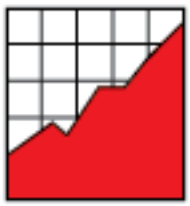
Technical Report 48

Student Think Aloud Reflections on Comprehensible and Readable Assessment Items: Perspectives on What Does and Does Not Make an Item Readable



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

Supported by:

U.S. Office of Special Education Programs

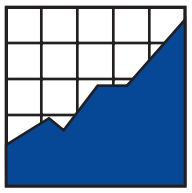
Technical Report 48

Student Think Aloud Reflections on Comprehensible and Readable Assessment Items: Perspectives on What Does and Does Not Make an Item Readable

Christopher Johnstone • Kristi Liu • Jason Altman • Martha Thurlow

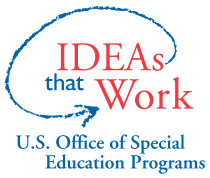
September 2007

Johnstone, C., Liu, K., Altman, J., & Thurlow, M. (2007). *Student think aloud reflections on comprehensible and readable assessment items: Perspectives on what does and does not make an item readable* (Technical Report 48). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Center is supported through a Cooperative Agreement (#H326G050007) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. This report was supported by a grant (#H324D020050) from the U.S. Department of Education, Office of Special Education Programs, Directed Research Division. The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.



NCEO Core Staff

Deb A. Albus
Jason Altman
Manuel T. Barrera
Laurene Christensen
Christopher J. Johnstone
Jane L. Krentz
Sheryl S. Lazarus
Kristi K. Liu
Ross E. Moen
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Martha L. Thurlow, Director

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/626-1530 • Fax 612/624-0879
http://www.nceo.info

Executive Summary

This document reports on research related to large-scale assessments for students with learning disabilities in the area of reading. As part of a process of making assessments more universally designed we examined the role of "readable and comprehensible" test items (Thompson, Johnstone, & Thurlow, 2002). In this research, we used think aloud methods to better understand how interventions to improve readability affected student performance. Decreasing word counts in items and making important words bold did not seem to make any difference in student achievement (although students preferred that important words were printed in bold). Vocabulary, however, was a notable factor. Non-construct vocabulary in both the stem and answer choices of items caused difficulty for students as well as words with negative prefixes (e.g., "dis"). Implications for this research are that readability is correlated with vocabulary (see Rand Reading Study Group, 2002) and that construct and non-construct vocabulary must be clearly defined in order to increase accessibility of assessments.

Table of Contents

Introduction

Recent changes in Federal legislation (including the No Child Left Behind Act of 2001) have placed greater emphasis on large-scale assessment data as a measure of student learning. Because of this focus, schools and school districts are accountable for ensuring student success on assessments. According to federal guidelines, all students must succeed on state-mandated assessments of learning, including students with disabilities.

In order to ensure that all students, including students with disabilities, are taking assessments that are accessible to a wide variety of student needs, Thompson, Johnstone, and Thurlow (2002) recommended that a "universal design" approach be used when designing assessments. Universal Design of Assessment (UDA) is broadly defined as assessments that are "designed and developed from the beginning to allow participation of the widest possible range of students, and to result in valid inferences about performance for all students who participate in the assessment" (Thompson et al., 2002, p.5).

The term universal design was originally used as an architectural concept (Center for Universal Design, n.d.). Philosophies of access and barrier removal were the foundations for early universal design approaches. As part of this design philosophy, ramps, elevators, expanded doorways, signs, bathrooms, and other features do not have to be added or modified at additional expense after the completion of a building. Rather, as part of universal design, these features are sketched into structures' blueprints from the beginning. The promise of universal design is that some of the same architectural features that accommodate people with disabilities also benefit many others, including senior citizens, families with young children, and delivery people.

Thompson et al.'s definition was also initially used to define UDA of general education assessments. The authors further explained UDA by defining seven "elements" of universal design, including: (1) inclusive assessment population; (2) precisely defined constructs; (3) accessible non-biased items; (4) amenable to accommodations; (5) simple, clear, and intuitive instructions and procedures; (6) maximum readability and comprehensibility; and (7) maximum legibility.

This study builds on a series of previous studies conducted by the National Center on Educational Outcomes. Beginning in 2002, Thompson et al., Johnstone (2003), Johnstone, Thompson, Moen, Bolt, and Kato (2005), and Johnstone, Bottsford-Miller, and Thompson (2006) all have sought to operationalize the term "universal design" as it applies to assessments, and to understand the practical implications of universal design processes. One common theme in all of the studies was that language was a particularly important characteristic of items that were and were not accessible.

This study focused specifically on maximum readability and comprehensibility of items in reading assessments. The study was particularly salient to ongoing research on the accessibility of

general education reading assessments (see http://www.narap.info) because it sought to better understand what particular characteristics make a test item readable and accessible.

Readable and comprehensible language was also selected as the focus for this study because it is a somewhat controversial concept in the field of measurement. Tampering with how items are written in order to increase readability has recently been challenged because of the potential effects that language changes may have on the intended constructs of items (Gong & Marion, 2003). In addition, language changes possibly make test items less authentic and reduce their similarity to real-world applications (Kahl, 2006).

Much of the research on the need for readable and comprehensible language as an aspect of universal design comes from literature on students who are English language learners. In some cases, readable text is operationalized as language that is concise and efficient, allowing readers to quickly access test items. For example, Abedi, Hofstetter, Baker, and Lord (2001) found that English language learners often have higher proficiency in mathematics than is reflected on large-scale mathematics tests, and recommended that language be clarified to reflect essential constructs. Likewise, Kopriva, Winter, Emick, and Chen (2006) noted that there are both target and non-target language skills required for assessment items and that by clarifying these language needs it is easier to create more concise language around relevant constructs. Once construct-relevant language is identified, the process of winnowing away superfluous language may improve assessment results and, more importantly, assessment validity.

Readable and comprehensible language can be quantified on readability scales. Shaftel, Belton-Kocher, Glasnapp, and Poggio (2006) applied formulae to predict how readable particular items were on a test. The authors found that students who typically struggle with reading the English language (English language learners and students with disabilities) may be at risk for items containing pronouns, complex verbs, and irrelevant language. In such cases, determining what students do and do not know is complicated by item language, which at times may not allow English language learners and students with disabilities to demonstrate their true capabilities.

As early as 1987, Rakow and Gee outlined a series of considerations for making test items more readable and comprehensible. These included:

- Students would likely have the experiences and prior knowledge necessary to understand the item
- Vocabulary is appropriate for the intended grade level
- Sentence complexity is appropriate for the intended grade level
- Definitions and examples are clear and understandable

- Required reasoning skills are appropriate for students' cognitive level
- Relationships are made clear through precise, logical connectives
- Content within items is clearly organized
- Graphs, illustrations, and other graphic aids facilitate comprehension
- Questions are clearly framed

The RAND Reading Study Group (2002) also stated that difficulty of text may vary due to "complex Boolean expressions." Such expressions are challenging because "the respondent needs to keep track of different options and possibilities" (p. 96). In the case of negative expressions, an unnecessarily high cognitive load may be added to items that employ negatives within items (e.g., "Which of the following is *not* a reason why the captain wanted to turn her ship around?").

The RAND Reading Study Group also found a strong correlation between the level of vocabulary on items and their readability. To address the danger in unnecessarily inflating the difficulty level of items, Haladyna and Downing (2004) suggested removing nuisance variables (such as challenging vocabulary that is not relevant to the construct tested) as a way of improving the validity of tests. Reduction in the level of non-construct vocabulary can be coupled with reducing the density of text (i.e., text that does not "pack too many idea units" or propositions into a single clause) (RAND Reading Study Group, 2002, p. 96) to create readable, comprehensible test items.

In summary, a variety of sources provide information on how to operationalize "readable and comprehensible" text. The purpose of this research was to take the cumulative information from these sources to create readable and comprehensible text in reading items and then to examine their impact on students with learning disabilities. In order to determine the qualitative impact of such interventions, we chose to use think aloud methods.

According to Ericsson and Simon (1994), think aloud verbalizations (where students say all their thoughts aloud while they are approaching an item) are an important data source for understanding cognitive activities. The reason why think aloud data are so rich is because all cognitive processes travel through short-term memory and are spoken at the time they are processed. Thus, think aloud methods provide researchers with qualitative information about the reasons why students may struggle on test items (Leighton, 2004). Such data are informative for purposes of addressing universal design issues in large-scale assessment items (Johnstone et al., 2006).

Methods

Overview

In this study, released National Assessment of Educational Progress (NAEP) items were used as reference or "standard" items from which changes were made to create items that fit our definition of readable and accessible. Specifically, Grade 8 passages and their corresponding items were examined. Researchers at the National Center for the Improvement of Educational Assessment (NCIEA) developed all protocols. Each protocol contained a passage (one informational and one literary passage were used for this study), followed by a series of items. For each passage, standard NAEP released items and modified items designed to be readable and comprehensible were interspersed with one another. Items from the standard set of NAEP released items were reviewed by NCIEA researchers for the following characteristics.

- **Use of pronouns.** A sentence with excessive pronoun use may cause a student to lose track of the main point of reference in an item. For example "Jake left his house to go to visit his grandmother at her house. What was the first thing she said?"

- **Use of negatives.** An unnecessarily high cognitive loading may be added to items that employ negatives within items.

- **Vocabulary.** There is strong evidence that suggests an inverse correlation between the level of vocabulary in text and its readability (RAND Reading Study Group, 2002). For the purposes of this study, we were not concerned with vocabulary in passages, but with the vocabulary level within items. Unless the *construct* of an item is to test vocabulary skills, we attempted to reduce vocabulary demands as a principle of access. For example, "The author's conclusion was a somber reminder of what?" might be rephrased as "The ending of the book was a sad reminder of what?"

- **Non-construct subject area language (specialized vocabulary).** English language arts has a specialized vocabulary of its own (e.g., characterization, denouement, prose, iambic pentameter, etc.). When these words are part of the intended construct, it is appropriate to include them in items. When these terms are extraneous to the intended construct, they may introduce "nuisance variables" (Haladyna & Downing, 2004). We attempted to remove any non-construct subject area language from items.

- **Complex sentences, dense text.** The RAND Reading Study Group (2002) defined sentences with "dense clauses" as sentences that "pack too many constituents or idea units (i.e., propositions) within a single clause" (p. 96). Likewise, the RAND Group defined sentences with "dense noun phrases" as sentences with "too many adjectives

and adverbs modifying the head noun" (p. 96). Items with either one of the above sentence types were rewritten for clarity.

When items contained such features, they were modified to reflect language that was more readable and comprehensible.

Sample

This study took place in the summer of 2007. Eight students participated in the study. All eight had recently completed eighth grade. All of the students had diagnosed specific learning disabilities in the area of reading and were participants in a summer reading enrichment program, but did not all have the same home school. All received services for their disability in public or private schools and were familiar with large-scale assessments. Because the students paid tuition for the private reading enrichment program, it is likely that they were all from middle or high socioeconomic status (SES), although we did not collect SES data. None of the students, however, reported familiarity with the NAEP assessment of reading (from which sample items were derived).

Among the participants, five students were male and three students were female. One of these students was African American, one was Asian American, and one was Hispanic. None was an English language learner.

Procedures

For this study, we followed a standard protocol for each student. First, each student participated in a practice activity in order to better understand how think aloud activities worked. Some of the students had experience with think aloud activities because their teachers had used them in reading and mathematics instruction. None of the students had difficulty with the process. When students did struggle, they were reminded to "keep talking" (Ericsson & Simon, 1994).

Each student read either a literary or informational text passage. Students were randomly assigned to passage types (informational or literary) and "forms" (each form had both standard and modified items, but ordering of items was different across forms). The different forms were designed so that every other question was standard or modified (e.g., Item 1 on Form A was a standard item, item 2 was a modified item, etc. For Form B, item 1 was a modified item, item 2 was a standard item, etc.). Thus, different forms had "matched pairs" of items that were written in standard/modified formats. To this end, every student was likely to answer both standard and modified items, but each student only read one particular passage type (informational or literary).

In this study, students read a passage either silently or aloud (students were given the choice of how they would like to read). Of the eight participants, only one participant chose to read the passage aloud. Once students completed reading the passage, they were asked to answer items one at a time while thinking aloud. Students were asked to read all items aloud while answering them. The interaction between researcher and student then took place in two phases.

In phase 1, the student engaged in a think aloud while the interviewer used only passive prompts to encourage the child to think out loud. During this phase, students read each item aloud and described their problem solving processes while attempting to correctly answer the item. In phase 2, the interviewer asked the child specific questions to probe his or her understanding of the child's cognitive process. Questions that were asked of the students included:

- How did you get that answer?
- What makes you believe that answer is the right one?
- Was there anything that seemed tricky about this question?
- Was there anything that confused you about this question?
- Were there any words in this question that you did not know?
- Could we do anything, change the item in any way, to make it clearer to you?
- Did you think this passage was easy or hard to read?
- Were there any words you did not understand?
- Was any part of it confusing to you?
- Could you find the answers to the questions in the passage?
- Other questions related to content of passage or item.

Researchers praised students throughout the process for thinking out loud and verbalizing, but did not provide any information to students about whether particular answers were correct or incorrect. The purpose of providing feedback to students was to encourage them to continue verbalizing, because this was the primary data source for this study.

Each think aloud activity took approximately one hour to complete. Students were videotaped at two angles to capture student verbalizations plus any activities related to pointing, circling, or other physical acts in which students engaged during the think-aloud session.

Analysis

Data were transcribed from videotapes and coded using a standard coding form. This form asked reviewers to note the following information:

- Student identification number
- Whether the question was standard or modified
- Student comments relevant to the item
- Whether the student answered the question correctly or incorrectly
- If the student answered the question incorrectly, an analysis of her or his error
- Data supporting conclusions made from the error analysis
- Any advice the student had for test makers

Once coding was completed, secondary analyses were conducted to (a) determine the number of correct and incorrect answers for individual items, (b) make comparisons of matched pairs of standard and modified items, and (c) find qualitative patterns in errors made by students.

There was no way to test for statistical significance with the small sample of students who participated in this research. Thus, descriptive statistical information was compared with qualitative information for each item.

Results

Not all students answered all items in this research. This was because students were randomly assigned to either informational or literary passages, and further randomly assigned to "standard" (odd numbered items that students completed were standard and even numbered items were modified) or "modified" (odd numbered items students completed were modified and even numbered items were standard) conditions. Furthermore, we stopped all think aloud sessions after one hour to prevent student fatigue; thus, some items at the end of protocols were not answered at all. Each item pair (its standard and modified versions) is reported below, along with quantitative information on correct/incorrect responses and qualitative information supplied by students. After all items are reported, overall information on the number of correctly and incorrectly answered items by type (standard vs. modified) is reported.

Overall Results

Among students who answered both standard and modified questions, only one student answered more standard questions correct than modified questions. Overall, students answered only 46% of standard items correctly and answered 72% of modified items correctly. Student responses ranged from one student answering three of four standard items (75%) correctly and only one of three (33%) modified questions correctly to one student answering only one of four standard questions (25%) correctly and three out of three modified questions correctly (100%). Although sample sizes were small, the general overall effect seemed to be that students achieved better results on modified items than on standard items. Table 1 represents overall results for standard and modified items.

Table 1. Overall Results for Standard and Modified Items

Item Type	Percent Correct Overall	Range Percent Correct by Participant
Standard	46	25%–75%
Modified	72	50%–100%

Item Level Results

Grade 8, Matched Pair 1 (Literary Passage)

For the first item, the standard version of the item asked the students to "Use the dictionary entry below to answer the question, what meaning of fast is used in line 4?" For this particular item, the modified version of the item simply asked "Which meaning of fast is used in line 4?" Therefore, the modifying process involved reducing the text load of the item.

Both the standard and modified version of the item were answered correctly 50% of the time (1 out of 2 students answered each version correctly). Two of the four students looked back at the passage to examine the vocabulary word in question. One student who looked back at the text (and had the modified item) answered correctly while another student (with the standard item) struggled with the entire item. Although she looked back, she said she was "confused" and wanted to go on to the next item. Students who answered the item correctly (both in modified and standard formats) used deduction to determine the correct answer. Students who answered the item incorrectly did so because they did not grasp the meaning of "fast" within the context of the story. These errors were errors in construct-relevant skills.

Grade 8, Matched Pair 2 (Literary Passage)

The standard version of the second item begins with "In lines 10–13, the poet uses a simile 'his brown skin hung in strips/like wallpaper' to emphasize the fish's." The modified version of the

phrase asked "In lines 10–13 'his brown skin hung in strips/ like ancient wallpaper' is used to emphasize the fish's." This modified version reduced the text load and decreased the vocabulary demands for the item. All students answered this item correctly in both versions (three correct answers, zero incorrect answers for the standard item and one correct answer, zero incorrect answers for the modified item). Two students (both with standard items) answered quickly from memory. The other two students looked back at the passage to answer correctly.

Grade 8, Matched Pair 3 (Literary Passage)

The standard version of the item asked "The image of the fish in lines 43 through 50 develops the fish as having a character that is?" The modified version asked "Lines 43–50 describe the fish as." The modified version contained only seven words and eliminated the word "character."

In both versions of the item, one student answered the item correctly (50%) and one student answered the item incorrectly (50%). One of the students who answered the item incorrectly (modified version) eliminated answers she thought were incorrect, but used faulty logic. The student who answered the standard item incorrectly did so because he could not read the word "character." In this case, an error analysis demonstrated that the use of the vocabulary "character" may have been the cause for one error.

Grade 8, Matched Pair 4 (Literary Passage)

The standard version of item 4 asked "When the poet says 'Like medals with their ribbons frayed and wavering' (lines 61–62) she is referring to." The modified version of this item simply asked " 'Like medals with their ribbons frayed and wavering' (lines 61 and 62) refers to." This item removed the pronoun "she" and reduced the word count on the item from 18 to 14 words.

None of the students who answered the standard question answered it correctly (0 of 2 students). One out of two students answered the modified version correctly. Error analysis of incorrect answers indicates that the two students who answered the standard version incorrectly were unable to infer meaning from the quote provided. The student who answered the modified version incorrectly was confused by a vocabulary word in the item. The student who answered the question correctly re-read the quote, then used deduction (eliminating perceived incorrect answers). It appears as if the elimination of "she" did not make any difference for students. The connection between students' inability to infer meaning in the standard version of the item and characteristics of the item are unclear for this item.

Grade 8, Matched Pair 5 (Literary Passage)

The standard version of this item asked students to "Reread the lines beginning with 'I admired' (line 45) and ending with 'aching jaw' (line 64). The speaker most admires the fish because she

thinks he." The standard version of this item contained two pronouns (she for the author and he for the fish) and 26 words. The modified version of this text contains 12 words and only one pronoun (for the fish). The modified version asks "The speaker **most** admires the fish (lines 45-64) because of his." The word "most" was also presented in bold print in the modified version.

The two students who answered the modified question incorrectly did so because they did not know one of the terms used in the response choices. One student could not pronounce two of the words in the response choices. The one student who answered the modified version of the item incorrectly did so because he depended on his prior knowledge rather than rereading the text to find the correct answer.

In this item, it appears as if the vocabulary level of the response choices in the item were questionable. It is unknown why the same response choices were less problematic for the modified version of the item, but evidence points to a possible compounding effect of challenging items and subsequent answer choices. This item also demonstrated the problematic nature of students depending on prior knowledge to answer questions. Although activating prior knowledge is an important tool in the reading process, if prior knowledge is not combined with information in actual text, it may lead students astray.

Grade 8, Matched Pair 6 (Literary Passage)

The standard item in matched pair 6 asks students "Which of the following best describes the person speaking in the poem?" The modified version of this item begins "Which **best** describes the speaker of the poem?" A second set of revisions in the item also reduces the text load in the latter part of the item by one word (making the total word count 12 words for the standard item and 8 words for the modified item).

All students answered this item correctly. Two students depended on memory and two students eliminated answer choices they felt were incorrect until they arrived at the correct answer. Overall, it appeared as if the reduction in word count (only four words) and the use of bold print for the word **best** did not make a difference for this item as the item was easy for the students in both versions.

Grade 8, Matched Pair 7 (Informational Passage)

The standard version of this item asked "What did the immigrants dislike most about their trip to America?" When modified, the item read "What upset the immigrants most about their trip?" For this item, the negative term "dislike" was replaced by the word "upset." There was clear evidence in this item that eliminating a negative term (dislike) changed how students interpreted the item. Both students who answered the standard item answered incorrectly while both students

who answered the modified item answered correctly. One student answered incorrectly because he relied on memory and did not revisit the passage; the other student read the word "dislike" as "like." The latter student clearly demonstrated the effect of negative words that have prefixes such as "dis" or "non" on readers' comprehension of test items.

Grade 8, Matched Pair 8 (Informational Passage)

The standard version of this item was worded "The statement that immigrants had to 'contend with border guards, thieves, and crooked immigration agents' means that the immigrants?" The modified version of this item was re-written for clarity of purpose, "Immigrants had to 'contend with border guards, thieves, and crooked immigration agents.' This means." The language for the item was simplified and the word "that" was removed twice to present a clearer message to students.

Two students answered the standard version of this item correctly. The student who answered incorrectly struggled to find the quote in the passage and claimed that the item was too difficult to understand. He suggested the item read "what was the main idea of the story." It is unknown whether the modified version of this item would have helped this student answer correctly, or if the item's requirements were simply too difficult for this student. The one student who answered the modified item did so correctly.

Grade 8, Matched Pair 9 (Informational Passage)

The standard version of this item asked students to determine "What most worried the immigrants about the medical examinations?" The modified version of the story attempted to use language for which Grade 8 students may be more familiar, and asked students "What were immigrants most afraid of during medical examinations?"

One out of two students answered the standard item correctly. The student who answered incorrectly looked for the section referred to in the answer choices, but was unable to find those sections in the text. It is unknown whether the modified version of this item would have helped this student, but his error appeared to be the result of his inability to recall features of the story, not an issue related to the item. Both students who answered the modified item did so correctly.

Grade 8, Matched Pair 10 (Informational Passage)

The standard version of this item asked students "The United States eventually reduced the number of immigrants allowed to enter the country because?" The modified version of this item asked "Why did the United States reduce the number of immigrants?" One of two students who answered the standard version of this question did so correctly. The student who answered

incorrectly did so because he answered based on his own prior knowledge of the issue, which tainted his response. Only one student took the modified version of this item and answered incorrectly. This student did so because he thought the word "reduce" meant to "make more." In this item, the version students answered did not seem to make a difference. It is clear, however, that defining the vocabulary term in the stem of the item (e.g., implying that reduction means making less of something) may have changed the outcome of this item.

Grade 8, Matched Pair 11 (Informational Passage)

The standard version of this item asked "In the passage, what is the main purpose of the subheadings?" The modified version of this question simply omitted the phrase "In this passage" and asked "What is the purpose of the subheadings?"

Three students took the standard version of this item, but only one answered correctly. The two students who answered incorrectly did so because they were unfamiliar with the terminology used in the item (subheadings). The one student who answered the standard item correctly and the one student who answered the modified item correctly did so because they were familiar with the expository convention of subheadings from school. In this item, design was not a factor because student success depended on knowledge of the particular convention in the item. Reducing the number of words in the item had neither a positive nor negative effect.

Grade 8, Matched Pair 12 (Informational Passage)

This item again referred to a convention used in expository text. The standard version of the text asked "In the article, how do the quotations by immigrants relate to the sections of the article?" The modified version of this item asked students "The author most likely included the quotations by immigrants to."

Results for this item were similar to the previous item. One student answered the standard version of this item and two students answered the modified version. All students answered the item correctly because they were familiar with the convention to which the item referred. In this case, modifying the item did not appear necessary because all students were able to find the key convention in question and answer based on their prior knowledge of the convention.

Grade 8, Matched Pair 13 (Informational Passage)

The final item related to the main idea of the passage. The standard version of this item asked students "The main point the author is making in this passage is about the?" The modified version asked "This passage is mostly about the."

Although the reduction in language complexity was hypothesized to make the item more accessible, all students who answered this item did so correctly (two students for the standard version and one student for the modified version). All students answered this item very quickly, indicating that finding the "main idea" or what a passage is "mostly about" is a strategy for which students are very familiar. In this case, simplified language was not necessary because of the ease with which students approached the construct.

Summary of Results

Results from this study indicate that modified items may be more accessible to students with learning disabilities. As a whole, students answered modified items correctly at a rate of 72% compared to answering standard items correctly at a rate of 46%. The differences indicate an overall positive effect on student outcomes when using items that are readable and comprehensible, but individual items provide further information on what specific item modification strategies were most meaningful.

For example, the text load (word count) was reduced for modified items 1–5 and 10–12, but reduction in word count did not make a difference in student results. Likewise, the modified version of items 5 and 6 contained bold print. Students commented that they liked the bold print, but it did not seem to affect results. Errors directly related to design of items, however, were present. For example, item 3 contained a vocabulary term that was unknown to the student. Likewise, item 7 (standard) had a negative prefix (dis) that was removed in the modified item. The negative prefix was the cause of student error. Items 5 and 10 both had item-related errors that were not addressed in our modification process, but are instructive in their results. Item 5 had answer choices that were too difficult for students to read and item 10 had a vocabulary word in the item's stem that was problematic in both the standard and modified versions of the item. Both items appeared to be comprehension items, but the vocabulary in the items served as an access point. It is unknown whether making changes to the vocabulary level of these items would affect the intended construct. Because the items were not vocabulary items, it is possible that parenthetical definitions of terms might have helped students, but further research is needed to determine the effects of parenthetical definitions on both student outcomes and test validity.

For most items, the source of student error or success was the particular student's knowledge of the content and story. For items 6, 9, and 11 students answered incorrectly because of flawed reading strategies. In contrast, there were no errors for either the standard or modified version of items 12 and 13 because all students understood the constructs tested. The sources of error were difficult to distinguish from student responses to item 8. Table 2 shows how each item was modified, and what the likely sources of error were.

Table 2. Item Modification Strategies and Student Errors

Item	Modification	Source of Error
1	Reduced text load (word count)	Student reading strategies
2	Reduced text load and vocabulary level	Student reading strategies
3	Reduced text load and vocabulary level	Item stem vocabulary*
4	Removed pronouns and reduced vocabulary	Student reading strategies
5	Reduced word count, important words bold	Vocabulary in answer choices*
6	Reduced text load, important words bold	Student reading strategies
7	Replacement of negative word	Item negative term*
8	Rewritten for clarity of purpose	Error source unknown
9	Rewritten with familiar language	Student reading strategies
10	Reduced text load	Undefined vocabulary in stem*
11	Reduced text load	Student reading strategies
12	Reduced text load	No error
13	Reduced text load	No error

*Indicates item-related sources of error.

Discussion

As states design tests that are intended to be accessible to the diverse general assessment population (e.g., universally designed assessments), continued research is important for defining how test items can become more "readable and comprehensible." To this end, specific research on what strategies are most effective in making test items more accessible can inform practitioners on ways to make items accessible to a wide variety of students, including students with learning disabilities.

This study was limited because of its sample size. Only eight participants engaged in think aloud activities, so results may not generalize to the wider population. Furthermore, the sample was drawn from public and private school students who were actively engaged in reading improvement programs over the summer. Thus, the particular students who participated may not represent the heterogeneity of students with learning disabilities in the United States. It is likely that the particular sample with which we worked was from a higher socioeconomic level than the general population of students with learning disabilities in the U.S. It will be important to conduct additional research, both increase sample size and to increase the diversity of students within the sample.

Despite limitations of this study, informative results emerged. Because think aloud methods can provide in-depth information about student problem-solving processes for test items, error analyses could be conducted on all items to determine the likely source of error. For many items, students answered correctly or incorrectly based on reading strategies and their knowledge of reading conventions. Item modifications such as reducing the number of words in an item and highlighting bold words did not seem to have any effect on student results.

Four items, however, provided information on ways that items can be made more readable and comprehensible. All of these items related to vocabulary. As noted above, English language arts has specific vocabulary and terms that are often tested in large scale assessments. In this research, we found that other words (not constructs that were tested) confused students. Because of this, it may be important to conduct further research on the effects of undefined, non-construct vocabulary in item stems; challenging non-construct vocabulary in item response choices; and words with negative prefixes (such as "dis").

In this research, it was not the number of words that challenged students, but the words themselves. As the assessment community continues to attempt to understand how to make test items as accessible as possible without changing intended constructs, consideration of the role of vocabulary within the items themselves is important. Findings from this research indicate that non-construct vocabulary may obscure what knowledge we can gain about student reading comprehension from tests. Such findings imply there is a need for both further and more comprehensive research on the effects of vocabulary, and continued review by content experts to ensure the level of non-construct vocabulary in items is appropriate.

References

Abedi, J., Leon, S., & Mirocha, J. (2001). *Validity of standardized achievement tests for English language learners*. Paper presented at the American Educational Research Association Conference, Seattle, WA.

Center for Universal Design (n.d.). *What is universal design?* Center for Universal Design, North Carolina State University. Retrieved January, 2002, from the World Wide Web: www .design.ncsu.edu.

Ericsson, K. A., & Simon, H. A. (1994). *Protocol analysis: Verbal reports as data (Revised edition).* Cambridge, MA: MIT Press.

Gong, B, & Marion, S. F. (2003). *Implementing Universal Design principles in educational assessment: Current challenges of construct clarity*. Retrieved June 18, 2006, from the World Wide Web: http://www.nciea.org/

Haladyna, T.M. & Downing, S.M. (2004). Construct irrelevant variance in high stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

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.

Johnstone, C.J., Thompson, S.J., Moen, R.E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design* (Technical Report 41). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kahl, S. (2006). *Universal design vs. depth of knowledge*. Presented at Chief Council of State School Officers Large Scale Assessment Conference, San Francisco, CA, June 25–28, 2006.

Kopriva, R., Winter, P., Emick, J. & Chen, C. (2006). *Achieving accurate results for diverse learners: Access-based item development.* Paper presented at American Educational Research Association Meeting, April 7–11, 2006.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6–15.

RAND Reading Study Group (2002). *Reading for understanding: Towards an R & D program in reading comprehension.* Retrieved March 10, 2007, from the World Wide Web: http://www.rand.org/pubs/monograph_reports/MR1465.pdf

Rakow, S. J. & Gee, T. C. (1987). Test science, not reading. *Science Teacher, 54* (2), 28–31.

Shaftel, J., Belton-Kocher, E., Glassnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11* (2), 105–126.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota,National Center on Educational Outcomes.