Exceptional Children

Vol. 77, No. 1, pp. 61-80.

Incorporating Student Input in Developing Alternate Assessments Based on Modified Academic Achievement Standards

ANDREW T. ROACH
Georgia State University

PETER A. BEDDOW

ALEXANDER KURZ

RYAN J. KETTLER

STEPHEN N. ELLIOTT
Peabody College of Vanderbilt University

ABSTRACT: *In developing alternate assessments based on modified academic achievement standards (AA-MAS), several states have modified existing test items with the aim of enhancing accessibility and reducing difficulty for students with disabilities. Using Grade 8 multiple-choice test items in unmodified and modified forms, two studies examined student perceptions of item modifications and their effects on accessibility. Study 1 used a think-aloud cognitive lab to explore the effects of modifications on student perceptions and performance. Study 2 examined student survey data from a field test of the items with students (N = 709) with and without disabilities. Data indicated students generally perceived item modifications as helpful or positive. Educators and test developers are encouraged to consider students' perceptions of test items when designing tests.*

In April 2007, the United States Department of Education (ED) revised regulations under the No Child Left Behind Act of 2001 (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 academic achievement standards (AA-MAS). According to the ED's *Modified Academic Achievement Standards: 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 ED 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).

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 include modifications (e.g., visual cues, fewer answer choices, key terms bolded) intended to make these tests more accessible. These new tests will be referenced to modified achievement standards developed by each state. A *modified academic 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 modified academic achievement standards are intended to be more challenging than states' alternate academic 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 group 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 individualized education program (IEP) teams.

In its *Modified Academic Achievement Standards: Non-Regulatory Guidance* document (2007), the ED indicated that states "may modify an existing assessment or develop a new assessment" for use as an AA-MAS (p. 24). The document provides examples of AA-MAS development strategies, including "modifying the same items that appear on the grade-level assessment by simplifying the language of the item or eliminating a 'distractor' in multiple-choice items" (p. 25). There has been considerable concern and debate over the use of the terms *modify* and/or *modification* in describing the changes made to items for use on an AA-MAS. Over the past 2 decades, research and policy documents describing testing accommodations typically have described modifications as changes to testing procedures that undermine the construct measured and lead to less valid inferences about student performance. However, the most recent version of *Standards for Educational and Psychological Testing* (*Standards for Testing;* American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) is less restrictive in its definition of modifications, describing them as "changes made in the content, format, and/or administration procedures of a test in order to accommodate test takers who are unable to take the unmodified test under standard test conditions" (p. 183). The term *modification* appears nearly a dozen times in the index of the *Standards for Testing*, with significant coverage in chapters on test administration, scoring, and reporting; testing individuals of diverse linguistic backgrounds; and testing individuals with disabilities. The *Standards for Testing*'s chapter that addresses assessment of individuals with disabilities indicates "No connotation that modification implies a change in the construct(s) being measured is intended" (p. 101). The *Standards for Testing*'s definition of modification reflects the item procedures undertaken in developing the assessment items used in our studies.

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 should facilitate the understanding of students with disabilities, or provide background information and support in ways that do not "compromise the validity and reliability of the test results" (ED, 2007, p. 26). In essence, test developers and policy makers 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 existing 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 devel-

opment. To understand the effects on test performance for items with modification intended to enhance accessibility, test developers must use a variety of methods that tap students' cognitions, problem-solving behaviors, and opinions.

USES OF STUDENT RESPONSE DATA IN THE STANDARDS FOR TESTING

The *Standards for Testing* (AERA, APA, & NCME, 1999) is intended to guide the development and validation of testing practices in education and psychology and also includes 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 for Testing.*

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" (AERA, APA, & NCME, 1999, 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 student 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 "retrofit" test forms and procedures during "live" testing.

The *Standards for Testing* suggests that 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 (unmodified vs. modified) and student groups (e.g., students with and without identified disabilities; AA-MAS eligible vs. noneligible students). The use of concurrent think-aloud protocols and follow-up questioning may allow researchers to "unpack" unexpected results. 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 for Testing* identifies 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" (AERA, APA, & NCME, 1999, 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.

Gathering student response data during pilot testing allows test developers to identify items with features students perceive as confusing.

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 modifications will result in tests that are more accessible and comprehensible, leading to improved student motivation and sense of efficacy. The *Standards for Testing* addresses 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 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 assessment strategies. Previous efforts to assess acceptability with children and adolescents generally have focused on less complex but more easily comprehensible concepts and terms (e.g., *like* and *fair*). Elliott (1986) suggested conceptual understanding of acceptability also requires some experience with the proposed intervention strategy. Therefore, in posttesting interviews and surveys about test items with accessibility-enhancing modifications, students who received the modifications might be expected to identify them as making the items seem 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; Moni, Van Kraayenoord, & Baker, 2002; Reay & Wiliam, 1999; Sambell, McDowell, & Brown 1997), with the primary purpose of eliciting student perceptions on constructs such as self-efficacy, effort, and classroom knowledge. 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.

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). 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, Yovanoff, & 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, Scarpati, & Li, 2005; Thompson, Blount, & Thurlow, 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; Kosciolek & Ysseldyke, 2000; Lang, Elliott, Bolt, & Kratochwill, 2008; McKevitt & Elliott, 2003; Nelson, Jayanthi, Epstein, & Bursuck, 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 (Kosciolek & 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. 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

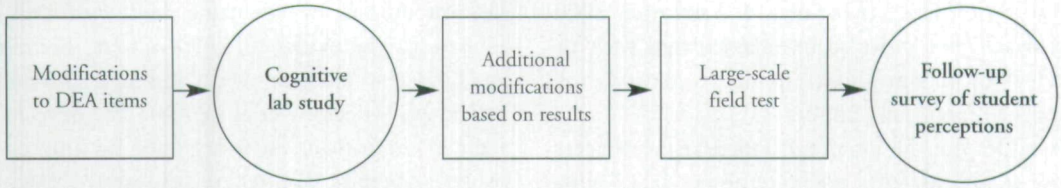
In July 2007, a team of educators and assessment specialists convened with the goal of generating a set of multiple-choice items for use in an experimental study as part of the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) project (for project details, see http://peabody.vanderbilt.edu/Faculty_and_Research/Peabody_Research_Office/About_Peabody_Research/Funded_Projects/CAAVES_Project_Home.xml). Using original reading and mathematics items provided by Discovery Education Assessment (DEA), content area groups used research on item development (Haladyna, Downing, & Rodriguez, 2002; Rodriguez, 2005) as well as an early version of the Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, & 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 eligible for an AA-MAS.

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 (Center for Applied Special Technology, 2008; Rose & Meyer, 2006), cognitive load theory (Chandler & Sweller, 1991; Clark, Nguyen, & Sweller, 2006; Mayer & Moreno, 2003), and research on test and item development (Haladyna et al., 2002; Rodriguez, 2005).

The goal of the item modification process was to reduce the difficulty of items for eighth-grade 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 unmodified items. The team generated a set of 39 items for each content area that included modifications intended to enhance accessibility and comprehensibility. 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. The decision to delete the least plausible distractor was based on Rodriguez's (2005) meta-analytic work on the optimal number of answer choices for multiple-choice items. Rodriguez's analyses indicated moving from four answer choices to three choices resulted in .04 reduction in item difficulty, a .03 increase in item discrimination, and a .02 increase in reliability. Rodriguez also compared the effects of randomly deleting answer choices versus deleting the least plausible choices and found "when eliminating random distractors to create 3-option items, reliability drops .06 on

FIGURE 1

Use of Student Response Data in CAAVES Item Development



Note. Student response elements are in bold font in circles. CAAVES = Consortium for Alternate Assessment Validity and Experimental Studies project; DEA = Discovery Education Assessment.

average, with no change if ineffective distractors are deleted" (p. 10).

Two studies evaluated 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 posttest questionnaire regarding students' perceptions of the item modifications. Figure 1 illustrates the integration of student response data into the CAAVES item development and validation process.

STUDY 1. THINK-ALOUD COGNITIVE LAB

Following the item development session, the think-aloud cognitive labs allowed us to study the influence of the aforementioned test item modifications on problem-solving and test-taking behaviors of students with and without identified disabilities. By requesting that students verbalize cognitive processes or "think aloud" while completing test items (of which half were modified), we were able to detect issues related to test design and accessibility.

The think-aloud cognitive labs allowed us to study the influence of the aforementioned test item modifications on problem-solving and test-taking behaviors of students with and without identified disabilities.

In their seminal book on the topic, *Protocol Analysis: Verbal Reports as Data,* Ericsson and Simon (1993) described the rationale for the development of their methods for obtaining concurrent and retrospective verbal reports. As Ericsson and Simon explained, 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 developed 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 sits 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. Students may find it difficult to verbalize their problem solving on test items that are too

TABLE 1

Test-Item Modification Distribution

	Reading								*Mathematics*							
Item #	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Test A	U	U	U	U	M	M	M	M	U	U	U	U	M	M	M	M
Test B	M	M	M	M	U	U	U	U	M	M	M	M	U	U	U	U

Note. U = unmodified version of item; M = modified version of item.

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 Reading, 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 (1997) 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 unmodified and eight modified items (Table 1). Using two versions of the test allowed us to make comparisons between student behavior and responses on modified and unmodified 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 an 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- or videotape, 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 1990s 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, participating eighth-grade students (n = 9) were representative of particular groups of students deemed important to the project: general education students without identified disabilities (SWOD); students identified with disabilities who were not likely to be eligible for an AA-MAS (SWD-NE) according to the participation criteria, which were developed for the project on the basis of ED's *Modified Academic Achievement Standards: Non-Regulatory Guidance* (2007); and students identified with disabilities who would be eligible for an AA-MAS (SWD-E) according to the participation criteria.

Eligibility according to participation criteria was determined via school records and information provided by each student's IEP team. To be considered for inclusion in the SWD-E group, 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; and (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 2 depicts the frequencies of participants by group and form.

Procedure. We individually administered one form of each content area test to each participant.

TABLE 2

Frequencies of Participants by Group and Test Form

Participants	*Test Form A*	*Test Form B*	*Total*
SWOD	2	1	3
SWD-NE	1	2	3
SWD-E	1	2	3

Note. SWOD = Students without disabilities; SWD-NE = Students with disabilities who likely would not be eligible for an AA-MAS; SWD-E = Students with disabilities who likely would be eligible for an AA-MAS. AA-MAS = alternate assessment based on modified academic achievement standards.

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 we modified from a protocol developed and used by Johnstone and colleagues (2006) in previous cognitive labs conducted with students with disabilities.

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 et al.'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 the students completed the test items for each subtest, we asked them 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 unmodified 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. Team members recorded student responses and behaviors during the think-aloud sessions on a prepared coding sheet. Initial interrater agreement for the coding was strong, ranging from 94.7% agreement on time spent (i.e., number of seconds) on each item to 97.2% for the number of researcher prompts. In cases where the coders were not in agreement on the codes for a particular test item following independent viewing of the video, the coders watched the corresponding section of the video together to reach consensus.

RESULTS

Table 3 provides descriptive data on each subgroup's performance on the reading items. 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 unmodified 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 unmodified 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. There are also noticeable differences in oral reading fluency on items that included reading passages. The SWOD group read much more fluently (158.3 words correct per minute; wcpm) than either of the groups with identified disabilities. Reading fluency generally

TABLE 3

Student Performance on Think-Aloud Reading Items

Participants		*Percentage of Items Correct*	*Time Spent per Item M*	*Miscues on Passages M*	*Fluency on Passages M*	*Researcher Prompts per Item M*
SWOD (n = 3)	Unmodified 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 (n = 3)	Unmodified 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 (n = 3)	Unmodified items	67	149.4 s	12.3	86.9 wcpm	.81
	Modified items	75	98.5 s	9.5	85.8 wcpm	.28

Note. SWOD = Students without disabilities; SWD-NE = Students with disabilities who likely would not be eligible for an AA-MAS; SWD-E = Students with disabilities who likely would be eligible for an AA-MAS; AA-MAS = alternate assessment based on modified academic achievement standards; wcpm = words correct per minute.

was similar among the two SWD groups: SWD-NE participants read 85.7 wcpm and SWD-E participants read 86.3 wcpm.

Table 4 provides descriptive data on each subgroup's performance on the mathematics items. 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 unmodified 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 unmodified item vs .08 prompts per modified item). SWODs were more likely than SWDs to use appropriate strategies for solving mathematics problems.

Use of Visuals and Other Graphics as an Item Modification Strategy. In follow-up questioning, most SWDs (67%) thought the visuals were helpful and provided 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 Bold-Font Vocabulary or Key Terms as an Item Modification Strategy. We also asked students 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, although 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."

TABLE 4

Student Performance on Think-Aloud Mathematics Items

Participants		*Percentage of Items Correct*	*Time Spent per Item M*	*Researcher Prompts per Item M*	*Problem-Solving Strategies*	
					Correct Strategy Used %	*Incorrect Strategy Used%*
SWOD (n = 3)	Unmodified items	67	65.8 s	.33	67 (8)	25 (3)
	Modified items	50	54.1 s	.08	50 (6)	33 (4)
SWD-NE (n = 3)	Unmodified items	50	125.2 s	.33	42 (5)	50 (6)
	Modified items	75	126.2 s	.08	42 (5)	50 (6)
SWD-E (n =3)	Unmodified items	33	102.5 s	.58	25 (3)	58 (7)
	Modified items	50	72.8 s	.08	8 (1)	58 (7)

Note. SWOD = Students without disabilities; SWD-NE = Students with disabilities who likely would not be eligible for an AA-MAS; SWD-E = Students with disabilities who likely would be eligible for an AA-MAS; AA-MAS = alternate assessment based on modified academic achievement standards.

Reducing the Number of Answer Choices (Distractors) as an Item Modification Strategy. 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 seem 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 unmodified versions. Students in the SWOD (67%) and SWD-NE (67%) groups generally indicated three-answer choices made the mathematics items seem 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 perceived as more difficult than a modified version (e.g., "meteor is to space as dolphin is to _____"). Most students (including two thirds of students identified as having a disability) stated the traditional format for the analogy was easier for them to understand. 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: POSTTEST QUESTIONNAIRE

The unmodified and modified versions of the 39 items were field-tested experimentally using DEA's online delivery system. We used the results of a follow-up survey of student participants to examine two questions, considered both for the total sample and for the subsample of students who would be eligible for an AA-MAS: (a) When considering all students who reported item modifications influenced item accessibility or perceived difficulty, did a majority of this group report that specific item modifications were positive or helpful versus not helpful? and (b) Did a majority of the total participant sample report that specific item modifications were positive/helpful compared to those who reported specific item modifications were not helpful or had no effect?

METHOD

Participants. A large sample of students in Grade 8 from four states (i.e., Arizona, Hawaii, Idaho, and Indiana; n = 694 in reading; n = 709 in mathematics) responded to seven survey questions following completion of each content area test. The sample comprised three groups: SWOD (n = 246 in reading; 255 in mathematics), SWD-NE (n = 220 in reading; 219 in mathematics), and SWD-E (n = 228 in reading; 235 in mathematics). These groups were identified using the same participation criteria applied in the think-aloud cognitive lab (Study 1). Based on recent classroom and standardized test data, the SWD-E group had the most persistent academic difficulties of the three groups.

Instruments. For each content area, all participants received 13 items in each of three conditions: unmodified, modified, and modified with reading support. In the modified with reading support condition, students received help from a voice recording 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 unmodified and modified versions of any particular item across the 39-item test forms. Actual field test results indicated the effect size between modified (without reading support) and unmodified conditions was .38 for reading and .21 for mathematics (Elliott, et al., 2010). Effect sizes for reading and mathematics between unmodified and modified with reading support conditions were .46 and .25, respectively. Effect sizes between modified and modified with reading support conditions were .07 for reading and .05 for mathematics. Considered within a Rasch model that equated group ability levels, the largest effect sizes were observed for the SWD-E group (see Kettler et al., 2009.) After each content area test, students completed a follow-up survey of seven questions about their perceptions of particular item modifications.

Data Analyses. We used pairwise tests of proportion to examine whether students selected the expected survey response(s) more frequently than the comparison response(s). Our expectation was that the majority of students would endorse modifications and modified items as being helpful (or "easier") instead of unhelpful and/or making no difference. Thus, for each test, the analyses examined whether the proportion of interest (i.e., modifications/modified items were helpful) was greater than 0.5.

The first item in both content areas required students to indicate whether they perceived the test as easier toward the beginning, in the middle, at the end, or whether the test seemed equally difficult all the way through. For this item, we collapsed the modified and modified with reading support conditions as the responses of interest. Thus, inferential analyses consisted of 4 one-tailed tests of proportion to examine (a) the proportion of students who indicated either the modified or modified with reading support condition seemed easier versus the proportion who indicated the unmodified condition seemed easier (i.e., the sample was restricted to only those students who selected one of the three conditions, eliminating from the analysis students who indicated the test was the same throughout); and (b) the proportion of students who indicated either the modified or modified with reading support condition seemed easier versus the proportion of students that indicated they perceived the unmodified condition as easier or the test was the same all the way through. Each test of proportion was conducted for the total sample and repeated for the SWD-E group. Thus, we selected a critical

p value of .0125 to correct for the number of comparisons ($\alpha = .05 / 4 = .0125$).

The remaining six questions in both content areas required students to consider specific item modifications or item pairs and indicate whether the modification(s) were positive, negative, or made no difference. Inferential analyses for each of these items consisted of 2 one-tailed tests of proportion to examine the two research questions: (a) the proportion of students who reported the modification(s) were positive versus the proportion who reported they were negative, and (b) the proportion of students who reported the modifications were positive versus those who reported they were negative or made no difference. Each pairwise comparison was conducted for the total sample and repeated for the SWD-E group. We selected a critical p value of .0125 to correct for the number of comparisons ($\alpha = .05 / 4 = .0125$). As with the first survey question, these analyses examined whether the proportion of interest (i.e., modifications/modified items were helpful) was greater than 0.5.

Results

Across survey questions, participants were more likely to report that modifications were perceived as positive/helpful versus negative/not helpful, but in all cases the greatest number of respondents reported no perceived difference between the two conditions. For analyses related to the first research question, the sample was restricted to respondents who reported either the modifications were positive/helpful or who reported modifications were negative/not helpful (i.e., students who reported the modifications made no difference were removed from the analyses.) The analyses related to the second question included the total respondent sample. Table 5 contains the response frequencies and percentages for each of the questions on the mathematics and reading surveys across groups and for students in the eligible (SWD-E) group only.

Perceived Difficulty of Each Condition. The first survey item asked whether the test seemed easier toward the beginning, easier toward the middle, easier toward the end, or the same all the way through. It should be noted that students were not made explicitly aware that the test was divided into three conditions, and the conditions were counterbalanced across the sample to control for order effects. Thus, aggregated student responses to this survey item are independent of the order of conditions. For both reading and mathematics, the proportion of the total student sample who reported the modified condition or the modified with reading support condition seemed easier was significantly larger than the proportion of students who reported the unmodified condition seemed easier ($p < .001$ for both content areas). Likewise, the proportion of students in the SWD-E group who reported the modified condition or the modified with reading support condition seemed easier was significantly larger than the proportion who reported the unmodified condition seemed easier ($p < .001$ for both content areas). The proportion of students who reported either the modified or modified with reading support condition seemed easier was not significantly larger than the proportions of students who either reported the unmodified condition seemed easier or that the test was the same all the way through ($p = 1.00$ for both content areas). Similarly, the proportion of students in the SWD-E group who reported either the modified or modified with reading support condition seemed easier was not significantly larger than the proportion of students in the SWD-E group who reported the unmodified condition seemed easier or that the test was the same all the way through (both $p = 1.00$).

Adding Visuals. When asked about the visuals that were included with some of the items, a significantly larger proportion of the total sample reported the pictures helped them understand the question compared to the proportion who reported visuals made the question more difficult to understand ($p < .001$ for both content areas). For students in the SWD-E group, a significantly larger proportion reported visuals were helpful ($p < .001$ for both content areas) compared to those who reported visuals were not helpful. For the full sample, the proportion of students who reported visuals were helpful was not significantly larger than the proportion of students who reported they were either not helpful or made no difference ($p = 1.00$ for both content areas). For the SWD-E group, however, the proportion of students who reported visuals were helpful was significantly larger than the proportion of stu-

TABLE 5

Frequencies and Percentages of Student Responses on Field Test Follow-Up Survey

		Reading Survey n *(%)*				*Mathematics Survey* n *(%)*			
Survey Question	*Sample*	*Unmodified*	*Modified*	*Reading Support*	*Same*	*Unmodified*	*Modified*	*Reading Support*	*Same*
Easier condition	Total	66 (10)	79 (11)	126 (18)	423 (61)	98 (14)	122 (17)	163 (23)	326 (46)
	SWD-E	36 (16)	32 (14)	47 (21)	113 (50)	42 (18)	36 (15)	61 (26)	96 (41)
		Positive/Easier		*Same*	*Negative/Harder*	*Positive/Easier*		*Same*	*Negative/Harder*
Visuals	Total	359 (52)		60 (9)	275 (40)	328 (46)		299 (42)	81 (11)
	SWD-E	142 (62)		33 (14)	53 (23)	136 (58)		62 (27)	36 (15)
Reading support	Total	390 (56)		220 (32)	84 (12)	395 (56)		212 (30)	100 (14)
	SWD-E	155 (68)		52 (23)	21 (9)	162 (69)		48 (21)	24 (10)
Bold font	Total	551 (80)		107 (15)	35 (5)	—		—	—
	SWD-E	165 (73)		38 (17)	24 (11)	—		—	—
		Three Choices		*Same*	*Four Choices*	*Three Choices*		*Same*	*Four Choices*
Answer choices	Total	392 (56)		235 (34)	67 (10)	413 (58)		216 (30)	80 (11)
	SWD-E	112 (49)		78 (34)	37 (16)	114 (49)		79 (34)	42 (18)
		Modified		*Same*	*Unmodified*	*Modified*		*Same*	*Unmodified*
Sample graph	Total	—		—	—	383 (54)		199 (28)	126 (18)
	SWD-E	—		—	—	117 (50)		73 (31)	45 (19)
Sample Item 1	Total	342 (49)		291 (42)	59 (9)	454 (64)		142 (20)	112 (16)
	SWD-E	94 (42)		99 (44)	33 (15)	118 (50)		60 (26)	57 (24)
Sample Item 2	Full	393 (57)		205 (30)	93 (13)	375 (53)		250 (35)	82 (12)
	SWD-E	102 (45)		77 (34)	47 (21)	111 (47)		81 (35)	42 (18)

Note. SWD-E = Students with disabilities who likely would be eligible for an AA-MAS; AA-MAS = alternate assessment based on modified academic standards.

dents who reported visuals were not helpful or made no difference ($p < .01$ for both content areas).

Number of Answer Choices. Across content areas, a significantly larger proportion of the total sample reported items with three answer choices seemed easier than the proportion who reported items with four choices seemed easier ($p < .001$ for both content areas). Similar results were observed when considering the responses of SWD-E group only ($p < .001$). Further, a significantly larger proportion of the total sample reported items with three answers seemed easier compared to the combined proportion of students who either reported items with four answers seemed easier or that the number of answer choices made no difference ($p < .001$ for both content areas). Conversely, these results were not observed for the SWD-E group for either content area ($p = .58$ for reading; $p = .68$ for mathematics).

Bold Font for Key Terms. The reading survey included a question regarding the use of bold font to help test-takers identify key words in passages. A significantly larger proportion of the total sample reported items that used bold font made the key words easier to find than the proportion who reported the bold font was distracting ($p < .001$). Similar results were observed when considering the responses of SWD-E group only ($p < .001$). Further, a significantly larger proportion of students reported items that used bold font made the key terms easier to find compared to the combined proportion of students who either reported bold font was distracting or that it made no difference ($p < .001$). These results also were observed for the responses of SWD-E group alone ($p < .001$).

Reading Support. For the total sample, a significantly larger proportion of respondents in both content areas reported reading support made the items seem easier compared to the proportion who reported the reading support made the items seem harder ($p < .001$ for both content areas). Similar results were observed when considering the responses of SWD-E group only ($p < .001$ for both content areas). Further, the proportion of the total sample who reported the reading support made the items seem easier was larger than the combined proportion of students who reported the reading support either made the items seem harder or made no difference ($p < .001$ for both content areas). This result also was observed when considering SWD-E group's responses ($p < .001$ for both content areas).

Relative Difficulty of Sample Reading Items. The final two questions on the reading follow-up survey presented sample items in unmodified 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 2). Upon examining both the modified and unmodified versions, a significantly larger proportion of the total sample reported the modified version of the item seemed easier compared to the proportion who reported the unmodified version seemed easier ($p < .001$). Similar results were observed when considering the responses of SWD-E group only ($p < .001$). The proportions of students who reported the modified version of the item seemed easier were not significantly larger than the combined proportions of students who reported the unmodified version seemed easier or that the items were about the same.

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 to the unmodified item included eliminating a distractor, using bold font for the vocabulary term, and increasing space between lines. A significantly larger proportion of the total sample reported the modified item seemed easier compared to the proportion reporting the unmodified item seemed easier ($p < .001$). This result also was observed when considering the responses of SWD-E group only ($p < .001$). Further, a significantly larger proportion of the total sample reported the modified item seemed easier compared to the combined proportion of students who either reported the unmodified item seemed easier or the items were about the same. This finding was not observed for the SWD-E group.

Relative Difficulty of Sample Mathematics Items. The final three questions from the mathe-

FIGURE 2

Sample Reading Assessment Item in Unmodified and Modified Forms

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 **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.

matics follow-up survey presented sample graphs or test items in unmodified and modified forms. The first of these included two versions of a graph and asked respondents to select which was easier to understand. The unmodified 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.

For the total sample, as well as for the SWD-E group only, significantly larger proportions of students reported the modified graph was easier to understand compared to the proportions reporting the unmodified version was easier to understand ($p < .001$ across groups; $p < .001$ for the SWD-E group). The proportion of students from the total sample who reported the modified graph was easier to understand was not significantly larger than the combined proportion of those who either reported the unmodified graph was easier to understand or that they were about the same ($p = .015$). The result also was nonsignificant for the SWD-E group's responses ($p = .53$).

The next mathematics survey item asked respondents to examine the unmodified and modified versions of an item (see Figure 3) and report which version seemed easier. Changes from the unmodified to the modified form included eliminating one answer choice, adding a visual, using bold font for essential words, and simplifying item text.

For the total sample, the proportion of students who reported the modified version of the item seemed easier was significantly larger than the proportion reporting the unmodified version of the item seemed easier ($p < .001$). This result also observed when considering the responses of SWD-E group alone ($p < .001$). Likewise, the proportion of students in the total sample who

FIGURE 3

Sample Mathematics Item in Unmodified and Modified Forms

ORIGINAL

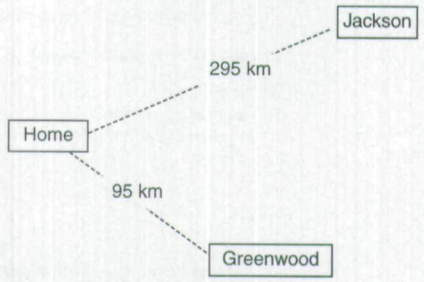
35. Mr. Jameson is a salesman who lives in Knoxville, Tennessee. He travels the distance from his home to Jackson and back twice a month. He also travels to Chattanooga and back home twice a month. If the distance from Knoxville to Jackson is 295 km and the distance from Knoxville to Chattanooga is 95 km, how many km does he drive on these trips each month?

- ○ A. 670 km
- ○ B. 780 km
- ○ C. 1,560 km
- ○ D. 14,103 km

MODIFIED

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.

The distance from his home to Jackson is 295 km **each way**. The distance from home to Greenwood is 95 km **each way**.



How far does he drive on these trips in a month?

- ○ A. 670 km
- ○ B. 780 km
- ○ C. 1,560 km

reported the modified item seemed easier was significantly larger than the combined proportion of students who either reported the unmodified version seemed easier or reported the items were about the same ($p < .001$). This result was not observed for the SWD-E group ($p = .47$).

The final mathematics follow-up survey item presented an item in both unmodified and modified forms and students were asked to indicate which was easier. Changes to the unmodified item included eliminating one answer choice, increasing white space, and using bold font for the word *not* in the item stem. For both the total sample and for the SWD-E group only, the proportion of students who reported the modified version seemed easier was significantly larger than the proportions who reported the unmodified item seemed easier ($p < .001$ for the total sample; $p < .001$ for the SWD-E group). The proportion of the total sample and of the SWD-E group who reported the modified version seemed easier was not significantly larger than the combined proportions of students who either reported the unmodified version seemed easier or the items were about the same ($p = .05$ for the total sample; $p = .78$ for the SWD-E group).

DISCUSSION

The studies presented 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 posttest questionnaires during the CAAVES field test provided important information about the validity and utility of item modifications.

Think-Aloud Cognitive Labs

The results from the think-aloud cognitive labs indicated reading fluency may be a barrier for students with disabilities (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 policy makers. 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 unmodified condition. However, the difference was most pronounced for the SWD-E group, providing 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 reading 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.

This study's conservative modifications resulted in generally modest effects on student performance. More "aggressive" modifications (e.g., reading passages aloud, simplifying or preteaching 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 modification strategies such as shortened passages, additional visuals, or more white space on the page are unlikely to provide needed supports or facilitate access.

Posttest Surveys

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. Across survey questions, students perceived modifications as positive and/or reported they made the test items easier. This was confirmed by actual field test data; namely, a moderate effect size was observed from unmodified to modified items (Elliott et al., 2010).

Although all students participated in all three conditions and received a combination of modified and unmodified items, a majority of students indicated they perceived the test as equally difficult all the way through. Because the conditions were counterbalanced across the sample and no cues were provided to indicate a shift in condition, it is likely many students were unaware there were three separate conditions, and thus did not identify whether either of the modified conditions was easier than the unmodified condition. Still, a significantly greater proportion of respondents across content areas reported one of the two modified conditions was easier than the unmodified condition.

In all cases, when asked about specific modification strategies (i.e., the use of visuals, reading support, and bold font for key words), a significantly greater proportion of students for the total sample and for the SWD-E group reported that modifications were positive or helpful compared to those who reported the modifications were negative or not helpful. This finding was consistent across survey questions. It should be noted, however, that analyses conducted on the unrestricted sample (i.e., including respondents who reported the modifications made no difference) generated typically nonsignificant results.

Moreover, student support of individual modifications sometimes contradicted actual performance data from the field test. Although students overwhelmingly reported the reading support was helpful (particularly AA-MAS-eligible students), the smallest effect sizes were observed between the modified and modified with reading support conditions (Elliott et al., 2010). Additionally, students endorsed the modification of adding visuals as being helpful. Field test data on the use of visuals, however, suggest this item modification strategy should be implemented with caution when working with reading items (Kettler et al. 2009). These findings highlight the importance of distinguishing student perceptions from actual performance data as two separate sources of information.

A significantly greater proportion of students in both reading and mathematics reported items with three answer choices were easier than items

with four answer choices seemed easier. 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 preserves (and in some cases improves) the psychometric properties of an item, while theoretically reducing the reading load of the entire test. Thus, eliminating distractors should be expected to be most beneficial to students who tend to perform poorly and take a longer amount of time in testing situations.

In all cases, when students were presented with an item or visual in unmodified and modified forms and asked to report which seemed easier, a significantly greater proportion of respondents reported the modified form 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

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.

A large and representative sample of students completed the posttest questionnaire; 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 posttest set of questions is unknown, and it is possible that some students completed the survey items with haste or limited understanding. Also, the sample was necessarily restricted in many comparisons as students who made neutral selections were excluded. Future posttest survey research might benefit from a forced-choice method (i.e., excluding the option for students to report the test as being the same in various conditions).

Implications for Policy and Practice

Since the passage of the Individuals With Disabilities Education Act of 1997, 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 actively involved in the development and evaluation 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, policy makers, and test developers may work from a paternalistic assumption—"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 examined are 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. More than 25 states have reported they are planning to develop an AA-MAS (Altman et al., 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, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Barnum, C. (2003). What's in a number? *Usability Interface, 9*(1). Retrieved 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, NY: Guilford.

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,* 221–242.

Center for Universal Design. (1997). *What Is Universal Design?* Raleigh, NC: North Carolina State University. Retrieved from http://www.design.ncsu.edu

Center for Applied Special Technology (2009). *Universal Design Guidelines - Version 1.0.* Wakefield: MA: CAST. Retrieved from http://www.udlcenter.org/aboutudl/udlguidelines.

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

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

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.

Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., . . . & Roach, A. T. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children, 76,* 475–495.

Elliott, S. N., & Marquardt, A. M. (2004). Extended time as testing accommodations: Its effects and perceived consequences. *Exceptional Children, 70,* 349– 367.

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.

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,* 309–334.

James, W. (1890). *The principles of psychology* (Vols. 1–2). New York, NY: 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. (2009). *Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost.* 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,* 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,* 583–600.

Moni, K. B., Van Kraayenoord, C. E., & Baker, C. D. (2002). Students' perceptions of literacy assessment. *Assessment in Education, 9,* 319–342.

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,* 343–354.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24,* 3–13.

Rose, D. H., & Meyer, A. (Eds.). (2006). *A practical reader in universal design for learning.* Cambridge, MA: Harvard Education Press.

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,* 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* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Titchener, E. B. (1912). The schema of introspection. *The American Journal of Psychology, 23,* 485–508.

U.S. Department of Education. (2007, July). *Modified academic achievement standards: Non-regulatory guidance.* Washington, DC: Author. Retrieved from http://www.cehd.umn.edu/nceo/2percentReg/NonregulatoryGuidance.pdf

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison, WI: University of Wisconsin–Madison, National Institute for Science Education.

Young, K. (2005). Direct from the source: The value of 'think-aloud' data in understanding learning. *Journal of Educational Enquiry, 6,* 19–33.

ABOUT THE AUTHORS

ANDREW T. ROACH (Georgia CEC), Assistant Professor, Department of Counseling and Psychological Services, Georgia State University, Atlanta. **PETER N. BEDDOW** (Tennessee CEC), Research Assistant in Special Education; **ALEXANDER KURZ** (Tennessee CEC), Dunn Family Scholar in Educational and Psychological Assessment; **RYAN J. KETTLER** (Tennessee CEC), Research Assistant Professor in Special Education; and **STEPHEN N. ELLIOTT** (Tennessee CEC), Professor of Special Education, Peabody College of Vanderbilt University, Nashville, Tennessee.

Correspondence concerning this article should be addressed to Andrew T. Roach, Department of Counseling and Psychological Services, Georgia State University, P.O. Box 3980, 30 Pryor St., Atlanta, GA 30302-3980 (e-mail: aroach@gsu.edu).

The current study was implemented as part of the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) project, funded by the U.S. Department of Education (awarded to the Idaho Department of Education; #S368A060012). The positions and opinions expressed in this article are those solely of the author team.

Manuscript received October 2008; accepted October 2009.