

Modified Multiple-Choice Items for Alternate Assessments:
Reliability, Difficulty, and the Interaction Paradigm

Ryan J. Kettler
Peabody College of Vanderbilt University

Michael C. Rodriguez
University of Minnesota, Twin Cities

Daniel M. Bolt
University of Wisconsin, Madison

Stephen N. Elliott, Peter A. Beddow, and Alexander Kurz
Peabody College of Vanderbilt University

Please direct all correspondence concerning this manuscript to Dr. Ryan J. Kettler, 405 Wyatt Center, Peabody #59, 230 Appleton Place, Nashville, TN 37203-5721, email: ryan.j.kettler@vanderbilt.edu, Telephone: (615)343-5702, fax: (615)322-4488.

Abstract

The final regulations for the *No Child Left Behind* act (U.S. Department of Education, 2007) indicate that a small group of students with disabilities may be counted as proficient through an alternate assessment based on modified achievement standards (AA-MAS). This new policy inspired the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) project, a four-state collaboration with the goal of investigating item modification strategies for constructing future alternate assessments. An experimental design was used to determine whether tests composed of modified items would have the same level of reliability as tests composed of original items, and also help reduce the performance gap between students who would be eligible for an AA-MAS and students who would not be eligible. Three groups of eighth-grade students ($N = 755$) defined by eligibility and disability status took Original and Modified versions of reading and mathematics tests. In a third condition, the students were provided limited reading support. Changes in reliability across groups and conditions for both the reading and mathematics tests were determined to be minimal. Mean item difficulties within the Rasch model were shown to decrease more for students who would be eligible for the AA-MAS than for non-eligible groups, revealing evidence of an interaction paradigm. Exploratory analyses indicated that shortening the question stem may be a highly effective modification, and that adding graphics to reading items may be a poor modification.

Modified Multiple-Choice Items for Alternate Assessments: Reliability, Difficulty, and the Interaction Paradigm

Recent changes in federal legislation (U.S. Department of Education, 2007) allow states and districts to use a modified version of the regular assessment for up to 2% of the total tested student population to be counted as proficient. These alternate assessments based on modified achievement standards (AA-MAS) are intended for students with disabilities whose Individual Education Plans (IEPs) refer to grade-level content goals, whose inability to reach proficiency is determined to be the result of their disability, and who are highly unlikely to attain proficiency on the regular assessment. Acceptable modifications to the standard test should be made with the intent of providing universal access and reducing cognitive load to ensure that scores on the AA-MAS permit valid inferences about the same constructs measured by the regular assessment. Empirical evidence of the success of modifications may be characterized by an interaction paradigm, historically used to evaluate testing accommodations (Phillips, 1994). Using this paradigm, eligible students should benefit more from modifications to the test than students who are not eligible. The change in federal policy, the subsequent push to identify students who are eligible, and the need for data on creating valid modifications to regular assessments all inspired the current experimental study of the effects of modifications on test score reliability, item difficulty, and the interaction paradigm for students who are eligible for an AA-MAS.¹

The Legal Policy

The final regulations for the *No Child Left Behind* act (*NCLB*; U.S. Department of Education, 2007) indicate that a group of students identified with disabilities, not to exceed 2% of the total tested population, may be counted as proficient through an AA-MAS. These students can take a version of the regular assessment test with modifications. By definition, this new version of the regular test would result in a different test and would also require new cut scores for determining proficiency levels for *NCLB* reporting. *Modifications* are changes to a test's content or item format that may make a test more appropriate for some students, but may also make the test easier. Much like testing accommodations, modifications are intended to facilitate access to the assessment for eligible students to enable meaningful comparisons of their scores with the scores of students who take the regular test. Unlike testing accommodations, modifications are changes to aspects of items that result in a less difficult test while maintaining the same grade level as the original.

The final regulations allow the use of an AA-MAS for a subset of students identified with disabilities for whom such an assessment may be more appropriate. The modified standards are intended to meaningfully include in the assessment process those students for whom: (a) the standards of the regular assessment are too difficult and (b) the standards of the alternate assessment for students with significant cognitive disabilities are too easy. These modifications are intended for students whose disabilities have prevented them from reaching proficiency and whose disabilities make it unlikely that they will reach proficiency by the same standards and within the same timeframe as students who are not eligible (U.S. Department of Education, 2007). They are meant for students with disabilities who have a consistent record of below grade-level performance on achievement tests and, while they are able to make progress, have not reached grade-level achievement standards in spite of quality instruction.

Students who meet these criteria may take the AA-MAS, but no more than 2% of tested students within a district or state may be counted toward proficiency reports for AYP calculations. An additional clause allows that additional students may be counted as proficient on an AA-MAS, but only if less than 1% are counted as proficient on the alternate assessment for students with significant cognitive disabilities. The combined proportion of students counted as proficient on the two alternate assessments may not exceed 3%. Individual Education Plan teams must decide on a student-by-student basis whether an AA-MAS is the appropriate assessment. While there is no limit on the number of students who can take the test, the district and state will only benefit by identifying the 2% to 3% of students whose scores will improve enough to meet proficiency. Accurately identifying those students will require a clear understanding of the modifications that are made to items from a regular assessment to create an AA-MAS.

Modification Strategies

Modifications used to develop tests for eligible students should increase the students' access to the tests, which are intended to be aligned with the content to which they have been exposed in the general curriculum. With the aim of increasing item accessibility, the current study employed principles of universal design, cognitive load theory (CLT), and research on item development. The Center for Universal Design (CUD) has recommended the consideration of seven principles for the universal design of environments, products, and services "to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design (CUD, 1997)." Table 1 includes a list of these principles, along with definitions and guidelines that correspond to each, quoted directly from the Center for Universal Design Website. Although not intended specifically for testing, these principles should be considered when developing any large-scale assessment. They serve as an appropriate starting point for developing items that are maximally accessible for students with disabilities who consistently fail to achieve proficiency. The principle of simplicity and intuitiveness, for example, means that the environment, product, or service should be easy to understand. One of the principle's guidelines suggests the elimination of unnecessary complexity, a goal which may be met in several ways: by changing a compound sentence into a simple one, converting unfamiliar notation in a mathematics problem to a more common form, or removing unnecessary graphics, among many other possibilities. These modifications are likely to be helpful for many students, and may be especially helpful for students who would be eligible for an AA-MAS.

Modifications to regular assessments that are designed to help students who consistently fail to meet proficiency can also be guided by CLT. Conceptualized by Sweller (Clark, Nguyen, & Sweller, 2006) and based on Miller's (1956) classic 7 ± 2 paper on the limitations of working memory, CLT has been applied to classroom instruction to improve efficiency and student learning. It also has application to test construction and student responses to items. The theory posits that there are three types of short term memory loads for learning tasks: intrinsic load, germane load, and extraneous load. Intrinsic load is characterized by the complexity of a task and is heavily influenced by the associated goals. Germane load is characterized by the additional work that is relevant to the associated goal, and although not necessary for meeting the goal, it is theorized that tasks requiring germane load improve outcomes by increasing the generalizability of the skills being learned. Extraneous load is memory load unrelated to the task.

Clark, Nguyen, and Sweller assert that tasks requiring extraneous load result in a waste of mental resources. They argue that learning is made more efficient by decreasing the extraneous load of learning tasks without affecting intrinsic or germane load.

Table 1

Center of Universal Design's Principles, Definitions, and Example Guidelines

1. Equitable Use: useful and marketable to people with diverse abilities (e.g., Provide the same means of use for all users: identical whenever possible; equivalent when not.)

2. Flexibility in Use: accommodates a wide range of individual preferences and abilities (e.g., Facilitate the user's accuracy and precision.)

3. Simple and Intuitive Use: easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level (e.g., Eliminate unnecessary complexity.)

4. Perceptible Information: communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities (e.g., Maximize "legibility" of essential information.)

5. Tolerance for Error: minimizes hazards and the adverse consequences of accidental or unintended actions (e.g., Discourage unconscious action in tasks that require vigilance.)

6. Low Physical Effort: can be used efficiently and comfortably and with a minimum of fatigue (e.g., Minimize repetitive actions.)

7. Size and Space for Approach and Use: appropriate size and space is provided for approach, reach, manipulation, and use regardless of user's body size, posture, or mobility (e.g., Accommodate variations in hand and grip size.)

Although it has historically been used to improve the efficiency of instruction, CLT can be applied to the efficiency of test items as well. Consider the item in Figure 1.

Figure 1.

Sample multiple-choice test item

1. Jane would like to purchase a candy bar that costs \$1.25. If she were to pay with two one-dollar bills, how much change would Jane receive in return?
 - a. \$.75
 - b. \$.25
 - c. \$.50
 - d. \$1.00

If the stated goal of the item is to assess triple digit subtraction, then the intrinsic load can be represented by $2.00 - 1.25 = .75$. Extraneous load may include use of the two syllable word *purchase* rather than *buy* or writing in the conditional tense. Other information in the problem used to make the context richer than $2.00 - 1.25 = .75$ is germane load, helpful for assessing abilities across multiple contexts, but not directly pertinent to the stated goal. As illustrated, cognitive load theory can be applied to tests by removing or reducing the extraneous load from an item such as this, and can even help remove much of the germane load while preserving the intrinsic load and the grade level of the item. Key CLT guidelines identified by Sweller et al. include (a) using cues to focus attention on content; (b) reducing content to essentials; and (c) eliminating extraneous visuals, text, and audio.

Some guidance regarding the modification process for AA-MAS has also been taken from the research literature on item development. For example, one popular item-modification strategy with empirical support involves the reduction of the number of response options to a multiple-choice (MC) question. By reducing the number of response options, it is assumed that both the complexity of the decisions to be made and the amount of reading are reduced as well, if not the difficulty of the item. Rodriguez's (2005) meta-analysis of 27 studies addressed the question, "What is the optimal number of response options for a multiple-choice test?" Using the psychometric criteria of item difficulty, item discrimination, and test score reliability, Rodriguez concluded that:

"Three options are optimal for MC items in most settings. Moving from 5-option items to 4-option items reduces item difficulty by .02, reduces item discrimination by .04, and reduces reliability by .035 on average. Moving from 5- to 3-option items reduces item difficulty by .07, does not affect item discrimination, and does not affect reliability on average... Moving from 4- to 3-option items reduces item difficulty by .04, increases item discrimination by .03, and increases reliability slightly by .02." (p. 10)

While many states are using this strategy for item modification for students with disabilities, Rodriguez's (2005) findings indicate that reducing the number of distractors does not harm the psychometric properties of the test within the general population, but does theoretically reduce the cumulative cognitive load of the test. Modifications such as this are truly in the spirit of making items more accessible to students with disabilities.

Following the principles of universal design and CLT, as well as research on item development, a number of specific strategies have become common practice for developing an AA-MAS. Based on a survey of six states, the National Center on Educational Outcomes (Lazarus, Thurlow, Christensen, & Cormier, 2007) released a report describing the most common modifications used. Removing a distractor from a MC item, reducing the number of items on the test, and simplifying language were the most common modifications, followed by reducing the number and length of reading passages. A separate set of modifications that might be considered testing accommodations was also included, with reducing the number of items per page and increasing the font size being the most commonly considered alterations. As we will discuss in the next section, modifications and accommodations can not only be used to make assessment scores more accessible for students with special needs, but they can increase their validity as

well. A number of the modifications and accommodations listed in Lazarus al.'s (2007) report were used for the current study.

Testing Accommodations versus Modifications

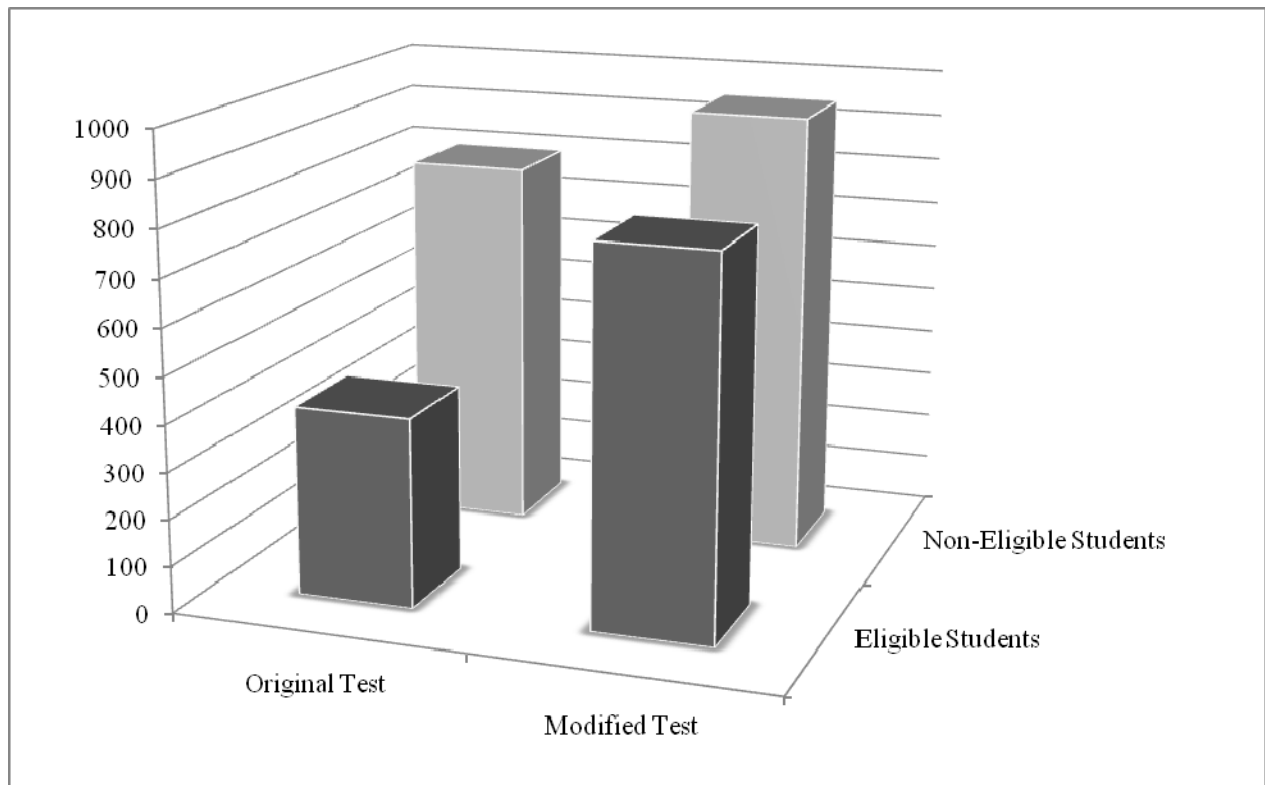
Testing accommodations and modifications both are alterations made to a test or testing administration that are intended to make the test more accessible. According to Hollenbeck (2002), appropriate accommodations (a) result in unchanged constructs being measured, (b) are based on individual needs, (c) lead to differential effects between students who need them and students who do not, and (d) can be used to make the same inferences as made from scores in original testing conditions. To be considered an accommodation rather than a modification, a test alteration should meet all four of these criteria. Hollenbeck cited the following as examples of accommodations that are usually appropriate: (a) manipulating the physical environment, (b) having a familiar person administer the test, (c) allowing the test-taker to mark in the test booklet, and (d) changing the pacing of the administration.

Phillips and Camara (2006) concurred with Hollenbeck's interpretation, indicating that the term *accommodation* should be used to refer to nonstandard test administrations that produce comparable scores, and that the term *modification* should be used to refer to nonstandard test administrations for which evidence of comparability is lacking. Koretz and Hamilton (2006) indicate that unlike accommodations, modifications may involve changes in test content. Although the test alterations in the current study were modifications in the sense that they were applied to groups and test content was changed, they were intended to be more similar to accommodations on three of the aforementioned attributes: unchanged constructs, sameness of inference, and differential effects. Our use of the term *modification* is intended to be conservative, reflecting a lack of evidence that the construct being measured has been preserved, rather than to imply that the test alterations have changed the constructs being measured or the inferences that can be made.

One empirical criterion for determining whether an alteration is an accommodation or a modification is whether an interaction paradigm (also known as differential effects or a differential boost) exists between the extent to which scores are increased for students with disabilities (SWDs) versus students without disabilities (SWODs). An interaction paradigm is the concept that if an accommodation is working, it should increase the scores of SWDs, but should generally not affect the scores of SWODs (Phillips, 1994). While accommodations are designed to make a test more accessible but not easier, modifications are allowed to make a test more accessible and easier, as long as the measured content is not considered to be off grade level. Therefore, an AA-MAS need not exhibit the same level of differential effects exhibited by valid testing accommodations, because modifications are allowed to result in a test that is marginally easier even for students that would not be eligible. However, it is reasonable to expect that valid modifications will help reduce the discrepancy (or at least not increase this gap) between the scores of students who would be eligible and those who would not be eligible if they took the same test. Based on these assumptions, Figure 2 depicts the ideal interaction paradigm based on modifications made to a regular assessment. Measurement of this interaction paradigm was a primary design influence for this project.

Figure 2.

Interaction paradigm on scores of a hypothetical test, from Original to Modified conditions, for students grouped by eligibility for an AA-MAS



The Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) Project

The CAAVES project focused on modified MC items and involved a 4-state collaboration (Arizona, Hawaii, Idaho, and Indiana) with the primary goal of investigating the feasibility of item modification strategies for future alternate assessments. This goal was accomplished by (a) developing a common set of items from existing Discovery Education Assessment (DEA) reading and mathematics tests using modification principles with the aim of facilitating reading access and valid responses, and (b) using a computer-based test delivery system to experimentally examine student preferences, score comparability, and statistics of the modified items for students with and without disabilities. Project leaders modified a common set of existing reading and mathematics items to create tests that are more accessible but still reflect the same grade-level content as the original items. A subset of project leaders then conducted a cognitive lab study with a small sample of students with and without disabilities, to gain their student insights into which item modifications are preferred and most likely to improve test access. During the second round of refinement, it was decided that a third condition should be added, in which students would receive limited reading support with the modified items. Although controversial within some state assessments, the modification was consistent with reducing extraneous cognitive load; therefore, the authors of the current study determined that a close study of its effects was warranted. The CAAVES team implemented a 4-state experimental study to compare the effects of tests with and without modified items on students' test

performances and test score comparability. The following research questions, testable as hypotheses, were addressed by the study:

1. *How do modifications affect the reliability of test scores?* We predicted no difference in reliability between the Original condition, the Modified condition, and the Modified with Reading Support condition, based on previous research regarding the elimination of one distractor, as well as the relatively conservative nature of our other modifications.
2. *How do modifications affect item difficulty across groups?* We predicted that modifications would decrease item difficulty across groups, and that there would be a group-by-condition interaction, such that the students with disabilities who are eligible (SWD-Es) would benefit more than will students without disabilities (SWODs) or students with disabilities who were not eligible (SWD-NEs). This prediction is based on the strong theoretical and research basis of the modification process.

Method

Participants

The sample included 755 eighth-grade students from four states, balanced across three groups: SWODs ($n = 269$), SWD-NEs ($n = 236$), and SWD-Es ($n = 250$). Eighth grade was selected because (a) students would be articulate enough to provide feedback on testing experiences, (b) in most states it is the final grade for testing prior to the high stakes high school assessment, and (c) participants were likely to be familiar with computers. Sample sizes by state were as follows: Indiana ($n = 463$); Idaho ($n = 164$); Arizona ($n = 74$); and Hawaii ($n = 54$). The sample included a higher number of male students ($n = 440$) than female students ($n = 315$). Participants were primarily European American (69%), Latino American (12%) and African American (11%) students. The study groups were well-balanced with regard to these demographic variables, with the exception that, concurrent with national parameters, males represented higher proportions of the SWD-NEs (65%) and the SWD-Es (64%) groups. Table 2 includes a summary of these demographics by group membership. Of the 755 students, 721 participated in the Reading test and 717 participated in the Mathematics test.

Among the students identified with disabilities, the largest percentage were identified with a Specific Learning Disability (60%), followed by Mental Retardation (14%). This pattern was largely the same between students in the SWD-NE and SWD-E groups, except that the former contained a larger percentage of students in the Specific Learning Disabilities category (SWD-NE = 65%, SWD-E = 46%), and the latter contained a larger percentage of students identified with Mental Retardation (SWD-NE = 3%, SWD-E = 23%). Table 3 includes a summary of disability category by group membership for the two groups of students who had disabilities.

Measures

CAAVES AA-MAS Participation Decision Criteria. To facilitate identification of students who would be eligible, project leaders developed the *CAAVES AA-MAS Participation Decision Criteria* document. The criteria provide examples of evidence that could be used to evaluate each

of the three criteria indicated in the federal policy. The first criterion was that a student have a current IEP with goals based on academic content standards for the grade in which the student was enrolled. For the CAAVES project, researchers recommended examining current IEP plans for goals and statements aligned to grade level content standards in reading, language, mathematics, or science; for IEP statements that show that the instructional material or curriculum contained grade level content; and for IEP team member statements that the IEP goals and instruction provided to the student aligned with grade level content standards. The second criterion was that the student's disability precluded her or him from achieving grade-level proficiency, as demonstrated by performance on the state assessment or another assessment that can validly document academic achievement. Evidence that a student met this criterion included previous years' general education test results with performance documented at the lowest proficiency level, or results from a recent achievement test known to accurately predict summative test performance equivalent to the lowest level on the statewide general education test. The third criterion indicated by the regulations was the student's progress to date in response to appropriate instruction designed to address individual needs and based on multiple measurements was such that, even if significant growth occurred, the IEP team was reasonably certain that the student would not achieve grade-level proficiency within the year covered by the student's IEP. Evidence relevant to this criterion included a written description of research-based instruction programs used to improve achievement, along with (a) two years of class performance records and materials (i.e., report cards and IEP progress reports toward the grade level annual goals and objectives), (b) three years of the student's scores on state achievement tests, or (c) multiple curriculum-based measurement scores and growth rates compared to grade-level national or local norms and proficiency levels. A copy of the *CAAVES AA-MAS Participation Criteria Decision Guide* is available at <http://peabody.vanderbilt.edu/x8312.xml>.

Table 2
Demographics by Group Membership

	SWODs <i>n</i> (%)	SWD-NEs <i>n</i> (%)	SWD-Es <i>n</i> (%)	Total <i>N</i>
State				
Arizona	28 (10%)	25 (11%)	21 (9%)	74 (10%)
Hawaii	20 (7%)	16 (7%)	18 (7%)	54 (7%)
Idaho	62 (23%)	43 (18%)	59 (24%)	164 (22%)
Indiana	159 (59%)	152 (64%)	152 (61%)	463 (61%)
Gender				
Male	127 (47%)	153 (65%)	160 (64%)	440 (58%)
Female	142 (53%)	83 (35%)	90 (36%)	315 (42%)
Ethnicity				
European American	183 (68%)	175 (74%)	165 (66%)	523 (69%)
African American	31 (12%)	18 (8%)	34 (4%)	83 (11%)
Asian American	6 (2%)	3 (1%)	3 (1%)	12 (2%)
Hawaiian/Pacific Islander	3 (1%)	3 (1%)	2 (1%)	8 (1%)
Native American	2 (1%)	2 (1%)	7 (3%)	11 (2%)
Latino American	33 (12%)	26 (11%)	31 (12%)	90 (12%)
Multiracial/Other	11 (4%)	9 (4%)	8 (3%)	28 (4%)
Total	269	236	250	755

Table 3

Disability Category by Group Membership for Students with Disabilities

	SWD-NEs <i>n</i> (%)	SWD-Es <i>n</i> (%)	Total <i>n</i> (%)
Autism	4 (2%)	8 (3%)	12 (3%)
Mental Retardation	7 (3%)	58 (23%)	65 (14%)
Specific Learning Disability	154 (65%)	116 (46%)	270 (60%)
Emotional Disturbance	20 (9%)	14 (6%)	34 (8%)
Traumatic Brain Injury	1 (0%)	2 (1%)	3 (1%)
Speech/Language Impairment	8 (3%)	10 (4%)	18 (4%)
Visual Impairment	2 (1%)	0 (0%)	2 (0%)
Deafness/ Hearing Impairment	2 (1%)	1 (0%)	3 (1%)
Orthopedic Impairment	4 (2%)	4 (7%)	8 (2%)
Other Health Impairment	18 (8%)	18 (8%)	36 (8%)
Multiple Disabilities	0 (0%)	1 (0%)	1 (0%)
Total	220	232	452

Note. Specific disability information was not available for the 54 students from Hawaii.

*Test Accessibility and Modification Inventory.*² The *Test Accessibility and Modification Inventory (TAMI)*; Beddow, Kettler, & Elliott, 2008) was designed to facilitate a comprehensive analysis of test items with the purpose of enhancing access and meaningful responses from all students. The *TAMI* was expressly influenced by principles of universal design (e.g., the Center for Universal Design, 1997), test accessibility (e.g., Johnstone, Thurlow, Moore, & Altman, 2006), cognitive load theory (e.g., Clark et al., 2006), and fairness (Educational Testing Service, 2006). With the aim of generating a complete list of considerations for designing accessible tests, Beddow, Kettler, and Elliott also consulted research on testing accommodations, item writing, and item modification (e.g., Clark et al., 2006; Rodriguez, 2005) when designing the inventory.

The *TAMI* includes six Item Accessibility Categories based on the key elements of a test item: Passage/Stimulus, Item Stem, Visuals, Answer Choices, Page and Layout, and Fairness. Each category contains approximately 10 considerations for designing items with a focus on accessibility. For each item under evaluation, raters use the *TAMI* with an accompanying worksheet to evaluate each element of the item on a 4-point scale (0 = not accessible; 1 = minimally accessible; 2 = moderately accessible; 3 = maximally accessible) and record suggested areas to consider for modification. Finally, the rater sums the Accessibility Category ratings to yield a Total Item Accessibility Rating. To facilitate documentation of modifications, *TAMI* worksheets also include sections for recording integrated summaries of suggested modifications as well as final changes to items. The version of the *TAMI* used in this study included only individual item considerations, but a revision is in development that includes considerations for designing accessible computer-based tests.

Reading and mathematics tests. In the Original condition, each test in reading and mathematics was composed of 39 computer-based multiple-choice items. The items were provided by DEA, from a pool of items that are used to meet assessment needs for clients nationwide. Tests in each content area were composed of items from two subscales. The reading test contained 19 vocabulary items that required students to define words and 20 comprehension items that required students to read passages and respond to related items. The mathematics test contained 20 numbers items that required students to decode mathematical symbols and perform basic operations and 19 data items which required students to perform basic arithmetic operations. Each test was further divided into three 13-item sets, each containing a balance of items from the two subscales.

The same test sets that were used for the Original condition were modified by a panel of educators and test design experts from the CAAVES team and the six states that were originally involved in the study. The group was trained in methods of item modification, and was provided the *TAMI* to improve and standardize the quality of modifications to each item. The group was then divided into teams assigned to modify items from one of each of the four subscales. Teams were provided extensive descriptive statistics for each item, based on previous use through DEA. These statistics included item difficulty and discrimination, frequency of response choice selection, depth of knowledge, readability based on eight different indices, frequency of omission, and reliability impact on scale. Teams then convened within the content areas for further item modification. Ultimately, each item was shared with the entire group for a third round of modifications. Items were then pilot tested with a small group of students within a cognitive lab, after which the entire group evaluated the items and generated final modifications. For more information on the cognitive lab pilot study, the reader is directed to Roach, et al. (2008). The most common modifications used throughout both tests included removal of a response option, simplification of language (in the item passage, stem, or response options), addition of graphic support, and reorganization of layout (e.g., breaking one paragraph into several, bolding of key words, adding white space between response options).

For the Modified with Reading Support condition, the same items from the Modified condition were used, except that each item appeared on a single screen. To reduce the cumulative reading load of the test, students were given limited reading support through a recorded voice which read item directions and stems automatically. Item options and graphics that contained words could also be played aloud by clicking on an audio file icon. On some parts of reading items, when reading support would have invalidated the construct being measured, reading support was not made available. For example, the recorded voice did not read key vocabulary words.

Coefficient alpha (reading = .89, mathematics = .85) across the sample and test-retest (reading = .78, mathematics = .65) with a subsample of students from Indiana ($n = 42$) indicated that the reading and mathematics tests had acceptable reliability for individual decision-making. Because the tests were administered in 13-item sets in randomly assigned conditions and orders, we also estimated the internal consistency of these smaller sets of items. The precision-weighted average coefficient alpha for reading was .68 (13-item forms), with an approximate standard deviation of .07 across sets. The alpha adjusted for score variance-heterogeneity was .77, with a standard deviation of .04. The alpha based on the Spearman-Brown (SB) adjustment to a 39-item test was .91. The precision-weighted average coefficient alpha for mathematics (13-item forms) was .58,

with an approximate standard deviation of .07. This alpha adjusted for score variance-heterogeneity was .70, with a standard deviation of .02. This alpha based on the SB adjustment to a 39-item test was .88. These results are based on the procedures described in the following section.

Procedure

Research personnel from state departments of education, who were also partners in the CAAVES project, selected a minimum of two school districts that collectively had at least 200 students with identified disabilities who participated in the previous year's general education achievement test. State leaders then identified the teachers of these students, shared information about the study and participation criteria, and asked them to apply it to the rosters of their morning classes. Participating teachers were asked to review the criteria for all of their students with IEPs and to indicate students who appeared to meet all three criteria for participation in an AA-MAS (SWD-Es) and those who clearly did not meet all three criteria (SWD-NEs). Teachers then submitted their rosters to special education directors, with students divided into these two groups. Special education directors then checked the lists of students with disabilities who were thought to meet criteria for a modified assessment and placed disagreements in the SWD-NE group. The revised list was then forwarded to research personnel from state departments of education. State leaders selected two male and two female students without an identified disability from each class roster for possible inclusion in the SWOD group. Finally, parent and student written consents were collected for all participants in the study.

The final list of students was submitted to DEA for random assignment to a set of testing packages. All three groups of students across four states completed computer-based reading and mathematics tests in February and March of 2008. Students in each of the three groups completed a 39-item reading test and a 39-item mathematics test during two separate sessions, either on a single day or on two consecutive days. No testing accommodations were provided in this study. This requirement did not result in omitting any students. All student answers were recorded online by DEA, which then provided scores to the CAAVES team for analysis.

Design

Study design. The current study used an experimental design, with students in three pre-determined groups completing all items on tests covering the same reading and math content. Group (SWODs, SWD-NEs, and SWD-Es) was the only between subjects variable. In order to control for any potential order effects, the items in reading and mathematics were broken into three sets (A, B, and C), and the order of the three sets across three parts of test administration (Part I being the first 13 items a student completed, Part II being the next 13 items, and Part III being the last 13 items) was randomized. Each student completed all three sets and worked in all three conditions (Original, Modified, and Modified with Reading Support) across the three parts of the test.

Set design. The order of sets and conditions was randomized to remove systematic error that could be attributed to potential order effects related to set or condition. The design resulted in 36 unique test forms, as defined by order of conditions and order of sets. Tables 4 and 5 depict the

frequencies of students taking each set by group and condition in reading and mathematics, respectively.

Table 4

Frequencies of Students Completing Reading Tests by Group, Condition, and Form

	Original			Modified			Modified with Reading Support		
	<u>Form</u>			<u>Form</u>			<u>Form</u>		
	A	B	C	A	B	C	A	B	C
SWODs	86	83	87	82	86	88	88	87	81
SWD-NEs	78	76	74	73	81	74	77	71	80
SWD-Es	86	71	80	81	84	72	70	82	85

Table 5

Frequencies of Students Completing Mathematics Tests by Group, Condition, and Form

	Original			Modified			Modified with Reading Support		
	<u>Form</u>			<u>Form</u>			<u>Form</u>		
	A	B	C	A	B	C	A	B	C
SWODs	82	88	86	85	87	84	89	81	86
SWD-NEs	77	76	70	75	71	77	71	76	76
SWD-Es	77	81	80	82	76	80	79	81	78

Data Analyses

Reliability analyses. Coefficient alpha was estimated for each of the 81 group x condition x set x part combinations for the reading and mathematics tests, resulting in 81 coefficients for analysis in each subject. The analyses were completed through a meta-analytic approach, treating each combination as a “study” of the measurement instrument, permitting the evaluation of the effects of group, condition, and set on the magnitude of score reliability. To facilitate this analysis, coefficient alpha was transformed using a statistical theory for the distribution of a set of k independent, asymptotically normally distributed estimates of coefficient alpha each with a large sample variance (see Rodriguez & Maeda, 2006). This transformation and variance is based on the sampling distribution theory of coefficient alpha and provides a way to conduct analysis on coefficient alpha that maximizes precision. These analyses are only approximate because subsets of alphas are dependent (each group of individuals took three 13-item sets); however, this dependence is fully balanced as the orders of sets and conditions were randomly assigned to groups.

The converted alphas were synthesized using a meta-analytic model, weighting each by its precision (a function of its variance), and used in a weighted-least-squares analysis to assess the impact of group, condition, and set on the magnitude of reliability. In addition, to recognize the relationship between alpha and sample variance, alpha was adjusted to a constant variance based

on the total observed-score variance within subject. To obtain the adjusted reliability given a referent score variance (adjusting for a constant score variance), $r_{\alpha}^* = 1 - \frac{S_X^2(1 - r_{\alpha})}{S_X^{2*}}$.

This approximates each alpha as though based on group scores with equal variance, S_X^{2*} , 9.33 for reading and 8.56 for mathematics. Average alphas were then adjusted following analyses to approximate 39-item tests based on the Spearman Brown Prophecy Formula (SB): $r_{\alpha}^{SB} = \frac{3r_{\alpha}}{1 + 2r_{\alpha}}$

Item difficulty and interaction paradigm analyses. Because of its capacity to represent item difficulty in a way that it not influenced by the ability level of the examinee, we used item response theory (IRT) to study item difficulty changes in relation to the modifications, as well as differential effects across the studied groups. To this end, we initially applied the Rasch model to all 117 item x condition combinations in each content area and for each group. The Rasch model characterizes each item by a difficulty parameter (usually referred to as a b parameter), such that items with more positive b parameters are harder items, and items with more negative b parameters are easier. Typically b parameters range from -3 to +3. By fitting the Rasch model to each of the groups separately, each of the 117 items had a unique b parameter for each group, and we could effectively study how the modifications applied to the items changed their difficulty within each group. To compare b parameters across groups, the calibrations for each of the three respondent groups were first linked. The 39 items in the Original condition in each content area were used as linking items to place the three calibrations across groups on a common metric. Specifically, we applied a mean and sigma transformation such that the difficulty parameters of the linking items were assumed to have a common mean and standard deviation for each group. Once the item difficulty parameters were on a common metric, we were able to examine how the modifications applied to the items resulted in differential difficulty changes (i.e., interaction paradigm) under each set of modifications for each group. By treating items as the units of analysis, we examined whether the mean change in difficulty across items tended to be greater for one group of examinees compared to another. We used paired t-tests and nonparametric Wilcoxon tests for all significance testing for these analyses.

Results

Collectively our analyses indicate that items can be successfully modified to improve access to tests for eligible students. Modified item sets remained comparable with regard to reliability, while the difference in difficulty for eligible students and students who were not eligible was reduced. One particularly effective modification in this regard was shortening the length of the question stem. With this overview of results as background, a detailed examination of the evidence for each research question is provided next.

Effect of Modifications on Reliability

We analyzed score reliabilities based on alphas resulting from the 13-item sets. Results represented the effect of design variables on the SB-adjusted alphas to illustrate the impact of the study variables on 39-item tests for reading and mathematics.

Reading test. The results of the weighted least squares ANOVA for reading indicated a significant condition x group interaction, $F(4, 48) = 5.18, p < .05$; and a significant condition x part interaction, $F(4, 48) = 6.23, p < .05$. The differences among study groups were nearly the same in each condition, except for the SWOD group, for whom the Modified condition yielded a slightly higher alpha (.932 and .936) than the Original condition (.917). Students in the SWD-E group tended to have scores with slightly lower alphas (lower by .03 to .05). The differences in alpha among study conditions also varied by part. There were no significant differences in part reliabilities for original tests. Differences increased in both conditions that featured modifications, with part III having the lowest reliability for the Modified condition (by .016) and parts II and III both being lower for the Modified with Reading Support condition (by .015 for part II to by .011 for part III). These differences are very small, and the full rotation of item-sets in all three parts of each set minimized the impact of part on the remaining results. Finally, sets displayed significant but small differences, amounting to differences in alpha less than .005. The four design variables (study group, condition, form, and part) accounted for 91% of the variation in alpha.

Mathematics test. The results of the weighted least squares ANOVA for mathematics indicated a significant main effect for part, $F(2, 48) = 5.32, p < .05$. Part III tended to have the lowest reliabilities, by .007 on average; a very small effect. The mathematics score results also indicated a significant condition x group interaction, $F(4, 48) = 4.19, p < .05$; and a significant set x group interaction, $F(2, 48) = 4.31, p < .05$. Mathematics sets displayed greater variation in alpha than did the reading sets. The differences among the groups were greater in set B (as large as .041) than in set A (differences as large as .021) or set C (differences as large as .025). On each form, students in the SWOD group had scores with slightly higher alphas than those in the other two groups. Students in the SWD-E group tended to have scores with slightly lower alphas compared to those in the SWOD or SWD-NE group, except in the Original condition. Unlike the reading scores, the differences among study groups were smaller under each condition (differences as large as .023 in the Original condition and .037 in the Modified condition). In each case, these differences are small. The four design variables accounted for 84% of the variation in alpha. Table 6 depicts mean coefficient alphas in reading and mathematics by group and condition, after the SB adjustment.

Effect of Modifications on Item Difficulty.

Table 7 shows means of the Rasch difficulty estimates by group and condition across the reading items. One item was removed from the reading analysis because all examinees in the SWOD group answered it correctly under the Modified condition, precluding estimation of a difficulty parameter for the item. Apparent from the first column in the table is the fact that the mean and standard deviation of the 38 reading item estimates are the same across groups, which is a result of the linking procedure applied to equate the IRT metric across groups. Consequently, the item difficulty parameters in the Modified and Modified with Reading Support conditions can be compared to the items in the Original conditions, as well as across groups, to evaluate whether differential changes in difficulty occurred when applying the item modifications.

Table 6
Mean Reading and Mathematics Coefficient Alphas by Group and Condition (95% Confidence Intervals)

Group	Reading			Mathematics		
	Mean	Lower	Upper	Mean	Lower	Upper
Original condition						
SWODs	.92	.91	.92	.89	.88	.89
SWD-NEs	.90	.90	.91	.86	.86	.87
SWD-Es	.89	.88	.89	.87	.86	.88
Modified condition						
SWODs	.93	.93	.94	.89	.89	.90
SWD-NEs	.91	.90	.91	.87	.86	.88
SWD-Es	.88	.88	.89	.85	.85	.86
Modified with Reading Support condition						
SWODs	.94	.93	.94	.90	.89	.90
SWD-NEs	.91	.91	.92	.87	.86	.88
SWD-Es	.88	.88	.89	.86	.86	.87

Table 7
Mean Rasch Difficulty Estimates for Reading Items by Group and Condition

	Original		Modified		Modified with Reading Support	
SWODs	0.51	(1.75)	-0.03	(1.60)	-0.16	(1.64)
SWD-NEs	0.51	(1.75)	-0.07	(1.38)	-0.17	(2.36)
SWD-Es	0.51	(1.75)	-0.40	(1.76)	-0.24	(1.75)

As expected, the mean item difficulties in all groups declined on average when the items were administered in their Modified or Modified with Reading Support conditions. More critical to the current analysis is that there was a statistically greater decrease in item difficulty for the SWD-E group compared to the SWOD group when items were changed from the Original to Modified condition, $t(37) = 2.38, p < .05$. There was also a statistically greater decrease in item difficulty for the SWD-E group compared to the SWD-NE group when items were changed from the Original to Modified condition, $t(37) = 3.12, p < .05$. Both of these significant results were also obtained under a nonparametric Wilcoxon test ($z = 2.183, p < .05$; $z = 2.937, p < .05$, respectively).

Table 8 shows the corresponding results for all 39 Mathematics items under each of the modification conditions. As with the reading items, the modifications on average reduced the difficulty of the items. Once again, there was a statistically greater decrease in item difficulty for students in the SWD-E group compared to students in the SWOD group when items were

changed from the Original to Modified condition, $t(38) = 2.11, p < .05$, as well as from the Original to Modified with Reading Support condition, $t(38) = 4.12, p < .05$. There was also a greater decrease in item difficulty for the SWD-E group compared to SWD-NE group when items were changed from the Original to Modified with Reading Support condition, $t(38) = 3.33, p < .05$. These latter two tests were also confirmed by the Wilcoxon test ($z = 3.48, p < .05$; $z = 2.99, p < .05$, respectively).

Table 8

Mean Rasch Difficulty Estimates for Mathematics Items by Group and Condition

	Original	Modified	Modified with Reading Support
SWODs	0.16 (.52)	-0.12 (.86)	-0.04 (.63)
SWD-NEs	0.16 (.52)	-0.15 (.70)	-0.08 (.59)
SWD-Es	0.16 (.52)	-0.34 (.41)	-0.43 (.60)

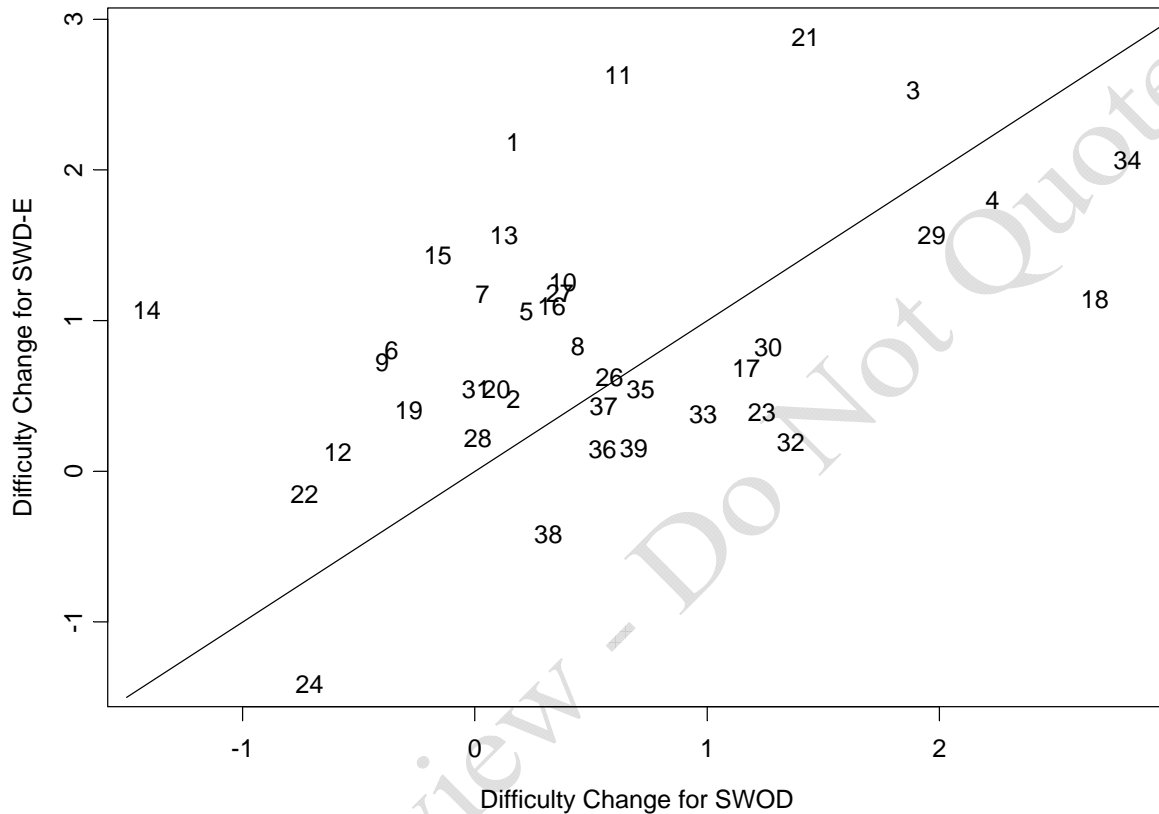
Graphic analyses were used to explore which modifications were common to very effective items, as well as which were common to very ineffective items. Figure 3 illustrates the difference in difficulty between each item in its Original condition and the corresponding item in its Modified condition, for the reading test. These differences are plotted with respect to the x-axis for students in the SWOD group and the y-axis for students in the SWD-E group. Items that fall above the diagonal show a greater reduction in difficulty due to modification for the SWD-E group, while items below the diagonal indicate a greater reduction in difficulty for the SWOD group. A positive value on either axis implies a decrease in difficulty when the item was changed from its Original condition to its Modified condition. An item such as number 21, for example, shows a much greater decrease in difficulty for students in the SWD-E group (about 3.0) compared to students in the SWOD group (about 1.3).

Based on visual inspection of these figures and like figures, comparing the SWD-E group with the SWD-NE and SWOD groups, we identified four reading items and five mathematics items with evidence of an interaction paradigm in the desired direction. We also identified three items from each content area with evidence of an interaction paradigm in the opposite direction. We then examined these items for common features.

Applying this framework, the modification that most consistently differentiated good items from bad items was shortening of the question stem. Among the nine highly effective items, seven were modified to have shorter question stems (as quantified by word count). In contrast, only one of the six highly ineffective items featured shorter question stems as modifications. No other modification, analyzed on its own, consistently distinguished items that reduced the gap from items that increased the gap. The only other consistent trend among these data indicates that adding images is an ineffective modification to reading items. Among the four highly effective items, none were modified with the addition of a graphic. In contrast, all three of the highly ineffective reading items were modified with the addition of a graphic.

Figure 3.

Rasch difficulty change for the SWD-E group by difficulty change for the SWOD group by item in Reading



Discussion

Recent changes in *NCLB* allowing that a small group of students may be counted toward AYP proficiency through an AA-MAS inspired the current study, by which we answered two questions about modified items and tests. Our hypotheses were supported, as no significant differences in reliability were found between tests in the Original condition and the Modified condition, and an interaction paradigm in item difficulty indicated that the collective effects of modifications were more helpful for those students who would be eligible for an AA-MAS than for those who would not be eligible. Exploratory analyses provided insight into individual modifications, applied within a set that was strongly grounded in theory, which consistently reduced the gap in performance between eligible students and non-eligible students.

Modifications and Reliability

We designed the current study in part to determine whether reliability varies as a function of condition or group. This is a similar question to what a meta-analyst might ask: Is a common

reliability parameter estimated by all studies of the measurement instrument? Do sample reliability coefficients appear similar across studies? In this context, each combination of condition and group might be considered an individual study or a replication of an attempt to measure reading or mathematics knowledge.

While we predicted no differences in reliability for the current study, results indicated some significant main and interaction effects. Both reading and mathematics tests yielded significant group by condition interactions, with similar patterns. The variation among reliabilities by group was larger in the Modified and Modified with Reading Support conditions than it was in the Original condition. Although these differences were significant, they were not large enough to be meaningful. When looking at sets of comparable length (39 items) and based on scores with comparable variability, the differences found in the significant interactions were typically cut in half, amounting to all differences being less than .06. In reading, the adjusted reliabilities ranged across groups and conditions between .88 and .94, acceptable magnitudes for use on an individual decision making level. The mathematics adjusted reliabilities ranged between .85 and .90, approaching that same standard. These are relatively minor differences in reliability, suggesting that systematic modifications can be made to a test without undermining the consistency of scores yielded by students from groups of various ability levels.

The additional effects found were also relatively small and of little interest, since they resulted from design characteristics that were balanced across sets and randomized across participants. The condition part interaction with the reading score reliabilities created differences among test part of less than .02 across study conditions; whereas the significant set effect resulted in differences of less than .005. Similarly, the set x group interaction found with mathematics score reliabilities resulted in study group differences of less than .05 (mostly about .02); whereas the significant part effect resulted in differences of less than .009. In both tests, set resulted in some effect, indicating the difficulty in establishing parallel 13-item sets—however, these effects were negligible. Part also indicated some effect on alpha, suggesting the importance of rotating experimental items in tests to avoid the impact of fatigue and effort.

Item Difficulty and Interaction Paradigm

We also designed the current study to determine whether modifications could help SWD-Es perform more like SWODs and like SWD-NEs. This issue was examined by using a linking procedure within the Rasch model, which allowed for difficulty levels in the Original condition to be equated across groups, controlling for differences in student ability. The mean and variance of items under each of the Modified and Modified with Reading Support conditions showed how the difficulty estimates on average changed for each group.

In support of our hypothesis, the difficulty estimates decreased in both content areas when comparing the Original and Modified conditions. This effect was particularly large for students in the SWD-E group, as their score improvement from the Original condition to the Modified condition in both reading and mathematics was significantly greater than the improvement of students in the SWOD group. Also in reading, scores of students in the SWD-E group improved significantly more than those in the SWD-NE group. In Mathematics, scores of students in the

SWD-E group improved significantly more than students in the other two groups when comparing the Original condition and the Modified with Reading Support condition.

These findings indicated that an interaction paradigm was occurring, as students who would be eligible to take the test experienced much larger reductions in item difficulty than students who would not be eligible experienced. Over the course of a test, these reduced item difficulties should result in higher scores for students who would be eligible. To the degree that this boost in performance is differential by eligibility, the modifications used in the current study appear to work like valid testing accommodations, providing access for eligible students to the same opportunity to show what they know and are able to do. Both groups of students in the current study who would not be eligible for an AA-MAS also experienced average reductions in item difficulty, but those reductions were much smaller than the reductions experienced by the SWD-E group. Some degree of reduction in difficulty across groups based on modifications is allowable within the current policy if the grade level of the test is maintained.

Modification Strategies

The success of the current study, as evidenced by stable reliabilities across conditions and by a differential reduction in item difficulty that favored those students who would be eligible, supports the modifications that were used and strengthens the theories on which those modifications were based. Multiple modifications were incorporated to convert each item from the Original condition to the Modified condition. The most common modifications included removal of a response option, simplification of language, addition of graphic support, and reorganization of layout to increase white space and facilitate the integration of visuals. These modifications were made with the intent of reducing cognitive load, maintaining or increasing student motivation, and providing accessibility to the test for students who would be eligible. Removing a response option, simplifying language, and reorganizing layout are all modifications that are theoretically consistent with cognitive load theory, as they are intended to reduce the amount of information that the student has to maintain in working memory. Using three response options rather than four and simplifying language clearly are modifications that reduce the amount and complexity of verbiage test-takers must store in memory while choosing an answer. Further, to the degree that the eliminated answer choice is implausible and that essential content is preserved during language simplification, the cognitive load that is reduced is extraneous. Reorganizing a layout is also intended to reduce cognitive load, by breaking into smaller chunks the amount of information that has to be held at a moment in time. Adding graphic support is a more complicated modification, intended to clarify complicated vocabulary and passages by providing context, and possibly to increase interest. Results of the current study indicate that adding graphic support to reading items may increase cognitive load. All of these modifications are intended address accessibility by making the modified test as experienced by eligible students more similar to the regular test as experienced by general education students and students identified with disabilities who would not be eligible.

Universal design and cognitive load theories are further invoked by the modification of added reading support, which yielded reliable scores and greater effects across almost all group and content area combinations. Reading support is highly effective for students with disabilities, particularly including those who would be eligible for an AA-MAS, because many have reading

problems. This was evident in our CAAVES project cognitive lab (Roach et al., 2008), and is replicated by the current findings. Reading support in the current study was used conservatively; it focused primarily on reading directions, question stems, and response choices, and was not used for reading passages or vocabulary words. While this modest support had a small impact across groups and content areas, it is likely that more extensive reading support would result in greater impact. The practical implications of findings with regard to reading support are complicated, however, by the diversity of state policies with regard to the perceived validity of reading support as an accommodation or modification.

Because the modifications used in the current study were always used in combination rather than individually, conclusions cannot easily be drawn about the validity of any one modification versus the others. We conducted an analysis to explore the effects of specific modifications made by examining 15 items across content areas, nine of which were particularly effective at reducing the gap in performance between eligible students and noneligible groups of students, and six of which actually increased this gap in performance. We analyzed the items for similarities and differences with regard to the modifications used on each. Theoretically, modifications that were consistently applied across effective items would be considered promising, while those consistently applied across ineffective items would be concerning. Two specific modifications – removal of a distractor and adding white space between response options – could not be evaluated through this analysis of modified items, only because they were applied uniformly to all reading and mathematics items.

Very few patterns emerged when comparing the modifications used among the items that were effective at reducing the discrepancy between groups versus the items that were ineffective and increased the discrepancy. This finding reinforces the thinking that modifications must be made thoughtfully on an item-by-item basis, because one change could make one item simpler and more accessible, while the same change could have the opposite effect when applied to a different item. Reduction of question stem length was the one modification that appeared to be effective across content area. Question stem length (as characterized by number of words) was reduced in seven of the nine highly effective items, but was only reduced in one of the six ineffective items. This finding indicates that students hold the question stem in their working memories while solving the item, and that shorter stems involve less working memory. In many cases, the stem was shortened by only a word or two, reinforcing the contention that eligible students tend to be very poor readers (Roach et al., 2008), and are therefore sensitive to small changes in reading load. Shortening the question stem should be considered a promising modification when developing an examination for students eligible for an AA-MAS.

The only other pattern apparent from this item level analysis was that adding graphic support to reading items may be a harmful modification for eligible students. Graphic support was added to all three of the ineffective items, and was not added to any of the four highly effective items. In many cases, the inclusion of graphics may add extraneous cognitive load to an item, reducing access for eligible students who tend to be poorer readers. While it is possible that in some cases adding graphic support might be helpful for students, this finding suggests that test developers should use caution when considering adding visuals to reading items. No similar pattern was found for the modification of adding graphic support to mathematics items.

It should be noted that for the reading items in this analysis, visuals were included primarily for interest and motivational reasons and did not include essential information for responding to the item. It is conceivable that visuals that support essential item content and thereby facilitate access could be added to reading items. This is consistent with the advisement of Clark et al. (2006), who indicate that optimal cognitive load is achieved when only necessary visuals are included, and extraneous visuals are eliminated.

Federal Policy and Participation Criteria

The findings from this study indicate that a modified examination can be made, and that appropriate students can be identified to take the test. The Modified condition was easier for all students, and particularly for students who would have been eligible. Our modifications were made by state level assessment leaders, who were familiar with the policy and guarded against changes that would take the test content off grade level. The likely impact of modifications would be that more students could meet proficiency, although the magnitude of that change would depend on the standard setting processes that follow tests like these. The findings also indicate that at an eighth-grade level, participation criteria directly abstracted from federal regulations can be used to identify three distinct groups of students, as indicated by the fact that the groups in our study performed at distinct levels across conditions (Elliott et al., 2008).

Limitations and Implications for Future Research

A sampling limitation in the current study was that it included only eighth grade students. The findings warrant replication at elementary and high school levels where participation decisions are potentially more challenging due to a lack of previous years' proficiency tests.

A second limitation of note is that testing accommodations were not used in the current study. Although this design allowed for the examination of the impact of modifications without accommodations having to be considered as a covariate, in practice testing accommodations would be allowable along with modifications for the AA-MAS. Future researchers should examine the relationship between modifications and accommodations and their effects on test scores for students with and without disabilities.

Conclusions

The current findings indicate that test developers can effectively meet the demand of new federal policy by designing modified versions of statewide achievement tests for eligible students which will provide access to assessment of grade level content of modified achievement standards. These tests can be just as reliable as the original versions of tests and can produce interaction effects whereby the item difficulties experienced by eligible students are reduced much more than they are for students who would not be eligible, ultimately yielding an interaction paradigm in test performance. These findings indicate that increasing accessibility to the grade level assessment is a necessary step in providing access to grade level content. This increase in accessibility is a primary purpose of developing an AA-MAS, and these findings support the theoretical and data-based processes used to modify test items to meet this goal.

References

- Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test Accessibility and Modification Inventory*. Retrieved September 20, 2008 from: <http://peabody.vanderbilt.edu/tami.xml>
- Center for Universal Design. (1997). *The principles of universal design*. Retrieved August 4, 2008 from <http://www.design.ncsu.edu/cud>
- Clark, R., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. San Francisco, CA: Pfeiffer.
- Elliott, S.E., Kettler, R.J., McGrath, D., Compton, E., Bruen, C., Hinton, K., Palmer, P., Beddow, P.A., & Kurz, A. (2008). Performance of students with persistent academic difficulties on original and modified multiple-choice items: Promising possibilities for alternate assessments and beyond. Unpublished manuscript. Peabody College of Vanderbilt University.
- Hollenbeck, K. (2002). Determining when test alterations are valid accommodations or modifications for large-scale assessment. In G. Tindal & T. Haladyna (Eds.), *Large scale assessment programs for all students* (pp. 109-148). Mahwah, NJ: LEA.
- Johnstone, C., Liu, K., Altman, J., & Thurlow, M. (2007). *Student think aloud reflections on comprehensible and readable assessment items: Perspectives on what does and does not make an item readable* (Technical Report 48). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Johnstone, C., Thurlow, M., Moore, M., & Altman, J. (2006). *Using systematic item selection methods to improve universal design of assessments* (Policy Directions 18). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Koretz, D.M. & Hamilton, L.S. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., 531-578). United States of America: American Council on Education and Praeger Publishers.
- Lazarus, S.S., Thurlow, M.L., Christensen, L.L., & Cormier, D. (2007). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2007* (Synthesis Report 67). National Center on Educational Outcomes.
- Miller, G.A. (1956). The magic number seven plus or minus two: some limits on our capacity to process information. *Psychological Review* 63: 81-97.
- Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, 7, 93-120.
- Phillips, S.E., & Camara, W.J. (2006). Legal and ethical issues. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., 733-757). United States of America: American Council on Education and Praeger Publishers.
- Roach, A.T., Elliott, S.E., Kettler, R.J., Beddow, P.A., Rodriguez, M.C., Compton, E., & Palmer, P. Using student responses and perceptions to inform item development for an alternate assessment based on modified achievement standards. Unpublished manuscript. Georgia State University.
- Rodriguez, M.C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13.
- Rodriguez, M.C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, 11(3), 306-322.

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (revised July, 2007). *Standards and assessments peer review guidance*. Washington, D.C.: Author.

Under Review - Do Not Quote

Footnotes

¹The current study was implemented as part of the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) project, a multi-state project funded by the U.S. Department of Education (Award to Idaho Department of Education; #S368A0600012). The positions and opinions expressed in this article are those solely of the author team. We wish to acknowledge the excellent state leadership and data collection efforts by Charles Bruen in Arizona, Kent Hinton in Hawaii, Elizabeth Compton in Idaho, and Dawn McGrath in Indiana. Without these individuals' coordination and support efforts this study would not have been possible.

²The *Test Accessibility and Modification Inventory* is the second version of an unpublished instrument originally titled the *Item Accessibility and Modification Guide*.

Under Review - Do Not Quote