

**Using Modified Items to Test Students with and without Persistent Academic Difficulties:
Effects on Groups and Individual Students**

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

Elizabeth Compton
Idaho Department of Education

Dawn McGrath
Indiana Department of Education

Charlie Bruen
Arizona Department of Education

Kent Hinton
Hawaii Department of Education

Porter Palmer
Discovery Education Assessment

Michael C. Rodriguez
University of Minnesota, Twin Cities

Daniel Bolt
University of Wisconsin-Madison

Andrew T. Roach
Georgia State University

Please direct all correspondence concerning this manuscript to Stephen N. Elliott, 404 Wyatt Center, Peabody #59, 230 Appleton Place, Nashville, TN 37203-5721, email: steve.elliott@vanderbilt.edu. Telephone: (615) 322-2538, fax: (615) 322-4488.

Using Modified Items to Test Students with and without Persistent Academic Difficulties:

Effects on Groups and Individual Students

Under Review - Do Not Quote Without Permission

Abstract

The goal of the study was to investigate the effects of using modified items in reading and mathematics achievement tests to enhance accessibility, simplify language, and reduce memory loads without changing the assessed depth of knowledge. An experimental design was used to determine whether tests composed of modified items would display the same level of reliability as tests composed of original (unmodified) items, and also help reduce the performance gap between students who would be eligible for an alternate assessment of modified achievement standards (AA-MAS) and students who would not be eligible. Three groups of eighth-grade students ($N = 755$) defined by eligibility and disability status from four states took original and modified versions of reading and mathematics tests. The findings indicated that the modified item conditions were significantly easier for all students, and particularly for students who met eligibility criteria for an AA-MAS. It was concluded the impact of our test modifications was that more students eligible for an AA-MAS would likely meet proficiency. Study limitations and need for follow-up research on the effect of item modifications on test accessibility and the performance of students with disabilities are discussed.

**Using Modified Items to Test Students with and without Persistent Academic Difficulties:
Effects on Groups and Individual Students**

Under federal law, states may report proficiency for a subset of students identified with disabilities using alternate assessments based on modified achievement standards (AA-MAS). As specified in amendments to the *No Child Left Behind* act (NCLB; U.S. Department of Education, 2007), students with Individual Education Plans (IEPs) focusing on 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 are eligible for grade-level achievement tests that are similar in content covered, but technically easier than the general education test required of classmates. To create an AA-MAS, acceptable changes or enhancements to the general education test result in enhanced universal access, reduced cognitive load, and improved validity of test score inferences about the same constructs measured by the general education achievement test.

The purpose of the current study was to investigate the effects of using modified items in reading and mathematics tests to enhance accessibility, simplify language, and reduce memory demands without altering the depth of knowledge the items assess. An experimental design was used with eighth graders to determine whether tests composed of modified items would display the same level of internal consistency as tests composed of original (unmodified) 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. This experiment was part of the research conducted by the Consortium of Alternate Assessment Validity and Experimental Studies¹ (CAAVES) and involved students from Arizona, Hawaii, Idaho, and Indiana.

The Educational Policy Contexts for the Use of Testing Modifications

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, if deemed eligible according to state guidelines by IEP teams, can take a version of the general education achievement test that their same grade-level peers take, but with modifications. By definition, this new version of the general education achievement test would constitute a different test and would also require new cut scores for determining proficiency levels for NCLB reporting. We believe acceptable modifications are changes to a test's content or item format that make a test more accessible for most students while continuing to assess grade-level content and skills at the same depth of knowledge (DOK) as unmodified items, resulting in an overall test that is less difficult. 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 an unaccommodated or unmodified general education 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 test (Kettler et al., 2008).

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. These modified assessments are for students for whom (a) the standards of the general education assessment are generally too difficult and (b) the standards of the alternate assessment for students with significant cognitive disabilities are generally too easy. 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 in the legislation allows that more 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. Accurately identifying students requires a clear understanding of the modifications that are made to items from a regular assessment to create an AA-MAS.

Item and Test Modifications: Theory to Practice

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 education curriculum. With the aim of increasing item accessibility, the current study employed principles of universal design (UD), cognitive load theory (CLT), and research on item development. Table 1 includes a list of salient UD principles, along with definitions and guidelines that correspond to each principle. Although not intended specifically for testing, these UD principles are wise to consider when developing any test. The principle of simplicity and intuitiveness, for example, means that the environment, product, or service should be easy to understand. One of the principles' 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 UD principles, along with empirical research on test item writing and cognitive load theory, guided the development of the Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, & Elliott, 2008), a tool that was used to guide the development of items for the tests in this research study.

Insert Table 1 about here

Modifications to general assessments that are designed to help students who consistently fail to meet proficiency can also be guided by CLT. Conceptualized by Sweller (Sweller, 1988),

CLT has been applied to classroom instruction and materials to improve efficiency and student learning (Clark, Nguyen, & Sweller, 2006; Sweller, van Merriënboer, & Paas, 1998). 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 of instruction. 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 (Paas & van Merriënboer, 1994). Extraneous load is memory load unrelated to the task. Clark et al. 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. The resulting CLT research base can be applied to the development of test items that students can process more efficiently.

Research literature on item development also influenced some of our item accessibility and modification tactics. In particular, one 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?" 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)

Several states are using this strategy for item modification for students with disabilities, and 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 these are truly in the spirit of making items more accessible to students with disabilities.

The National Center on Educational Outcomes (Lazarus, Thurlow, Christensen, & Cormier, 2007) released a report describing the most common modifications used in the first six states that implemented an AA-MAS. They indicated the most common modifications were: (a) removing a distractor from a MC item, (b) reducing the number of items on the test, (c) simplifying language, and (d) reducing the number and length of reading passages. A separate set of modifications that many professionals might consider testing accommodations was also included, with reducing the number of items per page and increasing the font size being the most commonly considered alterations. A number of the modifications and accommodations listed in the NCEO report were used for tests in the current study.

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 identified with disabilities (SWDs) versus students not identified with disabilities (SWODs). An interaction paradigm is the assertion that if an accommodation is working, it should increase the scores of SWDs, but should generally not affect the scores of SWODs (Phillips, 1994). Although

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 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. Our use of the term *modification* is conservative, reflecting no expectations related to change or consistency in the construct being measured or the inferences that can be made from resulting scores.

Recent Research Findings Leading to the Current Research Question

The current report is based on research from the CAAVES project. This project focused on modified multiple-choice items and involved a four state collaboration with the primary goal of investigating the feasibility of item accessibility and modification strategies for use in alternate achievement tests for students with disabilities and persistent academic difficulties. 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 indicated in our review 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 they hypothesized would be more accessible, but still reflect the same grade-level content as the original items. The CAAVES team implemented this multi-state experimental

study in the spring of 2008 to compare the effects of tests with and without modified items on students' test performances and test score comparability across three groups: students not identified with a disability (SWODs), students identified with a disability who would not be eligible for an AA-MAS (SWD-NEs) and students identified with disabilities who would be eligible for an AA-MAS (SWD-Es). Two main research questions that concerned the psychometrics of the modified tests were answered. The questions and a summary of supportive findings from the CAAVES project follow:

How do modifications affect the internal consistency of test scores? Based on previous research regarding the elimination of one distractor and the relatively conservative nature of our other modifications, we predicted no difference in reliability between the original condition, the modified condition, and the modified with reading support condition. This prediction was addressed and generally supported as documented by Kettler, et al. (2008). Specifically, the 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. In mathematics, adjusted reliabilities ranged between .85 and .90, approaching that same standard. These were relatively minor differences in reliability, suggesting that systematic modifications can be made to a test

without undermining the internal consistency of scores yielded by students from groups of various ability levels.

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 students in the SWD-E group would benefit more than those in the SWOD or SWD-NE groups. Kettler et al. (2008) examined this issue 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 this prediction, 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 the mean item difficulty for the group decreased from the original condition to the modified condition in both reading and mathematics to a significantly greater degree than the decreases experienced by other groups. 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.

Given the promising psychometric evidence for the CAAVES developed test forms

examined by Kettler and associates (2008), it is reasonable to take the next step and examine the empirical evidence that addresses practical questions about group performance differences and likely effects on the proficiency level for the students deemed eligible to take an AA-MAS. Thus, in the present study we examined the performances of students across the three groups to answer two questions: (a) Do eligible students perform better on tests comprised of highly accessible, modified items than on the original tests? and (b) If the performances of eligible students improve on tests comprised of modified items, what percentage of the students are likely to perform at a level deemed proficient in reading or mathematics?

Method

Participants

The sample included 755 eighth-grade students from four states, balanced across the three groups: SWOD ($n = 269$), SWD-NE ($n = 236$), and SWD-E ($n = 250$). Eighth grade was selected because (a) students were expected to 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 students in the SWD-NE group (65%) and the SWD-E group (64%). 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.

Insert Table 2 about here

Among students identified with disabilities, the largest percentage was identified as having a Specific Learning Disability (60%), and a representative subsample were identified as having 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 SWD groups.

Insert Table 3 about here

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 (a) goals and statements aligned to grade level content standards in reading, language, mathematics, or science; (b) IEP statements that show that the instructional material or curriculum contained grade level content; and (c) 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. The *CAAVES AA-MAS Participation Criteria Decision Guide* is available at <http://peabody.vanderbilt.edu/x8312.xml>.

Test Accessibility and Modification Inventory. The *TAMI* (Beddow et al., 2008) was designed to facilitate a comprehensive analysis of test items with the purpose of enhancing access and meaningful responses from all students. *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.

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 *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 *TAMI* used in this study included only individual item considerations, but the published revision includes considerations for designing accessible computer-based tests.

Reading and mathematics tests. Across conditions, each test in reading and mathematics was composed of 39 computer-based multiple-choice items. Original 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 identify meaning of words or phrases 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 data analysis with arithmetic operations. Each test was further divided into three 13-item sets 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 a pre-publication draft of *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. A subset of items was then pilot tested with a small group of students within a cognitive lab, after which the entire group evaluated all 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 automatically read item directions and stems. 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 internal consistency 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. The alpha adjusted for score variance-heterogeneity was .70, with a standard deviation of .02. The alpha based on the SB adjustment to a 39-item test was .88.

Procedure

Research personnel from the four state departments of education who were CAAVES partners selected a minimum of two school districts that collectively had at least 150 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). Research personnel from the four state departments of education who were CAAVES partners accessed their previous year's general education achievement test results for students with identified disabilities. State project leaders then targeted a representative sample of school districts to participate in the experimental study by identifying schools with 8th grade students with disabilities, sharing information about the study and participation criteria, and asking them to apply the criteria to rosters of current 8th grade students with disabilities in the school. Special education teachers and administrators were asked to review the criteria for all of their students with IEPs and to indicate students who met all three criteria for participation in an AA-MAS (SWD-Es) and those who clearly did not meet all three criteria (SWD-NEs). Administrators were also asked to identify another group of students without an identified disability who performed at the various levels of proficiency on the general education test. After special education teachers and their administrator agreed on the final list, it was forwarded to the research personnel from state departments of education. State leaders then conducted a final review and collected student demographics to ensure a representative sample of students. 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. No students were omitted as a result of this requirement. 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. 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.

Data Analyses

To address the questions motivating this study, we provide a set of descriptive statistics for the test performances of each of the three groups of students (i.e., SWOD, SWD-NE, and SWD-E) on the three types of test conditions (original, modified, and modified with reading support). We used two multivariate, repeated-measures analyses of variance (MANOVAs) to test for differences among the groups under these various test conditions. Group was a between-subjects variable, and condition was a within-subjects variable. Effect size for group status and condition was calculated as η^2 , the percent of variance in test scores explained by group status.

Effect sizes for condition were calculated as mean differences between each modified and referent condition, divided by the standard deviation of the referent condition, so that the impact of modifications could be interpreted in standard deviation units. For the effect of the modified and modified with reading support conditions, they were compared respectively with the original condition. For the additive effect of reading support, the modified with reading support condition was compared with the modified condition.

Finally, we provide descriptive information concerning individual student performance using two methods: (a) based on movement across a number of common percentile ranks as criteria or cut-scores proxies for proficient performance decisions, and (b) based on changes in student ability estimates across different conditions.

Results

Out of a possible 39 reading items, the total sample of participants in the study had a mean raw score of 23.93 ($SD = 8.02$), corresponding to 61% correct. Out of a possible 39 mathematics items, the total sample of participants mean score was 21.00 ($SD = 7.50$), corresponding to 54% correct. Findings for the three groups of students and individual students are discussed in the rest of this section.

Group Effect

Main effects for group were found in both reading, $F(2, 718) = 168.58, p < .05$, and mathematics, $F(2, 714) = 217.54, p < .05$. In reading, the size of the main effect for group was moderate, partial $\eta^2 = .32$. The average score for the SWOD group was 1.36 points higher on a 13-point scale than was the average score for the SWD-NE group, with the 95% confidence interval for the difference being between .96 and 1.75 points. The average score for the SWD-NE group was 2.27 points higher on a 13-point scale than the SWD-E group, with the 95%

confidence interval for the difference being between 1.87 and 2.68 points. This pattern was similar when data were disaggregated by state, except for Arizona, in which the difference between the SWD-NE and SWOD groups was nonsignificant. Table 4 and Figure 1 depict raw score means in reading across groups and conditions.

Insert Table 4 about here

In mathematics, the size of the main effect for group was large, partial $\eta^2 = .38$. The average score for the SWOD group was of 1.57 points higher on a 13-point scale than the average score for the SWD-NE group, with the 95% confidence interval for the difference being between 1.21 and 1.92 points. The average score for the SWD-NE group was 2.13 points higher on a 13-point scale than the average score for the SWD-E, with the 95% confidence interval for the difference being between 1.77 and 2.49 points. This pattern was similar when data was disaggregated by state, except for Arizona, in which the difference between the SWD-NE and SWOD groups was nonsignificant. Table 5 and Figure 2 depict raw score means in mathematics across groups and conditions.

Insert Table 5 about here

Condition Effect

Main effects for condition were found in both reading, $F(2, 1436) = 147.45, p < .05$, and mathematics, $F(2, 1428) = 119.53, p < .05$. In reading, the size of the main effect for condition was moderate, partial $\eta^2 = .17$. Students scored higher in the modified condition than in the original condition, with highest scores in the modified with reading support condition. The average difference between scores in the modified condition and scores in the original condition (i.e., the modification effect) was .38 standard deviations. The average difference between scores in the modified with reading support condition and scores in the original condition (i.e., the

modification plus reading support effect) was .46 standard deviations, and the average difference between scores in the modified with reading support condition and scores in the modified condition (i.e., the reading support effect) was .07 standard deviations. Results were similar when data were disaggregated by state. Table 6 depicts effect sizes by condition.

Insert Table 6 about here

In reading, the size of the main effect for condition was small, partial $\eta^2 = .05$. Students scored higher in the modified condition than in the original condition, with the highest scores in the modified with reading support condition. The average difference between scores in the modified condition and scores in the original condition (i.e., the modification effect) was .21 standard deviations. The average difference between scores in the modified with reading support condition and scores in the original condition (i.e., the modification plus reading support effect) was .25 standard deviations, and the average difference between scores in the modified with reading support condition and scores in the modified condition (i.e., the reading support effect) was .05 standard deviations. Results were similar when data were disaggregated by state. Table 6 depicts effect sizes by condition.

Group by Condition Interaction

The interaction between group and condition was not significant in either reading or mathematics. Figure 1 and Figure 2 depict average test scores by group and condition in reading and mathematics, respectively. Given the exploratory nature of this research, however, it is worth noting that the direction of effects (although not significant) are in the direction as expected. That is, SWD-E's showed the greatest increase in sample means when moving from the unmodified to modified conditions.

Insert Figure 1 and Figure 2 about here

Individual Effects

The practical impact of modifications on participants in the current study varied based on the criterion or cutoff proxy used for proficiency. From a descriptive analysis depicting the increase in the percentage of students who would become proficient at a series of common percentile ranks (30th, 40th, 50th, and 60th percentiles), it is clear that the effect of modification for SWD-Es is greatest at the lowest criterion. Percentile ranks were determined based on the distribution of scores for students in the SWOD group in the original condition. When using the 30th percentile as the criterion, 18% and 10% more students in the SWD-E group were proficient in the modified condition compared to the original condition in reading and mathematics, respectively. These percentage differences ranged from 3% to 11% in reading and 3% to 5% in math at the other percentile rank criteria. Increases for the SWOD ranged from approximately 15% to 20% across content areas and criteria, and the pattern for students in the SWD-NE group indicated there was a greater impact of the modified tests at the lower criteria, but not to the same extent that this impact was apparent among students in the SWD-E group. These patterns likely indicate that, because of group differences in performance and the basis of percentile ranks for the SWOD group, the higher criteria made proficiency too high for students in the SWD-E group to attain even with modifications. Table 7 depicts the percentage of students who would be considered proficient by criterion and group, for the original and modified tests.

Insert Table 7 about here

When controlling for individual differences in ability level, students in the SWD-E group were slightly more likely to experience large improvements based on modifications than were those in the SWOD or SWD-NE groups. Using a Rasch model to control for differences in individual abilities, the percentage of students who experienced large increases (greater than 1

standard error of measure) was higher for the SWD-E group than it was for other groups, regardless of content area or modification type (i.e., modification effect, modification plus reading support effect, or reading support effect). The percent of students experiencing large increases ranged from 12% to 18%. Table 8 depicts the percentages of students experiencing reliable increase by content area, group, and modification type.

Insert Table 8 about here

Discussion

The new NCLB policy that allows a small percentage of students with disabilities to be counted toward AYP proficiency through an alternate assessment of modified achievement standards or AA-MAS motivated this quasi-experimental study. We were interested in determining how students with and without disabilities functioned on grade-level tests enhanced with simplified and highly accessible items in comparison to unaltered tests (i.e., tests comprised of the same items prior to simplification and accessibility enhancements). In particular, we asked: (a) Do eligible students perform better on tests comprised of highly accessible, modified items than on the original tests? and (b) If the performances of eligible students improve on tests comprised of modified items, what percentage of the students are likely to perform at a level deemed proficient in reading or mathematics? Using tests known to be reliable and comprised of original items and modified items assessing the same depth of knowledge, we empirically answered the first question and provided data-based speculations for an array of proficiency cut-score options.

With regard to the first question, the summary findings indicated that modified on-grade level tests of reading and mathematics could be constructed, and that appropriate students with disabilities could be identified to take the test. The modified item conditions for both the reading and mathematics tests were significantly easier for all students, and particularly for students who met eligibility criteria based literally on the federal legislation. Our item modifications were made by state assessment leaders, who were familiar with the policy and guarded against changes that would result in test content that was off grade level. The modifications were presented as a package and could not be disaggregated to determine which ones account for the resulting effects. In a corresponding item analysis study, however, IRT analyses indicated that key modifications likely involve shorter item stems, boldfaced signal words, and language simplification for both reading and math items, and the use of clear supporting visuals for only math items (Kettler, et al., 2008). The modifications of increased white space and the use of 3 rather than 4 answer choices was used with all modified items and thus their effects on individual items could not be determined. Suffice it to say at this point in test and item modification research, this study contributes evidence to support that it is a viable approach to on-grade level testing for students with disabilities who have a history of non-proficient test performances.

Concerning the second question, the likely impact of modifications would be that more students eligible for an AA-MAS could meet proficiency, although the magnitude of that change would depend on the standard setting processes and actual cut scores that follow state-wide tests more comprehensive than those used in this study. As one would expect, the lower the cut-scores, the larger the percentage of students exceeding any given cut-score. Our effort to shed light on this aspect of large-scale assessment was admittedly limited in this study, but with the array of cut-scores provided for our reading and mathematics tests one can begin to understand that even

with an AA-MAS a significant portion of students with disabilities still are not likely to be deemed proficient.

Limitations and Implications for Future Research

A sampling limitation in the current study was that it included only eighth-grade students. The findings with our multi-state sample of eighth graders 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 concerned the fact that our tests were rather short when compared to typical achievement tests and the mathematics test only covered content for the domains of numbers and operations and algebra. These design features were necessary for the CAAVES project given time limitations and the desire to have test content that aligned well with the four participating states' content standards. It would seem that these design features may have functioned to minimize the differences between the modified and unmodified tests, but this hypothesis will need to be examined in a future study where comprehensive modified tests can be developed and compared with existing unmodified tests.

A final 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 findings from this study indicated that items that undergo comprehensive modifications to increase accessibility resulted in improved performances for most all students, in particular for students who meet the new federal policy for participating in grade level alternate assessments of modified achievement standards. Initial versions of such tests can be just as reliable as the original versions of tests and can produce effects whereby the item difficulties experienced by eligible students are reduced more than they are for students who would not be eligible. Moreover, evidence from the current study indicated students who would be eligible for an AA-MAS are the most likely to benefit from the modifications made with purpose of enhancing accessibility. This finding confirms the implicit assumption of current NCLB policy; namely, that students for whom current grade-level assessments do not permit unfettered access may benefit from the development of more accessible tests. 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. Whether the improvements in test performance on more accessible tests will be great enough that a majority of eligible participants meet proficiency standards remains to be determined. In other words, the important question of whether the difference in test performance on an AA-MAS is great enough to make a significant difference in the percentage of students who are proficient remains to be answered.

References

- Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). Test Accessibility and Modification Inventory (TAMI). Nashville, TN: Vanderbilt University. Retrieved November 28, 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.
- 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.
- Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (2008). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and the interaction paradigm. Unpublished manuscript. Peabody College of Vanderbilt University.

- 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., Beddow, P. A., Kurz, A., Kettler, R. J., & Elliott, S. E. (2008). 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.
- 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.

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.

Under Review - Do Not Quote Without Permission

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.)
-

Table 2

Demographics by Group Membership

	SWOD	SWD-NE	SWD-E	Total
	<i>n</i> (%)	<i>n</i> (%)	<i>n</i> (%)	<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

	SWD-NE	SWD-E	Total
	<i>n</i> (%)	<i>n</i> (%)	<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.

Table 4

Means, Standard Deviations, and Percent Correct for Reading by Group and Condition

	Original	Modified	Modified with Reading Support	Total
SWOD	8.83 (2.66)	9.96 (2.35)	10.00 (2.56)	28.80 (6.61)
(n = 256)	68%	77%	77%	74%
SWD-NE	7.36 (2.94)	8.52 (2.85)	8.85 (2.55)	24.73 (7.20)
(n = 228)	57%	66%	68%	63%
SWD-E	5.05 (2.21)	6.26 (2.58)	6.59 (2.56)	17.91 (6.03)
(n = 237)	39%	48%	51%	46%
Mean	7.13 (3.05)	8.29 (3.01)	8.52 (2.92)	23.93 (8.02)
(n = 721)	55%	64%	66%	61%

Table 5

Means, Standard Deviations, and Percent Correct for Mathematics by Group and Condition

	Original	Modified	Modified with Reading Support	Total
SWOD	8.36 (2.84)	8.81 (2.49)	8.95 (2.53)	26.14 (6.50)
(<i>n</i> = 256)	64%	68%	69%	67%
SWD-NE	6.68 (2.66)	7.31 (2.64)	7.45 (2.60)	21.44 (6.42)
(<i>n</i> = 223)	51%	56%	57%	55%
SWD-E	4.44 (2.13)	5.25 (2.07)	5.38 (2.21)	15.05 (4.63)
(<i>n</i> = 238)	34%	40%	41%	39%
Mean	6.54 (3.03)	7.16 (2.83)	7.30 (2.87)	21.00 (7.50)
(<i>n</i> = 717)	50%	55%	56%	54%

Table 6

*Effect Sizes for Reading and Mathematics by Group and Modification Condition**

	Modified	Modified with Reading Support	Reading Support Over Modified
Reading			
SWOD	.37	.38	.01
SWD-NE	.38	.49	.11
SWD-E	.40	.50	.11
Total	.38	.46	.07
Mathematics			
SWOD	.15	.20	.05
SWD-NE	.21	.25	.05
SWD-E	.26	.31	.04
Total	.21	.25	.05

* To determine the effect sizes for the Modified and Modified with Reading Support condition, the Original condition served as the point of comparison.

Table 7

Percentile Cut Score Statistics by Group and Condition

Percentile	Group	READING		MATHEMATICS			
		Cut Score	Original	Modified	Cut Score	Original	Modified
			# Above (%)	# Above (%)		# Above (%)	# Above (%)
30th	SWOD	8	185 (72.2%)	224 (87.5%)	7	162 (63.3%)	211 (82.4%)
	SWD-NE		142 (62.3%)	159 (69.7%)		115 (51.6%)	133 (59.6%)
	SWD-E		31 (13.1%)	74 (31.2%)		37 (15.5%)	60 (25.2%)
40th	SWOD	9	155 (60.5%)	201 (78.5%)	8	138 (53.9%)	177 (69.1%)
	SWD-NE		115 (50.4%)	131 (57.5%)		82 (36.8%)	111 (49.8%)
	SWD-E		23 (9.7%)	49 (20.7%)		26 (10.9%)	32 (13.4%)
50th	SWOD	9	155 (60.5%)	201 (78.5%)	9	107 (41.8%)	146 (57.0%)
	SWD-NE		115 (50.4%)	131 (57.5%)		59 (26.5%)	74 (33.2%)
	SWD-E		23 (9.7%)	49 (20.7%)		9 (3.8%)	20 (8.4%)
60th	SWOD	10	120 (46.9%)	171 (66.8%)	9	107 (41.8%)	146 (57.0%)
	SWD-NE		88 (38.6%)	97 (42.5%)		59 (26.5%)	74 (33.2%)
	SWD-E		19 (8.0%)	26 (11.0%)		9 (3.8%)	20 (8.4%)

Note. Group samples sizes were as follows: SWOD ($n = 256$ reading; $n = 256$ mathematics); SWD-NE ($n = 228$; $n = 223$ mathematics); SWD-E ($n = 237$ reading; $n = 238$ mathematics).

Each test was comprised of 13 items.

Table 8

Percentage Experiencing Reliable Increases by Group and Modification Condition

	Modified	Modified with Reading Support	Reading Support Over Modified
Reading			
SWOD	13%	7%	7%
SWD-NE	15%	14%	13%
SWD-E	15%	16%	17%
Total	14%	12%	12%
Mathematics			
SWOD	13%	16%	15%
SWD-NE	15%	14%	12%
SWD-E	16%	17%	18%
Total	14%	16%	15%

Figure 1. Mean test scores by group and condition in reading.

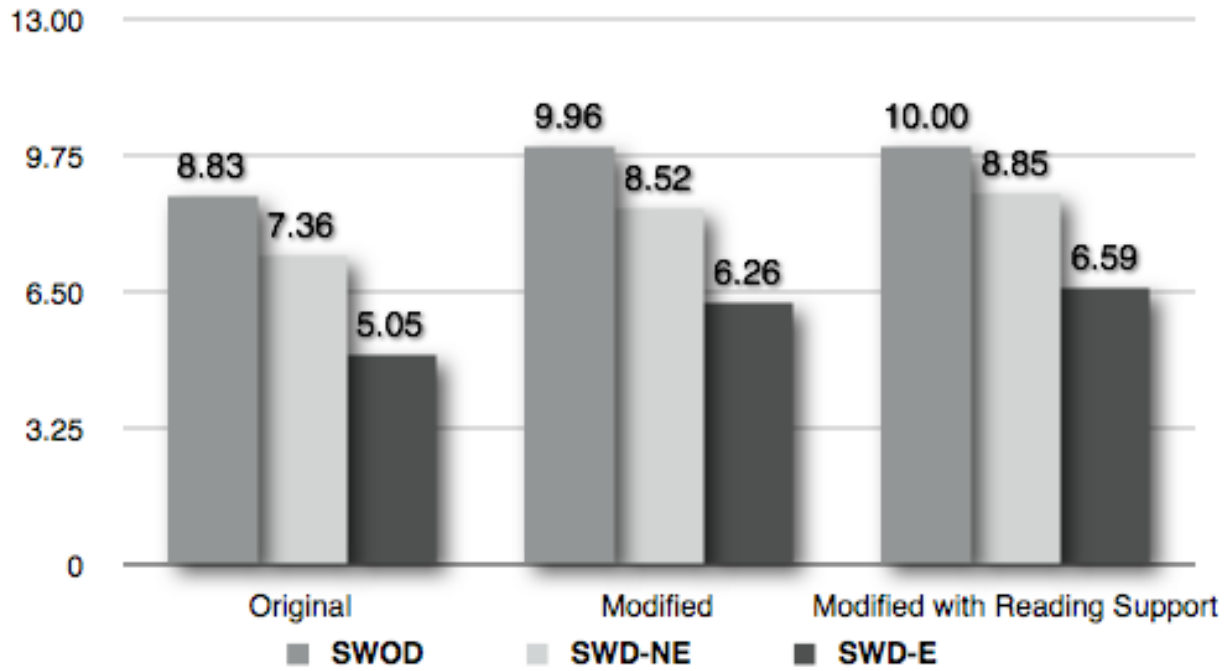


Figure 2. Mean test scores by group and condition in mathematics.

