for us and let us listen to how you do that. We are not as interested in the answer you come up with as we are with how you are thinking and what you are thinking about while you’re answering the test questions. If for any reason you would like to stop or take a break at any time, please ask to stop or to take a break.”

The researcher will strive to be general and honest about our interests in student responses and respectful of the contribution each student can make to the research process.

The student should not feel that he/she is being judged or that he/she needs to perform in a certain way or to a specific standard. Any extraordinary pressure may affect his/her behavior and introduces a bias.

The researcher will ask the student to “parrot” back what he or she was told about today’s session by the recruiting person or teacher. If the researcher finds that the student has been given information that is biasing or confusing that might affect the session, he or she will re-explain the process in order to rectify student understanding.

“What were you told we were going to do today?”

The researcher will demonstrate interest about what the student does and why he/she does it. Also the researcher should tell the student that we will be videotaping the session and let him/her know when the camera is turned on.

“What you say is really important, so we are going to run this camera to make sure that we don’t forget anything.”

Provide practice task: Each student will be given (at least one) practice task to familiarize him or her with thinking aloud while working through an item. First, the researcher will answer an item and then ask the student to answer one. (The camera is not turned on for the practice.) The researcher will give the following instruction:

“I’m going to think out loud while I answer this problem. That means I’m going to say everything that goes through my mind.” (Researcher completes a practice item while thinking out loud.)

“Now I’m going to ask you to answer an item the same way. Just say everything that goes through your mind while you figure out the item.” (Student completes a practice item(s) while thinking out loud until they can clearly demonstrate the process before proceeding to the research items.)

“Remember, I am not as interested in the answer to the item as much as how you are thinking about the task. Do you have any questions about what we just did?”

FIGURE 4 Think-aloud protocol script.

How hard did you have to work to answer the reading/math item above?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Very</td>
<td>Hard</td>
<td>Very</td>
<td>Hard</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 5 Sample Cognitive Effort/Cognitive Ease question.
perceived as helpful or similar to original items. In particular, modifications did no harm and may even be helpful from the students’ perspectives.

Making generalizable conclusions from data based on relatively small samples of students and collected at one point in time during a cognitive lab must be done cautiously but can facilitate the next steps in the test development process. In the case of Arizona, the results of the cognitive lab lead to the general endorsement of a collection of item modification tactics (bolding, embedding of visuals and questions, distractor elimination, and language simplification) for reading items and the further refinement of some mathematics items where these tactics could be used more consistently.

**Experimental Research on Item Modification**

As part of the CAAVES project, one large scale \( n = 756 \) study has been completed examining the effects of packages of modifications (often including language simplification, adding graphic support, bolding key words, and eliminating one distractor) on the performance of three groups of eighth-grade students across conditions in reading and mathematics. A rubric with wording taken directly from the guidelines, coupled with the presence or absence of an Individualized Education Program, was used to define the three groups: students with disabilities who would be eligible for an AA-MAS (SWD-Es), students with disabilities who would not be eligible for an AA-MAS (SWD-NEs), and students without disabilities (SWODs). Every student in the study completed items in an original condition, in a modified condition, and in a modified condition with limited reading support. The reading support condition was extensive in mathematics but was limited to the directions and the item stem or question in reading.

Main findings of the CAAVES experimental study included that (a) the reliability of forms could be preserved across modification conditions, (b) modification improved the performance of all three groups of students, and (c) eligible students experienced a differential boost in mean item difficulty when ability was controlled within a Rasch model (Elliott et al., 2009; Kettler, Rodriguez, et al., 2009). Across the two content areas (reading and mathematics), three groups, and three modification conditions, coefficient alphas ranged from .85 to .94 based on 39-item forms. No meaningful differences were found based on group status or condition, indicating that the modification process did not substantially decrease reliability. As would be expected, across conditions, SWODs scored the highest. Also across conditions, SWD-NEs scored higher than did SWD-Es (Elliott et al., 2009). Students with disabilities who would be eligible for AA-MAS exhibited slightly higher effect sizes (Cohen’s \( d \) for reading = .40, mathematics = .26) than did SWD-NEs (reading = .38, mathematics = .21) or SWODs (reading = .37, mathematics = .15). An interaction paradigm, which was not apparent when characterizing performance based on mean scores, was present when characterizing performance in terms of mean item difficulty. When using a Rasch model to equate the ability levels of the three groups, the impact of modifications on item difficulty was significantly larger for SWD-Es than for the other two groups, indicating that the modifications were especially helpful for providing access to students who would be eligible (Kettler, Rodriguez, et al., 2009).

Exploratory analyses conducted as part of the CAAVES experimental study addressed which modifications were included in particularly effective packages, and which were included in packages that were not particularly effective. Nine particularly effective packages were identified
as having resulted in an interaction between group and condition, reducing the gap between eligible and noneligible students. Five particularly ineffective packages were also identified, based on items increasing this gap. These analyses revealed that shortening the stems of items was an effective modification across content areas, and that adding graphics to reading items was a particularly ineffective modification. No pattern was found for the modification of adding graphics to mathematics items. The elimination of a distractor and the increase of white space could not be evaluated in these analyses, because they were a part of all packages, effective and ineffective. In a posttest survey of this sample, Roach et al. (2009) found that students from all groups reported that items were equally difficult across conditions, but that SWD-Es and SWD-NEs were more likely to find items toward the beginning of the test easier, perhaps indicating perceived fatigue toward the end of the test. With regard to specific modifications, students endorsed the use of bold font for key terms, as well as the removal of a distractor.

SUMMARY AND CONCLUSIONS

Precautions to Modifying Items and Tests

The research that has been done so far has raised a number of precautions to modifying items and tests for an AA-MAS. One modification that is often considered is allowing portions of a test to be read aloud to students, either by a proctor or by computer voiceover. In the CAAVES experimental study, the modified condition that featured reading support yielded even larger effect sizes than did the modified condition without reading support, when both were compared with the original condition. These effects sizes were very similar for SWD-Es (Cohen’s $d$ for reading = .50, mathematics = .31), SWD-NEs (reading = .49, mathematics = .25), and SWODs (reading = .38, mathematics = .20). Although a reading support modification is generally considered acceptable in content areas other than reading, states vary greatly in their interpretation of whether and to what degree reading support can be used as a modification on a reading test. The concern with providing read aloud support to passages on reading items is that the modification will go beyond reducing just content irrelevant variance, to also reducing content relevant variance. In other words, some of the meaningful information about students’ abilities to read would be lost. This is not a concern when considering a mathematics or science test, for which reading is an access skill. However, before eliminating the possibility of any reading support on a reading test, one should consider that the sample of students who would be eligible for an AA-MAS are typically very slow readers. Students in our Arizona Cognitive Lab study who were in the noneligible group typically read about twice as many words per minute, compared to students from the eligible group, regardless of grade level. This characteristic would appropriately affect their performance on items designed to measure fluency. However, on a broader reading test that includes items designed to indicate comprehension and vocabulary, fluency is an access skill. The decision about providing reading support in these cases is not as clear, as differences in fluency are likely to affect students’ abilities to show what they know and can do in other areas of reading.

A couple of precautions about the practice of eliminating one distractor from an item must be considered. One precaution that was previously mentioned is that properties of multiple-choice items and tests are best preserved when the least plausible distractor is eliminated (Rodriguez, 2005). Eliminating a distractor that is chosen a high percentage of the time is likely to affect more
than just the reading load of the item, and previous characteristics such as reliability, difficulty, and discrimination will likely be very different for the modified version. When considering the distractor that is least plausible, it is helpful to have item statistics that are calculated based on the population for whom the test is intended (Abedi, Leon, & Kao, 2007). Another precaution related to distractor elimination is that content validity may be changed in the process. In some cases, each of the answer choices may contribute to the content standard to which the item was written, and removal of a distractor might change alignment. One final precaution related to distractor elimination is that three answer choices is not the optimal number in all cases. Some items are written in such a way that the answer choices are balanced, and eliminating one answer choice would cue students toward either a correct or incorrect answer choice. A simple example is a problem that asks the student to multiply 2 by –2, with the following response options: (a) –4, (b) –2, (c) 2, (d) 4. In this case, there is no way to remove a response option without leaving an answer that shares one trait (either the sign or the value) with each of the other two, a cue that is likely to attract a test taker. Such a modification would likely change the properties of an item.

One practice that is allowable but has not been used yet by states is the option to retest students and use the best of multiple scores. Such a practice would not change the construct being reflected by the scores, except to the degree that a student might improve from practice, or that the two forms of a test might not be equivalent. On an individual basis, retesting and using the higher score would result in an equal or higher score being reported, and on a group basis the mean score would be higher. The drawbacks to this practice include a relatively high cost, both fiscally and with regard to time, and that it does not address the goal of developing tests that are more appropriate for the eligible population. Although retesting and using the average would yield a more reliable set of scores, based on calculation from a larger number of items, using the higher score effectively eliminates the contribution of all of the items on the test that is not used. Measurement would only be improved in this case if students systematically received higher scores on a test that was a superior measure.

Lessons Learned and Questions Generated

In a relatively short time doing research inspired by the AA-MAS, we have learned a number of key lessons that have in turn evoked a number of new questions. One of the first lessons that we learned was that the process by which these AA-MAS are to be developed is, as yet, unnamed. Although we prefer modification as a conservative term, others prefer the more positive enhancement, or the more general alteration. Through research, we have learned that the criteria for eligible students in the federal regulations will identify a population of students who achieves at a lower level (Elliott et al., 2009) and reads at a slower rate than do SWODs or SWD-NEs. We have also learned that a theoretical and data-based approach to modification can yield items and tests that are as reliable as the originals (Kettler, Rodriguez, et al., 2009). These modified items and tests are easier for all three groups (Elliott et al., 2009), and the effect is differential in favor of the eligible group, when using a design that controls for differences in ability. Also, we have learned that reducing the number of words in an item stem may be a helpful modification, while caution should be used before adding pictures to reading tests (Kettler, Rodriguez, et al., 2009). Last, through research projects and interactions with colleagues on the topic of designing AA-MAS, we have witnessed the development of a modification paradigm. Although unlikely to solve all of the issues raised by the AA-MAS policy, this evolving paradigm promises to advance
the conversation by providing a common framework and language for testing professionals who are considering these issues.

Questions that remain unanswered include (a) how to define and maintain a grade-level test during modification, (b) whether grade-levelness applies solely to content standards or extends also to item difficulty, (c) whether the modified tests reflect the same theoretical structure that is reflected by the tests in their original conditions, (d) whether reading load should be reduced on a reading test and how to do so, (e) what are the long term consequences of modified examinations and the AA-MAS policy, and of course (f) what will we name the process used to create an AA-MASs? These questions reflect a commitment to evaluating AA-MAS with the same types of evidence used to evaluate any educational measurement tools, and pursuing their answers through a theoretical and data-based modification paradigm, as our best strategy for designing tests that better measure eligible students’ achievement and yield access to grade-level content.

REFERENCES


Hess, K., McDivitt, P., & Fincher, M. (2008, June). Who are the 2% students and how do we design items and assessments that provide greater access for them? Results from a pilot study with Georgia students. Paper presented at the
MODIFYING ACHIEVEMENT TEST ITEMS


Rodriguez, M. C. (2009, April). The role of item response attractors in modified items for students with disabilities. In S. E. Elliott (Chair), Designing and evaluating modified items for students with disabilities: Research results. Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego, CA.


