

**Using Student Responses and Perceptions to Inform Item Development for an Alternate  
Assessment based on Modified Achievement Standards**

Andrew T. Roach

Georgia State University

Peter A. Beddow

Alexander Kurz

Ryan J. Kettler

Stephen N. Elliott

Peabody College of Vanderbilt University

Submitted: October 7, 2008

Revision Submitted: February 6, 2009

Please direct all correspondence to Andrew T. Roach, Department of Counseling and  
Psychological Services, Georgia State University, P.O. Box 3980, 30 Pryor St., Atlanta, GA  
30302-3980; or email: [aroach@gsu.edu](mailto:aroach@gsu.edu). Phone: 404-413-8176; Fax: 404-413-8013.

**Using Student Responses and Perceptions to Inform Item Development for an Alternate  
Assessment based on Modified Achievement Standards**

*Under Review - Do Not Quote Without Permission*

**Abstract**

Recent changes to NCLB permit the development of alternate assessments based on modified achievement standards (AA-MAS) for students with disabilities who are not expected to reach proficiency on general grade-level assessments. In developing their AA-MAS, several states have modified existing test items with the aim of enhancing accessibility and reducing difficulty for students with disabilities. Using a pool of grade 8 multiple-choice test items in original and modified forms, we conducted two studies to examine student perceptions of item modifications and their effects on accessibility. Study 1 used a think-aloud cognitive lab to explore the effects of modifications intended to enhance item accessibility on student perceptions and performance. Study 2 describes student survey data from a large-scale field test of the items with students ( $N = 698$ ) with and without disabilities. The authors conclude that the combination of think-aloud cognitive labs and post-test questionnaires provided important information about the validity and utility of item modifications and enhancements.

## **Using Student Responses and Perceptions to Inform Item Development for an Alternate Assessment based on Modified Achievement Standards**

In April 2007, the United States Department of Education revised regulations under the No Child Left Behind Act (NCLB) to create additional flexibility for states in facilitating the appropriate measurement of the achievement of certain students with disabilities. These revisions allowed states to develop alternate assessments based on modified achievement standards (AA-MAS). According to the USDOE *Non-regulatory Guidance* (2007), AA-MAS “are intended... for a limited group of students whose disability has prevented them from attaining grade-level proficiency” (p. 20). The USDOE has capped the number of students who may demonstrate proficiency via AA-MAS at 2% of a state’s or school district’s tested student population at a specific grade level (Bolt & Roach, 2008).

The AA-MAS are intended to measure the same grade-level content as states’ general large-scale assessments, but may include less difficult items or items that are modified or enhanced (e.g., visual cues, fewer answer choices, key terms bolded) to make these tests more accessible. These new tests will be referenced to modified achievement standards developed by each state. A modified achievement standard is “an expectation of performance that is challenging... but may be less difficult than a grade-level academic achievement standard. Modified academic achievement standards must be aligned with a State’s academic content standards for the grade in which a student is enrolled” (Bolt & Roach, 2008, p 14). It is important to note that these modified achievement standards are intended to be more challenging than states’ alternate achievement standards, which may feature content that is simplified in form and narrower in scope than the general grade-level standards. It is also important to understand that a modification to an item that all eligible students take is different than an accommodation for an

individual student. Both an accommodation to the testing procedures or response mode and a modification to an item are intended to enhance access for students, but accommodations are customized to an individual student's needs, whereas modifications are made to the actual "anatomy" of items. Modifications that enhance accessibility are not based on the individual needs of a particular student, but rather the class of students with disabilities and persistent academic difficulties. Thus, item modifications are more structural in nature and controlled by test developers. Conversely, accommodations are procedural in nature and controlled by IEP teams.

A goal of item development strategies used in creating an AA-MAS is to design assessment tasks that support, rather than inhibit, students' ability to show what they know and can do. Features included in AA-MAS items might facilitate the understanding of students with disabilities, or provide background information and support for understanding that does not undermine the construct being measured. In essence, test developers and policymakers expect that AA-MAS eligible students' experiences with, and cognitions while completing AA-MAS items will be different from what happens when the same students take original or unmodified items on the general large-scale assessments. Some information to support this assumption can be gathered from statistical analyses of test results (e.g., differential item functioning), but these methods can only provide quantitative evidence to support test item development. To understand the effects on test performance for items with modification intended to enhance accessibility, test developers are encouraged to use a variety of methods that tap students' cognitions, problem solving behaviors, and perceptions.

The central focus of this article is to report on two possible approaches to examining students' self-reported test-taking behaviors as well as their perceptions of test item features and

the relation of these features to the perceived difficulty of test items. Specifically, we provide an overview of support for the use of student response data in the recent version of the *Standards for Educational and Psychological Testing* (i.e., *Standards*; American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999), as well as a review of the use of student response data in the assessment literature. We will then present the results of two recent studies that used different research methodologies to gather and analyze students' responses and opinions regarding item modifications. We conclude by outlining implications and recommendations for policy and practice regarding the use of student response data in developing and validating states' AA-MAS procedures.

#### ***Uses of Student Response Data in the Test Standards***

The *Standards* (AERA, APA, & NCME, 1999) are intended to guide the development and validation of testing practices in education and psychology. The *Standards* are a relatively comprehensive overview of the rights and responsibilities of various stakeholder groups, including test developers and test users. The value of information regarding student responses and perceptions in supporting the development of assessments, including states' AA-MAS, is addressed at multiple points in the *Standards*.

Standard 10.3, in the chapter on testing individuals with disabilities, indicates "Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications" (p. 106). Because pilot testing often occurs with a smaller sample of participants, collecting information regarding student behaviors and cognitions during testing and their perceptions of assessment tasks may be more manageable than during actual

implementation. Gathering response data during pilot testing allows test developers to identify items with features students perceive as confusing. Identifying items and item features that may unintentionally influence and inhibit the performance of students with and/or without identified disabilities during pilot testing can reduce the unnecessary costs required to make changes to test forms and procedures during “live” testing.

The *Standards* suggest information about student response processes and test-taking behaviors can provide evidence to support the construct validity of an assessment. “Questioning test takers about their performance strategies can yield evidence that enriches the definition of a construct...” (AERA, APA, & NCME, 1999, p. 12). In the case of AA-MAS items, student response data can provide important information about the reasons for observed differences in performance across item types (original vs. enhanced/modified) and student groups (students with and without identified disabilities; AA-MAS eligible vs. non-eligible students). The use of concurrent think-aloud protocols and follow-up questioning may allow researchers to “unpack” unexpected results. For example, differential item functioning may indicate a particular item was difficult for students with identified disabilities in comparison to their peers. Recording students’ concurrent verbalizations while solving the item in question and questioning students following completion of the task may illuminate item features that contributed to the observed results. The *Standards* identify the latter as an important potential contribution of student response data: “Process studies involving examinees from different subgroups can assist in determining the extent to which capabilities irrelevant or ancillary to the construct may be differentially influencing (student) performance” (p. 12). This type of evidence is central in the development of an AA-MAS, where test items generally include features that are intended to reduce or eliminate construct-irrelevant influences on student outcomes.

Test developers also may collect data about student perceptions to provide consequential validity evidence. One desired outcome of the development and implementation of AA-MAS strategies is that the use of test item enhancements/modifications will result in tests that are more accessible and comprehensible, leading to improved student motivation and sense of efficacy. The *Standards* address this claim, indicating “Educational tests...may be advocated on the grounds that their use will improve student motivation....Where such claims are central to the rationale of testing, the direct examination of testing consequences necessarily assumes even greater importance” (AERA, APA, & NCME, 1999, p. 17). Follow-up questioning and surveys of student perceptions can provide important information about the influence of test item modifications on motivation and efficacy.

Similarly, item enhancements or modifications could be conceptualized as a form of educational intervention. In this case, student perceptions are essential evidence about the acceptability of these assessment strategies. Acceptability refers to an individual’s perceptions regarding the appropriateness, fairness, and reasonableness of an intervention (Kazdin, 1981). Evaluating the acceptability of proposed AA-MAS approaches requires surveys or interviews to understand the perceptions of students who qualify for these inclusive assessments strategies. Previous efforts to assess acceptability with children and adolescents generally have used focused on less complex, but more easily comprehensible concepts and terms (e.g., “like” and “fair”). Elliott (1986), however, suggested conceptual understanding of acceptability also requires some experience with the propose intervention strategy. Therefore, in post-testing interviews and surveys about test items with accessibility-enhancing modifications, students who received the modifications would be expected to identify them as making the items easier or less confusing than items without modifications.

### *Previous Reports of Using Students' Responses to Improve Testing Practices*

To date, the collection of student response data concerning assessments and testing practices in educational research has been largely qualitative and descriptive in nature (e.g., Brookhart & Bronowicz, 2003; Reay & Wiliam, 1999; Sambell, McDowell, & Brown 1997; Moni, Van Kraayenoord, & Baker, 2002) with the primary purpose of eliciting student perceptions on constructs such as self-efficacy, effort, and classroom knowledge. Educational psychologists have been especially interested in the relation of these constructs to motivation, a known academic enabler and contributor to student achievement (e.g., Brophy, 1999; DiPerna, Volpe, & Elliott, 2002; Hidi & Harackiewicz, 2000). Brookhart and Bronowicz used multiple case studies to investigate the relationships among three “constellations” of student perceptions relevant to classroom assessment: (a) responses related to the assessment task (e.g., interest, importance), (b) responses related to self-efficacy (e.g., ability to successfully complete the assessment task), and (c) responses related to personal goals (i.e., goal orientations). One of the primary findings was that students across different classroom assessments and settings consistently referenced their own needs and interests in their comments on task importance, ability to complete tasks, and overall work goals.

Some educational researchers have collected student response data for the purpose of evaluating the consequences of testing on learning style and learning behaviors (e.g., Gijbels & Dochy, 2006; Sambell, McDowell, & Brown, 1997). Such student response data can provide important evidence on the consequential validity of an assessment. Sambell and colleagues, for instance, situated student responses about fairness and changes in learning style in an overall validity argument. Student self-reports of adjusting learning behaviors before or after taking a test have provided further quantitative confirmation, indicating that differences in assessment

performance are correlated with differences in approaches to learning (Gijbels & Dochy, 2006).

The consequences of testing can thus have intended or unintended effects on classroom instruction and student behaviors (e.g., Wilson & Fowler, 2005; Dochy & McDowell, 1997).

The largest body of research concerning students with disabilities and inclusive assessment strategies is focused on the use of testing accommodations and their effects on test performance (see Sireci, Li, & Scarpati, 2003). Based on two literature reviews, the most frequently researched accommodations for students with disabilities were *oral administration* (i.e., read aloud) and *extended time* (Sireci, Scarpati, & Li, 2005; Thompson, Blount, & Thurlow, 2002). Thompson, Blount, and Thurlow found a positive effect on test scores for students with disabilities for oral administration and extended time in a majority of the reviewed studies. Sireci and colleagues examined much of the same literature, but focused on an interaction between student group and test administration condition, the so-called *interaction hypothesis* (i.e., accommodations lead to greater score improvements for students with disabilities in comparison to students without disabilities). Sireci et al. concluded that oral accommodations on mathematics tests were associated with increased test performance for some students with disabilities (i.e., mostly students with learning disabilities) and that extended time tended to increase the performance of all students, albeit to a greater extent for students with disabilities. The later interaction was found in a majority of the studies reviewed by Sireci et al. providing empirical support for the concept of *differential boost* (i.e., accommodations lead to greater score improvements for students with disabilities than for students without disabilities).

Previous research findings about the effects of accommodations on test performance for students with disabilities, however, have been inconsistent and difficult to interpret due to (a) variations in research designs, (b) differences in accommodation implementation procedures, and

(c) limitations of student samples (Ketterlin-Geller, Yovanof, & Tindal, 2007).

Recommendations for future research echoed across studies have included a suggested focus on (a) the interaction between item features and student characteristics, (b) the decision-making process for assigning accommodations, (c) studies beyond elementary school and on subjects other than mathematics and reading, (d) the application of universal design to assessments, and (e) investigations of student responses about the desirability and perceived usefulness of accommodations (Ketterlin-Geller et al., 2007; Sireci et al., 2005; Thompson et al., 2002).

Some researchers examining testing accommodations have used student response data to gain insights about students' perceptions of the fairness and utility of certain accommodations (e.g., Elliott & Marquardt, 2004; Fulk & Smith, 1995; Lang, Elliott, Bolt, & Kratochwill, 2008; McKevitt & Elliott, 2003; Nelson, Jayanthi, Epstein, & Bursuck, 2000; Kosciolik & Ysseldyke, 2000). The information provided through interviews, questionnaires, rating scales, and open-ended questions mostly indicated that students with and without disabilities perceived testing accommodations (for students with disabilities and students who might need them) as fair (e.g., Lang et al., 2008; Nelson et al., 2000; Polloway, Bursuck, Jayanthi, Epstein, & Nelson, 1996). In addition, perceptions about testing accommodations were sometimes found to be congruent with test performance (Kosciolik & Ysseldyke, 2000) and at other times to be unassociated with the effect of accommodations on student performance (Lang et al., 2008).

Despite the aforementioned research enterprises, the use of student response data for the purpose of test construction is virtually absent in the research literature. The perceptions of parents and teachers as users of assessments also have received little research attention (Roach, Berndt, & Elliott, 2007). This paucity of research belies the current understanding of validity and best practice recommendations for educational and psychological testing as advocated for by the

AERA, APA, and NCME (1999). The development and validation of AA-MAS strategies provides a context for affording users and consumers a greater voice in the test development process and the evaluation of test item modifications. In many cases, stakeholders (including students) can provide actionable information that leads to stronger validity evidence and greater test accessibility.

### **Using Student Response Data in the Development and Validation of an AA-MAS**

In July 2007, a team of educators and assessment specialists convened with the goal of generating a set of multiple-choice items for use in an experimental study as part of the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) project (for project details: <http://peabody.vanderbilt.edu/x8312.xml>). Using original reading and mathematics items provided by Discovery Education Assessment (DEA), content area groups used research on item development (Haladyna, Downing, and Rodriguez, 2002; Rodriguez, 2005) as well as an early version of the Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, and Elliott, 2008) to guide their efforts to modify the items with the aim of enhancing accessibility and improving measurement of intended constructs, particularly for students who would be eligible for an AA-MAS.

The TAMI is a research tool designed to facilitate a comprehensive analysis of tests and test items with the purpose of enhancing their accessibility, defined as “the extent to which a [test] permits equal access to all components and services for all individuals”(Beddow et al., 2008). The development of the instrument was guided by universal design principles (Center for Universal Design, 1997) and universal design for learning (UDL; Rose & Meyer, 2006; Center for Applied Special Technology, 2008), cognitive load theory (Chandler & Sweller, 1991; Mayer

& Moreno, 2003; Clark, Nguyen, & Sweller, 2006) and research on test and item development (Downing, Haladyna, & Rodriguez, 2002; Rodriguez, 2005).

The goal of the item modification process was to reduce the difficulty of items for students with special needs by reducing barriers to access and reducing extraneous cognitive load. Teams were instructed to make item modifications that did not change the target constructs or reduce the depth-of-knowledge (Webb, 1997) or grade level of the original items. Common item-specific modifications included simplification of language (in the item passage, stem, or response options), addition of graphic support, and using bold text for key terms in vocabulary and comprehension items. In addition to making specific modifications to individual items, the team applied a set of standard modifications across the entire item pool, including eliminating the least plausible distractor (i.e., reducing the number of answer choices from four to three) and increasing white space between response options. Ultimately, the team generated a set of 39 items for each content area that included modifications intended to enhance accessibility.

Two studies were undertaken to evaluate the effects of these item modifications: (a) a think-aloud cognitive lab in which students were asked to verbalize about their cognitions while completing a subset of the developed items; and (b) a large-scale field test of the items that included a post-test questionnaire regarding students' perceptions of the item modifications. Figure 1 illustrates how student response data were integrated into the CAAVES item development and validation process.

\*\*\* Insert Figure 1 about here \*\*\*

### ***Study #1. Think-Aloud Cognitive Lab***

Following the item development session, think-aloud cognitive labs were conducted to study the influence of the aforementioned test item modifications on problem-solving and test-

taking behaviors of students with and without identified disabilities. Students were asked to verbalize as they answered a series of assessment items, of which half were modified. With the high stakes placed on large-scale assessments like AA-MAS, there is a critical need for states to have valid information about how the design of assessments affects student performance (Johnstone, Bottsford-Miller, & Thompson, 2006). By requesting that students verbalize cognitive processes or “think aloud” while completing test items, the study allowed the research team to detect issues related to test design and accessibility.

In their seminal book on the topic, *Protocol Analysis: Verbal Reports as Data*, Ericsson and Simon (1993) describe the rationale for the development of their methods for obtaining concurrent and retrospective verbal reports. As Ericsson and Simon explain, due to difficulties with early attempts to use introspection in psychological research (e.g., James, 1890; Titchener, 1912) and the corresponding rise of behavioral psychology, many researchers came to view verbal reports as “useful for the discovery of psychological processes; (but) worthless for verification” (p. 2). However, drawing on more recent research on information processing, Ericsson and Simon were able to develop an approach to collecting concurrent and retrospective verbal reports that demonstrated minimal influence on subjects’ problem-solving and cognitions. Because of their desire to create “hard data” about individuals’ cognitive processes, Ericsson and Simon’s approach to gathering data is somewhat restrictive. For example, the experimenter provides limited prompting or encouragement and often seats him/herself behind the subject to discourage interaction. Moreover, “it is important that subjects verbalizing their thoughts while performing a task do *not* describe or explain what they are doing—they simply verbalize the information they attend to while generating the answer” (p. xiii).

***Test item selection and test construction.*** Young (2005) indicated assessment tasks or items chosen for use in think-aloud studies can dramatically impact the validity of the data generated by these studies. Ericsson and Simon (1993) suggested that think-aloud verbalization during assessment tasks reflects the cognitions simultaneous happening in participants' short-term memory (STM). Students may find it difficult to verbalize their problem solving on test items that are too simple and involve skills and concepts at a level of automaticity (i.e., stored in long-term memory). Conversely, items that are too difficult and complex may result in confusion and frustration on the part of student respondents.

Our study featured 16 items (8 Language Arts, 8 Mathematics) from the CAAVES item pool that had moderate (i.e., mid-range) difficulty based on student performance data from the existing DEA database. The selected items also were at Webb's depth-of-knowledge (DOK) level 2 or 3 (using Webb's original DOK descriptors). Two versions of each test were developed. Each version included eight original and eight modified items. Using two versions of the test allowed us to make comparisons between student behavior and responses on modified and original versions of each item.

### ***Method***

***Participants.*** According to Johnstone, Bottsford-Miller, and Thompson (2006), the sample size involved in cognitive interview research often is small (in comparison to other assessment research) because of the labor-intensive nature of the method. Because students spend several minutes (sometimes more than one hour) working their way through a series of items, a relatively small number of students results in extensive data sets (e.g., hours of audio tape, numerous pages of transcribed responses). Some researchers have suggested that a sample size as small as five participants per subgroup of interest can provide sufficient data for making

inferences. “The ‘Magic Number 5’—five participants will yield 80% of the findings from a usability test—comes from research conducted in the 1990’s by Nielsen, Virzi, Lewis, and other human factors engineers” (Barnum, 2003, ¶ 4). Unfortunately, it is unclear whether the methods implemented in usability research on software and other forms of technology can be generalized to assessment validation. In our study, participants ( $n = 9$ ) were chosen as representatives of particular groups of students deemed important to the project: general education students without identified disabilities (SWODs); students identified with disabilities who were not likely to be eligible for an AA-MAS (SWDs-NE) according to the participation criteria, which were developed for the project on the basis of USDOE’s *Non-regulatory Guidance* (2007); and students identified with disabilities who would be eligible for an AA-MAS (SWDs-E) according to the participation criteria.

Eligibility according to participation criteria was determined via school records and information provided by each student’s Individualized Education Plan (IEP) team. To be considered a SWD-E, the student’s IEP team had to determine evidence for all of the following criteria: (a) IEP goals were based on academic content standards for the grade in which the student was enrolled; (b) grade-level proficiency had not been achieved due to the student’s disability, as demonstrated by assessments that could validly document academic achievement; (c) academic progress in response to individualized instruction and assessed by multiple measures was judged to be insufficient to result in grade-level proficiency within the year covered by the IEP, even if significant growth was to occur. Students identified with disabilities who did not meet all three criteria were assigned to the SWD-NE group. Table 1 depicts the frequencies of participants by group and form.

**Procedure.** We individually administered one form of each test to each participant. The think-aloud sessions and follow-up questioning were videotaped and audiotaped for subsequent coding and analyses. A member of the research team explained the think-aloud procedures and modeled the process of verbalizing while answering test items. We used a script for explaining the process that was modified for the protocol used in a study by Johnstone, Bottsford-Miller, and Thompson (2006).

\*\*\* Insert Table 1 about here. \*\*\*

After explaining instructions and providing a short demonstration of how to “think aloud,” we asked students to engage in a series of sample items to practice verbalizing their thoughts. Students practiced the think aloud process until they were able to describe their problem solving behaviors and cognitions. At this point, students began completing the research items/tasks. Most students understood the directions and completed the sample items with little difficulty.

Following Johnstone, Bottsford-Miller and Thompson’s (2006) recommendation, we prompted students only when they were silent for approximately 10 consecutive seconds. If students verbalized infrequently while working on test items, we reminded them to “keep thinking aloud” or “keep talking.” Other than these prompts, we remained silent when students were thinking aloud to avoid disrupting or influencing their thought patterns (Ericsson & Simon, 1993). After completing the test items for each subtest, students were asked a series of follow-up questions to (a) clarify any inconsistencies or confusion regarding their think-aloud responses; and (b) gather additional information about their perceptions of the modified and original test items. Student responses to these follow-up questions also were videotaped for coding and analysis.

**Data analysis.** Videotapes of each student's think-aloud session and follow-up questioning were viewed and analyzed by at least two members of the research team. Student responses and behaviors during the think-aloud sessions were recorded by team members using a prepared coding sheet (see Figure 2). When the coders were not in agreement on the codes for a particular test item following individual viewing of the video, the corresponding section of the video was watched by the coders together to reach agreement on the appropriate code.

### **Results**

Descriptive data on each subgroup's performance on the reading items is provided in Table 2. Although the sample size does not lend itself to inferential statistics, a number of the results merit consideration. For example, although the SWOD and SWD-NE groups demonstrated relative consistency in the percentage of original and modified items answered correctly, the SWD-E group showed modest improvements in their performance on the reading items with modifications (i.e., a greater percentage of items were answered correctly in this condition). Although all three groups spent less time completing items with modifications and made fewer miscues in reading the modified passages, the differences on these measures were most dramatic for the SWD-E group (i.e., 34% less time and 23% fewer miscues in the modified condition compared to the original condition). Similarly, the number of research prompts (e.g., "keep talking," "tell me what you're thinking") per item was lower for the items with modifications; the reduction in number of prompts per item was greatest for the SWD-E group. Noticeable differences in oral reading fluency also were observed on items that included reading passages. The SWOD group read much more fluently (158.3 words correct per minute) than either of the groups with identified disabilities. Reading fluency generally was similar among the

two SWD groups: SWD-NE participants read 85.7 wcpm and SWD-E participant read 86.3 wcpm.

\*\*\* Insert Table 2 about here. \*\*\*

Descriptive data on each subgroup's performance on the mathematics items are provided in Table 3. Like the results from the reading cognitive lab items, the sample size does not lend itself to inferential statistics, but a number of the findings are intriguing. Modifications to the mathematics items appeared to contribute to improved performance (i.e., percentage of items answered correctly) for both groups of SWDs. A similar effect was not observed for general education students, who actually performed better on the original items. The SWOD and SWD-E groups spent less time completing items that had been modified, but no noticeable difference was observed in completion times for students in the SWD-NE group. The number of research prompts required was lower on modified items for each of the student groups with the largest difference observed with the SWD-E group (.58 prompts per original item versus .08 prompts per modified item). SWODs were more likely than SWDs to use appropriate strategies for solving mathematics problems.

\*\*\* Insert Table 3 about here. \*\*\*

*Use of visuals and other graphics as an item modification strategy.* In follow-up questioning, most SWDs (67%) saw the visuals as being helpful and providing support on reading questions and passages. Conversely, SWODs indicated the pictures made no difference in understanding the reading questions or passages. Most students (50% of SWDs; 67% of SWODs) saw the visuals and graphs as being helpful and providing support on mathematics items. A student in the SWD-E group commented "the (item) talking about the \$100 bills...well (the picture) showed me, and I was understanding how it goes with what it was talking about,

and I looked at it and it helped me even more.” However, 33% of SWDs indicated that the visuals or graphs were distracting or made it harder to answer the questions. As one student from the SWD-NE group said, “When people do math, they're working on a sheet and what's the point of looking at a picture. It doesn't really help you. For example, on (questions) #1 and #2, those two pictures were really messing me up.”

***Use of bolded vocabulary or key terms as an item modification strategy.*** Students also were asked about their perceptions of the use of bold font to highlight key terms or vocabulary in questions and reading passages. The majority of students from all groups (78% of the total) felt the use of bold type was helpful in answering the reading items. One student in the SWD-NE group indicated that, while this item modification helped draw students’ attention to key terms, it did not necessarily make the reading passages more accessible: “The bold type made (the answer) easier to find, but it didn’t help to understand the passage.”

***Reducing the number of answer choices (distractors) as an item modification strategy.*** Students were asked if they perceived items with three possible answers as less difficult than items with four answer choices. SWDs (with one exception) perceived no difference in difficulty between items having three or four possible answers on reading items. Conversely, 67% of SWODs identified the three-answer modification as making the reading items easier. As one student in the SWOD group indicated, on the modified items, “If you didn't get the answer right the first time, you (knew) you only had three choices to go back and look at three, instead of four.” This item modification strategy, however, generally did not affect either group’s performance on reading items; only one reading item demonstrated a discernable difference in student accuracy between modified and original versions. Students in the SWOD (67%) and SWD-NE (67%) groups generally indicated three answer choices made the mathematics items

easier. Some students in these groups appeared to use the possible answer choices to help solve mathematics items, but it was not clear that they used this same strategy in reading. For the students in SWD-E group, the three-answer choice modification was less likely to be identified as helpful, but it did seem to make a difference in performance on one particular item that dealt with scientific notation.

***Changing analogy formats as an item modification strategy.*** In an attempt to make vocabulary items on the reading test more accessible, we modified the format used in analogy questions. The CAAVES item development team anticipated that the original analogy format (e.g., “meteor: space:: dolphin: \_\_\_\_\_”) would be more difficult than a modified version (e.g., “meteor is to space as dolphin is to \_\_\_”). Most students (including 2/3 of students identified as having a disability) stated the traditional format for the analogy was easier. In follow-up questioning, some students indicated they had been taught analogies using this format and it was familiar to them. This was supported by the results, as SWDs correctly answered all the traditional analogy items but missed items with a modified analogy format 40% of the time.

The results from Study #1 (the think-aloud cognitive lab) were presented to the CAAVES leaders. Data on the effects of various modifications were then used to revise and finalize the modifications to the 39 items on the CAAVES reading and mathematics tests.

### ***Study #2: Post-Test Questionnaire***

The original and modified versions of the 39 item tests were field tested experimentally using DEA’s online test delivery system. The results from this field test and a follow-up survey of student participants were intended to answer three questions: (a) How did students perceive specific item modifications in reading and mathematics? (b) Did student perceptions about item

modifications differ across the three participant groups? (c) What was the correspondence between student perceptions about item modifications and actual test results?

### **Method**

**Participants.** A large sample of students in grade eight from the four states (Arizona, Hawaii, Idaho, and Indiana;  $N = 755$ ) participated in the CAAVES study. The sample was comprised of three groups. The first group consisted of SWOD ( $n = 269$ ), the second group of SWD-NE ( $n = 236$ ), and the third group of SWD-E ( $n = 250$ ). These groups were identified using the same participation criteria previously applied in the think-aloud cognitive lab (Study #1). Based on recent classroom and standardized test data, the latter group had the most persistent academic difficulties of the three groups. The majority of students who participated in the field test completed follow-up surveys after completing the reading and mathematics tests ( $N = 694$ ).

**Instruments.** For each content area, all participants received 13 items in each of three conditions: Original, Modified, and Modified with Reading Support. In the Modified with Reading Support condition, students were given help 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 icon. The ordering of the conditions was counterbalanced across the sample, and no student received both the original and modified versions of any particular item across the 39-item test forms. After the test, students were presented with a follow-up survey that contained seven questions about their perceptions of particular item modifications.

## **Results**

The authors of the current study examined response frequencies for each answer choice, focusing specifically on the two most frequently-selected answer choices as indicators of the most common views across the sample, to evaluate students' perceptions of various item modifications. It should be noted that the nature of the following analyses was exploratory rather than following from clearly specified *a priori* hypotheses. Therefore, contingency analyses (i.e., chi-square, Kramer's *V*) were excluded in favor of reporting the trends in the data observed in the current sample. Table 4 contains the results of the five questions that were common across the mathematics and reading surveys.

\*\*\* Insert Table 4 about here \*\*\*

***Perceived progression of difficulty.*** The first survey item required students to reflect on whether the test was equally difficult throughout or whether the test seemed easier toward the beginning, middle, or end. Across student groups, the majority of respondents reported the test had about the same difficulty all the way through (61% for reading; 46% for mathematics). Of the remaining students, most reported the test was easier toward the beginning (19% for reading; 29% for mathematics), despite the fact that some students received the Modified or Modified with Reading Support conditions first. For reading, compared to the overall sample, fewer students in the SWD-E group reported the test was the same difficulty throughout (49% versus 71% of SWODs). For mathematics, fewer SWDs reported the test was the same difficulty all the way through (42% and 41% of students in the SWD-NE and SWD-E groups, respectively, compared to 54% of SWODs). For all three student groups, the second most common response was that the items on both the reading and mathematics tests were perceived as easier toward the beginning. Actual field test results showed decreases in student performance for each successive

part across groups for both content areas, independent of the order of conditions (i.e., Original, Modified, or Modified with Reading Support).

**Adding visuals.** Students also were asked about the visuals that were included with some of the items. Across student groups, the majority of respondents reported the pictures helped them to understand the question. For the reading test, more SWD-Es (62%) reported the visuals provided helpful clues compared to SWD-NEs (50%), and SWODs (44%). On the mathematics test 58% of students in the SWD-E group reported visuals gave helpful clues, compared to 37% of students in the SWOD group and 44% of students in the SWD-NE group. The second most common response across groups and content areas was that visuals made no difference (56% of students in the SWOD group, 44% of students in the SWD-NE group, and 26% of students in the SWD-E group).

**Number of answer choices.** The majority of students reported items with three answer choices were easier than items with four choices (56% and 58% of all students for the reading and mathematics tests, respectively). For reading, fewer students in the SWD-E group reported items with three answer choices were easier when compared to students in the SWOD group. A similar pattern of survey responses was observed in regard to number of answer choices on the mathematics test items; fewer students in the SWD-E group identified this modification as helpful when compared to students in the SWOD group. Additionally, fewer students in the SWD-E group identified items with fewer answer choices as easier compared to students in the SWD-NE group.

**Bold font for key terms.** Across the three student groups, 80% of students reported the use of bold font helped them understand the key terms in reading passages. It should be noted that this result may be limited by the fact that the wording in the question (i.e. “Did [bold] make

[key terms] easier to find?") did not correspond properly with the answer choices (e.g., "Yes, the bold type helped me to understand the word in the passage"). We expected this modification to be most strongly endorsed by the SWD-E group, but fewer students in the eligible group reported bold type as helpful for vocabulary items (73%) compared to SWD-NEs (81%) and SWODs (84%). Actual data indicated that for the 17 items with key vocabulary terms in bold type, difficulty was lower for the Modified condition than for the Original condition. Several of these items had large decreases in difficulty from Original condition to Modified (e.g., one item had change in difficulty of .19) condition. These findings are confounded, however, by the fact that passages typically were shortened and simplified along with key terms being bolded.

**Reading support.** Across student groups, a majority of respondents reported reading support made the items easier (56% for both reading and mathematics). The magnitude of difference was largest between the SWOD and SWD-E groups for both reading and mathematics. Specifically, more students in the SWD-E group reported reading support made the items easier (67% on the reading test, 68% on the mathematics test) compared to students in the SWOD group (41% for reading; 40% for mathematics). Field test results in both content areas, however, indicated only small differences in student performance between the Modified condition and the Modified with Reading Support condition (effect sizes of .07 for reading and .05 for mathematics items).

**Relative difficulty of sample reading items.** The final two questions on the reading follow-up survey presented sample items in Original and Modified conditions to students and asked them to judge the items' relative difficulty. Modifications to the first sample reading item included eliminating one distractor, increasing space between answer choices, adding a visual, replacing underlined text for the vocabulary word with bold font, and adding a clarifying word

before the vocabulary word to provide additional context (see Figure 3). Upon examining both the modified and original versions, 52% of students reported the modified version was easier, whereas 44% indicated the two versions were about the same. We expected this package of item modifications to be strongly endorsed by students in the SWD-E group, but fewer SWD-Es (42%) reported the modified item was easier compared to students in both the SWOD (52%) and SWD-NE (53%) groups. Actual student performance data for this item, characterized in terms of the percentage who answered correctly, indicated that the item in the Modified condition was easier than the item in the Original condition ( $p = .53$  for the Original condition;  $p = .61$  for the Modified condition).

\*\*\* Insert Figure 3 about here \*\*\*

The second set of sample items on the reading follow-up survey included a stimulus which read “Do not desert me in my hour of need” accompanied by the stem, “What does **desert** mean when the stress is on the first syllable instead of the second?” Changes made to the original item included eliminating a distractor (i.e., incorrect answer choice), using bold font for the vocabulary term, and increasing space between lines. Across groups, 64% of students indicated the modified item was easier. Student performance data from the field test indicated the item was very difficult ( $p = .17$  for the Original condition;  $p = .21$  for the Modified condition) and suggested reading support, in tandem with the aforementioned modifications, actually increased the difficulty of the item ( $p = .15$ ). Although we would have expected the modifications to be preferred by SWD-Es, fewer students in the eligible group identified the modified item as easier (47%) compared to students in the SWOD group (64%).

*Relative difficulty of sample mathematics items.* The final three questions from the mathematics follow-up survey presented sample graphs or test items in Original and Modified

conditions. The first of these included two versions of a graph and asked respondents to select which was easier to understand. The original graph contained two data series and had a colored background. The modified version was larger, used a white background, and contained only one data series. Across student groups, 57% of respondents indicated the modified graph was easier to understand. Field test data for the item containing the original and modified versions of the graph indicated the item with the modified graph was easier ( $p = .44$  for the Original condition;  $p = .55$  for the Modified condition.) By contrast, another mathematics item that was included in the field test used the first (more complex) version of graph in both Original and Modified conditions, and only the standard set of modifications was used for the modified item (i.e., increasing white space and eliminating one distractor.) For this item, the difference in difficulty between conditions was negligible ( $p = .47$  for the Original condition;  $p = .45$  for the Modified condition).

\*\*\* Insert Figure 5 about here \*\*\*

The next survey item asked respondents to examine the original and modified versions of an item (see Figure 4) and report which version seemed easier. Changes from the Original to the Modified condition included eliminating one answer choice, adding a visual, using bold font for essential words, and simplifying item text. Across student groups, 64% of respondents indicated the modified item was easier. We expected that SWD-Es would be most likely to endorse the modified version as easier. Further analysis, however, indicated only 50% of students in the SWD-E group reported the modified item was easier compared to 70% of students in the SWD-NE group and 72% of students in the SWOD group. Approximately 50% of the students in the SWD-E group indicated the original item was easier or the items were about the same. Field test data for the item indicated the item was easier in the Original condition than in the Modified

condition for the SWOD group ( $p = .47$  and  $p = .37$ , respectively), whereas the item in the Modified condition was easier for both groups of students with disabilities. The difference in difficulty between conditions was greatest for the SWD-E group ( $p = .21$  for the Original condition,  $p = .35$  for the Modified condition).

The final mathematics follow-up survey item presented an item in Original and Modified conditions and asked students to indicate which was easier. Changes to the original item included eliminating one answer choice, increasing white space, and using bold font for the word *not* in the item stem. Across groups, 53% of students indicated the modified item was easier, and 35% indicated the items were about the same. Again, we expected students in the SWD-E group to be most likely to identify the modifications as helpful. Fewer SWD-Es (47%), however, reported the modified item was easier compared to students in the SWOD group (55%). Field test data for the item indicated the Modified condition was comparatively easier for all groups. The effects of the modifications were most apparent for the SWOD group, whose mean performance reflected the greatest magnitude of change in difficulty across conditions ( $p = .55$  for the Original condition;  $p = .66$  for the Modified condition). Performance on the item improved with the addition of reading support for the SWD-E group ( $p = .37$  for the Modified condition;  $p = .56$  for the Modified with Reading Support condition), but data indicated reading support made little difference for the other two student groups.

### Discussion

The studies presented in this article provide examples of ways in which students' perceptions can be integrated into the development of AA-MAS and other types of achievement tests. The combination of think-aloud cognitive labs and post-test questionnaires during the

CAAVES field test provided important information about the validity and utility of item modifications and enhancements.

### ***Think-Aloud Cognitive Labs***

The results from the think-aloud cognitive labs indicated reading fluency may be a barrier for students identified with a disability (regardless of eligibility for an AA-MAS). In some cases, these students' slower rates of reading resulted in testing sessions that were almost twice as long as the sessions experienced by their general education peers. These results raise a number of questions for test developers and policymakers. For example, should reading passages on an AA-MAS be briefer in an effort to reduce reading load? Could technology be used to address this barrier? Both of these inclusive assessment strategies were integrated into the CAAVES field test forms: passages were shortened as much as was feasible without altering the concepts and grade level difficulty; and, in one condition, directions and item stems were read aloud via a computerized voiceover.

During the cognitive lab sessions, students in all three groups spent less time and required fewer prompts on the items that included modifications. This difference may be explained, in part, by the items in the Modified condition being shorter than the items in the Original condition. However, the difference was most pronounced for the SWD-E group, which would provide some support for a differential boost effect from the modification strategies. Conversely, oral reading fluency did not appear to be influenced by the modifications made to reading passages, suggesting that shortened passages might be a necessary step towards improved accessibility.

Students with and without identified disabilities expressed support for some item modification strategies, including adding visuals to support comprehension of reading passages,

and using bold type to highlight key terms or vocabulary. SWD-E, however, were less likely than their peers to endorse a reduction in distracters (i.e., three answer choices rather than four) as making items easier to answer.

“Conservative” modifications were used in this study and the effects on student performance generally were modest. More “aggressive” modifications (e.g., reading passages aloud, simplifying or pre-teaching content) might result in more robust effects. Specifically, students in the SWD-E group often appeared unfamiliar with concepts (e.g., percentages, scientific notation), and often incorrectly applied problem solving strategies on mathematics items. In these cases, item modifications strategies like shortened passages, additional visuals, or more white space on the page are unlikely to provide support or facilitate access.

### ***Post-Test Questioning of Students***

Results from the follow-up surveys administered to students in the field test provided additional information about students’ perceptions of particular item modifications in reading and mathematics. Moreover, this study allowed us to examine the correspondence between student perceptions of item modifications and actual test results.

Although all students received at least some combination of modified and original items, a majority of students indicated they perceived test items as equally difficult across the test. Across student groups, data indicated performance decreased significantly from the beginning of the test to the end, regardless of the order of presentation for the different conditions (Original, Modified, and Modified with Reading Support). Students in the SWD-E and SWD-NE groups were more likely to indicate the test was easiest at the beginning than students in the SWOD group, however, suggesting that students with disabilities may be more susceptible to the effects of test fatigue than their peers.

During the field test, more students in the SWD-E group reported visuals were helpful, suggesting that the addition of visuals may improve the accessibility of some items for students with identified disabilities who would be eligible for an AA-MAS. Actual data for items for which visuals were added, however, indicate that this item modification strategy should be implemented with caution (Kettler, et al. 2008). Preliminary analyses suggest student performance on items improved only when the added visuals contained information essential for understanding and responding to the item. This finding suggests test developers should consider refraining from adding visuals when the primary goal is making items more interesting or to increase student motivation.

Rodriguez's (2005) meta-analysis of over 80 years of item-writing research indicated that reducing the number of answer choices for a multiple-choice item from four to three does not harm the psychometric properties of an item, but does theoretically reduce the reading load of the entire test. Thus, eliminating distractors should be expected to be most beneficial to students who tend to perform poorly in testing situations. Overall student survey responses indicated a majority of students in the field test perceived items with three answer choices as easier than items with four answer choices, but fewer students in the SWD-E group identified this modification as beneficial. Similarly, a large majority of students indicated the use of bold font was helpful for finding the key terms in the reading passages, but fewer students in the SWD-E group perceived this modification as helpful, compared to the other groups. These findings suggest that students with persistent academic difficulties may not be aware of features intended to enhance the accessibility or reduce the difficulty of items. By contrast, students in the SWD-E group were the most likely to report reading support as helpful. Results from the field test, however, indicated reading support had a negligible effect on student performance on modified

items. This suggests reading support may provide a level of comfort for students who tend to perform poorly on tests, thereby reducing anxiety but only minimally affecting their performance on items.

In all cases, when students were presented with an item in Original and Modified conditions and asked to report which item seemed easier, the majority of respondents reported the modified item was easier. This suggests students across a broad range of abilities and needs are able to perceive when modifications have been made to items to enhance their accessibility.

### ***Limitations and Areas for Future Research***

The two studies presented in this article were part of a series of studies designed to advance the development of a set of 8th-grade achievement tests with items modified to increase accessibility and reduce item difficulty in response to the need to develop AA-MAS strategies for use students with persistent academic difficulties. Think-aloud cognitive labs typically involve small samples of students, but given the diversity of students identified with disabilities who have persistent academic difficulties, it would have been desirable to have included a broader sample of students with academic disabilities. The cognitive lab study also would have been enhanced had we collected a direct measure of reading fluency or been able to use eye tracking measures to document a key nonverbal behavior of students. Additional think-aloud studies also could be conducted to understand elementary or high school students' perceptions of test item modifications.

The post-test questionnaire was taken by a large and representative sample; however, unlike the cognitive lab study where a researcher was present and able to prompt and probe students' responses to follow-up questions, the questionnaire required students to respond without support or direction. The motivation for students to complete this post-test set of

questions is unknown, and it is possible that some students completed the survey items hurriedly or randomly in order to finish quickly.

Across the two studies reported in this article, student perceptions of the acceptability and utility of item modifications did not consistently align with the effects of these modifications on performance. Although this was an unanticipated outcome, these results point to the need for future investigations of the effects of item modifications, especially for students with disabilities. For example, data from field tests and post-test questionnaires indicate that students' performance (regardless of disability status) might be affected by fatigue during testing sessions. Future research might investigate whether modifications, including shortened reading passages or reduced test length, can address students' fatigue and improve their performance. Similarly, additional research is warranted to determine if item modifications that are viewed by students as helpful, improve their engagement and sense of efficacy during the testing sessions.

### ***Implications for Policy and Practice***

Over the past decade since the passage of the Individuals with Disabilities Education Act of 1997 (IDEA), we have completed a number of validity studies involving inclusive assessments -- accommodated testing and alternate assessments -- for students with disabilities. As we have advanced our understanding of inclusive assessment strategies, and concurrently embarked on new validity research with items designed for AA-MAS instruments, we have found it essential to involve students with disabilities more directly and actively in our work. Given limited research on item modifications, advances in cognitive load theory and mental load analyses (Clark et al., 2006), and ongoing concerns about improved accountability for students with disabilities, it seems logical and appropriate to invite students to be more involved in the development of more accessible tests. The involvement of students is not required by policy, but

we believe that it is an essential component of a comprehensive item development and refinement process, and that it will lead to more accessible items and tests.

When students' perspectives regarding assessments and modifications are not considered, educators, policymakers, and test developers may work from a "paternalistic" assumption (i.e., "acting upon (our) own idea of what's best for another person without consulting that other person" (Marchewaka, cited in Smart, 2001, p. 200). Although there are some cases in which ascertaining student perspectives and preferences would be difficult (e.g., students with significant cognitive disabilities with no reliable mode of communication), we believe most students with and without disabilities are fully capable of expressing their opinions regarding the accessibility and acceptability of testing practices.

The accessibility-enhancing item modifications we have examined are currently part of research efforts to improve the quality of assessments for students with disabilities who have experienced persistent academic difficulties. Not all such eligible students are likely to realized improved test scores from item modifications; however, it appears test design practices that affect many students with and without disabilities can be improved. Given that more than 25 states have reported they are planning to develop an AA-MAS (Altman, Lazarus, Thurlow, Quenemoen, Cuthbert, & Cormier, 2008), it is important to give the students most likely to be affected by such assessments a voice in the development process. We encourage educators and test developers to include these data as part of their test development and validation efforts.

## References

- Altman, J.R., Lazarus, S.L., Thurlow, M.L., Quenemoen, R.F., Cuthbert, M., & Cormier, D.C. (2008). *2007 survey of states: Activities, changes, and challenges for special education*. Minneapolis, MN: University of Minnesota, National Center on Education Outcomes.
- American Educational Research Association, American Psychological Association, and the National Center on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.
- Barnum, C. (2003). What's in a number? *Usability Interface*, 9 (1). Available from: <http://www.stcsig.org/usability/newsletter/0301-number.html>
- Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test Accessibility and Modification Inventory (TAMI)*. Nashville, TN: Vanderbilt University.
- Bolt, S. E., & Roach, A. T. (2008). *Including Diverse Learners in Standards-Based Accountability: Promoting Access to Assessment and Instruction*. New York: Guilford Press.
- Brookhart, S. M., & Bronowicz, D. L. (2003). 'I don't like writing. It makes my fingers hurt': Students talk about their classroom assessments. *Assessment in Education*, 10(2), 221-242.
- Brophy, J. (1999). Toward a model of the value aspects of motivation in education: developing appreciation for particular learning domains and activities. *Educational Psychologist*, 34(2), 75-88.
- Center for Universal Design (1997). *What is universal design?* Center for Universal Design, North Carolina State University. Retrieved from the World Wide Web: [www.design.ncsu.edu](http://www.design.ncsu.edu).

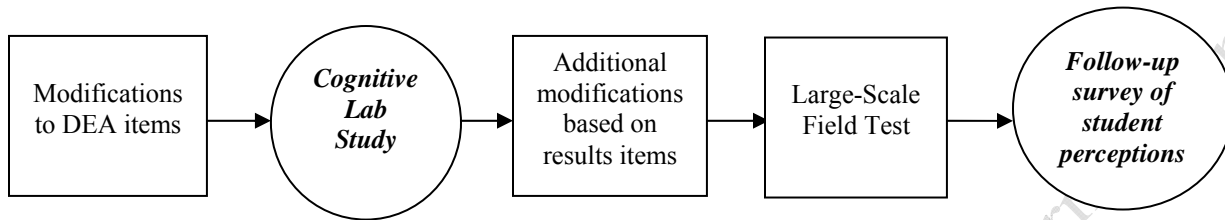
- Clark, R., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. San Francisco, CA: Pfeiffer.
- DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2001). A model of academic enablers and elementary reading/language arts achievement. *School Psychology Review, 31*, 298–312.
- Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*(4), 279-298.
- Elliott, S. N., & Marquardt, A. M. (2004). Extended time as testing accommodations: Its effects and perceived consequences. *Exceptional Children, 70*, 349-367.
- Elliott, S. N. (1986). Children's ratings of the acceptability of classroom interventions for misbehavior: Findings and methodological considerations. *Journal of School Psychology, 24*, 23–35.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.
- Fulk, C. L., & Smith, P. J. (1995). Students' perceptions of teachers' instructional and management adaptations for students with learning or behavior problems. *The Elementary School Journal, 95*, 409-419.
- Gijbels, D., & Dochy, F. (2006). Students' assessment preferences and approaches to learning: can formative assessment make a difference? *Educational Studies, 32*(4), 399-409.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

- Hidi, S. & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issues for the 21st century. *Review of Educational Research*, 70, 151-179.
- James, W. (1890). *The principles of psychology* (Vols. 1-2). New York: Dover.
- Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Kazdin, A. E. (1981). Acceptability of child treatment techniques: The influence of treatment efficacy and adverse side effects. *Behavior Therapy*, 12, 493–506.
- Ketterlin-Geller, L., Yovanoff, P., & Tindal, G. (2007). Developing a new paradigm for conducting research on accommodations in mathematics testing. *Exceptional Children*, 73, 331-347.
- Kettler, R. J., Rodriguez, M. R., Bolt, D. M., Elliot, S. N., Beddow, P. A., & Kurz, A. (2008). *Modified multiple-choice items for alternate assessments: Reliability, difficulty, and the interaction paradigm*. Manuscript submitted for publication.
- Kosciolek, S. & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Technical Report 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Lang, S. C., Elliott, S. N., Bolt, D. M., & Kratochwill, T. R. (2008). The effects of testing accommodations on students' performances and reactions to testing. *School Psychology Quarterly*, 23(1), 107-124.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38, 43-52.

- McKevitt, B. C., & Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review, 32*(4), 583-600.
- Moni, K. B., Van Kraayenoord, C. E., & Baker, C. D. (2002). Students' perceptions of literacy assessment. *Assessment in Education, 9*(3), 2002.
- Nelson, J. S., Jayanthi, M., Epstein, M. H., & Bursuck, W. D. (2000). Student preferences for adaptations in classroom testing. *Remedial and Special Education, 21*, 41-52.
- Polloway, E. A., Bursuck, W. D., Jayanthi, M., Epstein, M. H., & Nelson, J. S. (1996). Treatment acceptability: Determining appropriate interventions within inclusive classrooms. *Interventions in School & Clinic, 31*, 133-144.
- Reay, D., & Wiliam, D. (1999). 'I'll be a nothing': Structure, agency and the construction of identity through assessment. *British Educational Research Journal, 25*(3), 343-354.
- Roach, A. T., Elliott, S. N., & Berndt, S. (2007). Teacher perceptions and the consequential validity of an alternate assessment for students with significant cognitive disabilities. *Journal of Disability Policy Studies, 18*(3), 168-175.
- 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.
- Sambell, K., McDowell, L., & Brown, S. (1997). 'But is it fair?': An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation, 23*(4), 349-371.
- Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Research Report 485). Amherst, MA: Center for Educational Assessment.

- Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75, 457-490.
- Smart, J. (2001). *Disability, Society, and the Individual*. Gaithersburg, MD: Aspen.
- Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Tech. Rep. No. 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Policy14.htm>
- Titchener, E. B. (1912). The schema of introspection. *The American Journal of Psychology*, 23(4), 485-508.
- U.S. Department of Education. (April, 2007). *Modified academic achievement standards: Non-regulatory guidance*. Washington, D.C.: Author.
- Wilson, K. & Fowler, J. (2005). Assessing the impact of learning environments on students' approaches to learning: Comparing conventional and action learning designs. *Assessment and Evaluation in Higher Education*, 30(1), 87-101.
- Young, K. 2005. Direct from the Source: The value of 'think-aloud' data in understanding learning. *Journal of Educational Enquiry*, 6(1), 19-33.

Figure 1. *Use of student response data in CAAVES item development.*



*\*\*Note: Student response elements are in bold font in circles.*


Under Review - Do Not Quote Without Permission

Figure 2. Example of coding sheet for think-aloud cognitive lab data.

<b>Think Aloud Data Coding Sheet</b>				
Student ID #	SWOD	SWD-NE	SWD-E	Coder initials
<p><b>Describe Researcher Introduction</b> (if included on video)</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Researcher explained think aloud procedure</li> <li><input type="checkbox"/> Researcher modeled thinking aloud on practice item</li> <li><input type="checkbox"/> Researcher asked the student to re-state/paraphrase their understanding of the study (“<b>What were you told we were going to do today?</b>”)</li> </ul> <p>Verbatim recording of student response/paraphrase of directions:</p>  <p>Record any additional questions or comments student made about think-aloud directions, purpose, or practice items:</p>				
<p><b>Reading #</b> _____ <input type="checkbox"/> <b>Modified</b> <input type="checkbox"/> <b>Original</b></p> <p>Time started: _____ Time completed: _____</p> <p><i>If this item included a reading passage, record the amount of time student spent orally reading the item:</i></p> <p>_____</p> <p><i>Number of words read correct in passage:</i> _____ <i>Number of miscues:</i> _____</p> <p><i>Reading fluency:</i> _____ <i>wpm</i></p> <p>Number of researcher prompts: _____</p> <p>Verbatim recording of prompts:</p>  <p>Other student comments (verbatim recording):</p>				
<p><b>Mathematics Item #</b> _____ <input type="checkbox"/> <b>Modified</b> <input type="checkbox"/> <b>Original</b></p> <p>Time started: _____ Time completed: _____</p> <p>Number of researcher prompts: _____</p> <p>Verbatim recording of prompts:</p>  <p>Student approach the question/problem solving strategies:</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Correct/appropriate process for answering question or solving problem</li> <li><input type="checkbox"/> Incorrect approach or problem solving process <input type="checkbox"/> Appeared to guess</li> <li><input type="checkbox"/> Not Apparent <input type="checkbox"/> Did not attempt</li> </ul> <p>Other student comments (verbatim recording):</p>				

Figure 3. Reading survey question, item comparison.

Look at the following two example problems. Which of these is easier for you to do?

 Play

47. **Problem #1**

The young driver was given a permit after passing the written test.


In this sentence permit means \_\_\_\_\_.

A. to consent to  
 B. an oral agreement  
 C. an official document  
 D. a ticket

**Problem #2**

The young driver was given a learner's **permit** after passing the written test.

In this sentence **permit** means \_\_\_\_\_.



A. to consent to  
 B. an oral agreement  
 C. an official document

A. Problem #1  
 B. Problem #2  
 C. About the same.

Under Review

mission

Figure 4. Sample mathematics item in original and modified forms.

ORIGINAL	MODIFIED
<p>35. Mr. Jameson is a salesman who lives in Knoxville, Tennessee. He travels the distance from his home to Jackson and back twice a month. He also travels to Chattanooga and back home twice a month. If the distance from Knoxville to Jackson is 295 km and the distance from Knoxville to Chattanooga is 95 km, how many km does he drive on these trips each month?</p> <p> <input type="radio"/> A. 670 km  <input type="radio"/> B. 780 km  <input type="radio"/> C. 1,560 km  <input type="radio"/> D. 14,103 km                 </p>	<p>35. Mr. James travels from his home to Jackson and back twice a month. He also travels to Greenwood and back home twice a month.</p> <p>The distance from home to Jackson is 295 km <b>each way</b>. The distance from home to Greenwood is 95 km <b>each way</b>.</p> <p>How far does he drive on these trips in a month?</p> <p> <input type="radio"/> A. 670 km  <input type="radio"/> B. 780 km  <input type="radio"/> C. 1,560 km                 </p>

Under Review - Do Not Quote

Table 1.

*Frequencies of Participants by Group and Form*

	Form A	Form B	Total
SWOD participants	2	1	3
SWD-NE participants	1	2	3
SWD-E participants	1	2	3

Under Review - Do Not Quote Without Permission

Table 2.

*Student Performance on Think Aloud Reading Items*

Group		Percent of Items Correct	Time Spent per Item (mean)	Miscues on Passages (mean)	Fluency on Passages (mean)	Researcher Prompts per Item (mean)
SWOD	Original Items	83%	79.6 s	2.7	153.3 wcpm	.49
	Modified Items	83%	51.0 s	1.5	163.3 wcpm	.29
SWD-NE	Original Items	83%	123.8 s	9.8	92.6 wcpm	.65
	Modified Items	75%	100.5 s	9.0	78.7 wcpm	.28
SWD-E	Original Items	67%	149.4 s	12.3	86.9 wcpm	.81
	Modified Items	75%	98.5 s	9.5	85.8 wcpm	.28

Table 3.

*Student Performance on Think Aloud Mathematics Items*

Group		Percent of Items Correct	Time Spent per Item (mean)	Researcher Prompts per Item (mean)	Problem Solving Strategies	
					Correct Strategy Used	Incorrect Strategy Used
SWOD	Original Items	67%	65.8 s	.33	67% (8)	25% (3)
	Modified Items	50%	54.1 s	.08	50% (6)	33% (4)
SWD-NE	Original Items	50%	125.2 s	.33	42% (5)	50% (6)
	Modified Items	75%	126.2 s	.08	42% (5)	50% (6)
SWD-E	Original Items	33%	102.5 s	.58	25% (3)	58% (7)
	Modified Items	50%	72.8 s	.08	8% (1)	58% (7)

Table 4.

*Follow-up Survey Results and Group Comparisons*

Question	Most Common Answer	Content Area	Total ( <i>N</i> = 694)	SWOD ( <i>n</i> = 245)	SWD-NE ( <i>n</i> = 220)	SWD-E ( <i>n</i> = 229)
1. Thinking about the different questions on this test, were the questions:	D. about the same all the way through	Reading	61%	71%	62%	49%
		Mathematics	46%	54%	42%	41%
2. Some of the questions included pictures. Did the pictures help you understand the question?	A. Yes, the pictures gave me clues that helped me understand the question	Reading	52%	44%	50%	62%
		Mathematics	46%	37%	44%	58%
3. Some of the questions had three possible answers and some had four possible answers. Which format did you think was easier?	A. The questions with 3 possible answers seemed easier.	Reading	56%	58%	61%	49%
		Mathematics	58%	59%	68%	49%
4. On some of the items, parts were read aloud to you. Did hearing the items read aloud make the items easier?	A. Yes, having parts of the items read aloud made them easier.	Reading	56%	41%	62%	67%
		Mathematics	56%	40%	61%	68%
5. On some of the reading passages, the key vocabulary terms were written in <b>bold</b> type. Did that make them easier to find?	A. Yes, the bold type helped me understand the word in the passage.	Reading	80%	84%	81%	73%