Chapter 1

Adapting Reading Test Items: Decreasing Cognitive Load to Increase Access for Students with Disabilities

Caroline E. Parker
Education Development Center, Inc.

Joanna Gorin
Arizona State University*

Sue Bechard
Measured Progress

This manuscript was supported, in part, by the U.S. Department of Education Office of Elementary and Secondary Education (Grant No. S368A070004). However, the opinions expressed do not necessarily reflect the position or policy of the U.S. Department of Education and no official endorsement should be inferred.

* *Joanna Gorin is now Research Director of the Cognitive & Learning Sciences Group at Educational Testing Service, though all work on this research was completed while at ASU.*

Abstract

The Adapting Reading Test Items to Increase Validity of Alternate Assessments Based on Modified Academic Achievement Standards (ART 2%) project used multiple research methods to explore the impact of item modifications on the ability of high school students with disabilities to access a large-scale reading assessment. Beginning with item difficulty modeling and exploratory cognitive interviews, researchers identified aspects of the assessment that increased the cognitive load for students with disabilities. In consultation with test developers, item modifications were developed that attempted to lessen the cognitive load, break down cognitive barriers, and reduce the item difficulty for students, while maintaining the grade level content of the assessment. A second round of research was conducted using the modified items; a pilot study of all the items was administered to 1051 students, and 32 students participated in confirmatory cognitive interviews to explore the impact of the item modifications on cognitive load. Case summaries of the students who participated in the confirmatory cognitive interviews provided information about the characteristics of students who are not well served by a regular assessment. Students who participated in the cognitive interviews were notable for their diversity of reading levels, of levels of success on both unmodified and modified assessment items, of access to appropriate learning opportunities, and of their ability to communicate their own cognitive processes. Results of the study indicated that while it was possible to identify cognitive barriers in assessment items, it was difficult to develop modifications within the constraints of a grade-level paper and pencil assessment that would sufficiently address the barriers and increase access to the assessment for students with disabilities. Some item and test features that were found to improve test accessibility could be incorporated in all tests as principles of universal design, while some manipulations that reduced item difficulty for the target population would not be appropriate for students without disabilities. Implications for both instruction and assessment are discussed.

Introduction

Funded in 2007, the Adapting Reading Test Items to Increase Validity of Alternate Assessments Based on Modified Academic Achievement Standards (ART 2%) study was the third in a series of Enhanced Assessment Grants awarded to four New England states and Montana. The first grant, from 2003-2005, focused on a study of specific technological adaptations to large-scale assessment mathematics items in order to increase accessibility for students with disabilities. The second, from 2005-2007, focused on identifying the characteristics of the subgroup of students with disabilities who are not well-served by large-scale assessments because the assessments do not effectively measure what they know and can do. The third, ART 2%, from 2007-2010, used cognitive modeling strategies to identify the cognitive barriers to accessing the assessment and cognitive interviews with students to understand more about the cognitive characteristics of students eligible for the AA-MAS. The goal of the study was to identify manipulations to items that would lower the cognitive load for students, thus increasing the ability of students with disabilities to show what they know and can do on large-scale assessments. It was not the intent to produce an operational assessment. By using cognitive modeling strategies with the assessment items and cognitive interviews with students with disabilities, the study identified linguistic and formatting barriers in items, manipulated the items to reduce the cognitive load, and studied the impact of the manipulations through a pilot study of 1,000 students and in-depth cognitive interviews with a sub-sample of 32 students. This dual approach to cognition, using cognitive modeling strategies with the items themselves, and directly observing students' cognitive practices when interacting with the items, allowed researchers to both better describe the cognitive characteristics of assessment items and manipulations, and also to describe the cognitive characteristics of this elusive population of "2%" students.

Literature review

The Students

The regulations that established the "2% option," published in 2007, identified the following criteria to determine eligibility for the 2% assessment: students with disabilities, from any of the 13 disability categories; students who are addressing grade level content standards on their Individualized Educational Programs (IEPs), but are not expected to meet grade level achievement standards in the current year; and students who need less difficult test items, covering the same breadth of content (U.S. Department of Education, 2007). This definition was problematic. Previous research on students for whom the large-scale assessments may not be valid with or without accommodations found two groups of students who are not accurately assessed through the large-scale assessment (Bechard & Godin, 2007; Parker & Saxon, 2007). Students in the

first gap perform proficiently in the classroom, according to teachers, but do not demonstrate proficiency on the assessment. These students are more likely to be in general education and without identified disabilities. Students in the second gap consistently perform well below grade level in the classroom, and perform no better than chance on the large-scale assessment; they are more likely to have identified disabilities. A third group of students, both with and without disabilities, also performs below grade level in the classroom, and their low performance on the large-scale assessment confirms their classroom performance. For these students, the assessment is an accurate measure of their below-proficiency performance, and their needs can be addressed through instruction rather than assessment design.

This description of the types of students not served well by the large-scale assessment system indicated that the group most in need of different options for large-scale assessment is the group of students with disabilities performing far below grade level in the classroom and no better than chance on the assessment. However, the regulations as released in 2007 defined the students and the assessment differently; the students have IEPs addressing grade level content, and the assessment "must cover the same grade-level content as the regular assessment" (U.S. Department of Education, 2007). Thus, the students of most concern to teachers, students with disabilities far below grade level but above the 1% alternate achievement performance level, would likely not benefit from the 2% assessment.

A Cognitive-Modeling Approach to Item Modification

A cognitive model of item or test performance is a representation of the cognitive processes, skills, abilities, or knowledge required to answer a question (Leighton & Gierl, 2007). One approach to the development of a cognitive model is via item difficulty modeling. An item difficulty model (IDM) is a list of variables describing features of test questions that vary across items; the variability of these features is shown to explain variability in item statistics such as item difficulty. By establishing a relationship between variability in item features and variability in item difficulty, it is presumed that the item processing has been explained by the processes associated with the features. When the impact of item features can empirically and mathematically be linked to item parameters, such as difficulty and discrimination, then items can be written to target specific ability levels and testing populations most efficiently. Further, once the factors affecting difficulty have been identified and evaluated empirically, assessment developers can use this information to alter item content and format to target the cognitive processes one *wants* to measure while controlling for, isolating, or removing those one *does not want* to measure.

Cognitive Models of Reading Comprehension Assessments

To correctly answer passage-based multiple-choice reading comprehension test items it is presumed that students read the passage, or passages, then read the question and select an answer from among the possible options. Research on similar assessment items has shown that the

variables influencing item difficulty include characteristics of the passage, the question, the response options, and the interaction between the three (Embretson & Wetzel, 1987; Gorin & Embretson, 2006; Sheehan & Ginther, 2001; Sheehan, Kostin, & Perskey, 2006). The processing components of prime interest vary across models and include propositional density (Embretson & Wetzel, 1987; Kintsch & vanDijk, 1978), context of to-be-learned information (Landauer, 1998; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998), location of relevant information (Freedle & Kostin, 1992; Sheehan & Ginther, 2001), and correspondence between passage and question information (Alderson, 1990; Embretson & Wetzel, 1987; Freedle and Kostin, 1993; Sheehan & Ginther, 2001).

Embretson and Wetzel (1987) developed a cognitive processing model of reading comprehension to describe the processing difficulty of the Armed Services Vocational Aptitude Battery (ASVAB) items. To validate the model with these items they identified stimulus features that were theoretically related to processing components of the model, and then scored items in terms of the features. The model describes sources of cognitive complexity derived from two general processes: Text Representation and Response Decision. *Text Representation* processes consist of the encoding and construction of the passage for a set of items. The difficulty of encoding is controlled by linguistic features of the passage, particularly vocabulary difficulty (Drum, Calfee, & Cook, 1981; Graves, 1986). Passages with high levels of vocabulary are more difficult to encode, and consequently more difficult to retrieve when responding to comprehension questions. Construction involves processes of connecting word meanings and propositions into a meaningful, coherent representation of the text. Kintsch and vanDijk (1978) described text comprehension as an iterative process of construction and integration wherein text is processed as propositional units that are continuously integrated with prior knowledge. The constuction-integration theory (Kintsch, 1988, 1998), derived from earlier work by Kintsch and vanDijk (1978) characterizes text comprehension as cyclic propositional processing. During each processing cycle propositions are retrieved from the text and arranged into a network. At the integration phase, activation spreads throughout the network and accumulates primarily at points of high interconnectivity. Following each of these cycles, the most highly activated propositions are carried over into the next cycle with working memory for further processing (Kintsch & vanDijk, 1978). The difficulty of construction processes is most strongly influenced by the propositional density of the text, which is the ratio of the number of propositions to the total length of the passage. Several studies have concluded that propositionally dense text is difficult to process and integrate for later recall and comprehension (Kintsch, 1994; Kintsch & Keenan, 1973; Kintsch & vanDijk, 1978). This finding may be related to limitations in working memory capacity that preclude holding large amounts of information simultaneously. If propositions are not well integrated into the working knowledge representation, then the information may not be available for later recall (Kintsch, 1994 ; Kintsch & vanDijk, 1978).

The remainder of the model describes three *Decision Processes*: encoding and coherence, text mapping, and evaluating the truth status of the response alternatives. Encoding and coherence are the same as in text representation except that they apply to questions and response alternatives rather than the passage. Text mapping is the process of relating the propositions in the question and response alternatives to the information retrieved from the passage. Difficulty in text mapping is partially influenced by the amount of information needed from the text to answer the question. According to Embretson and Wetzel (1987), as the amount of text relevant to answering a question increases, so do the demands on memory, encoding, and item difficulty.

Evaluating truth status involves a two-stage process of falsification and confirmation of response alternatives. The decision processes of falsification and confirmation were the strongest predictors of item difficulty in the Embretson and Wetzel study (1987). These two decision processes describe the extent to which information given in the passage could be used to make decisions regarding the response options. Items with correct responses that were directly confirmable or distractors that were explicitly contradicted by the text required little processing. Their findings are consistent with other research suggesting that the overlap or matching between the text and a question can affect response processes (Alderson, 1990; Freedle & Kostin, 1992, 1993).

Embretson and Wetzel postulated that decision processes were also affected by vocabulary difficulty of the response options. The vocabulary level of the response alternatives affected the likelihood that an examinee would consider the alternative as a potential correct response. Distractors with difficult vocabulary were less likely to be processed for consideration as potential alternatives and required less processing than low-difficulty vocabulary distractors as measured by response time and item difficulty. The reverse effect was found for the vocabulary level of the correct response. Examinees were less likely to confirm a response alternative if the vocabulary level was high.

In addition to vocabulary level, the phrasing of the information in the alternatives also affects decision processes. The reasoning level of the response alternatives represents the relationship between the structure of the propositions in the alternatives and those in the passage. Anderson (1982) proposed a taxonomy describing the levels of transformation needed to match a question to text. The lowest level is verbatim, in which the exact words used in the question are found in the passage. These items are assumed to be easy because little to no transformation of information must be conducted to identify the location of the item answer. The highest level question is transformed paraphrase, in which neither the order nor the wording of information in the question matches that of the passage text. Items with transformed paraphrase questions are assumed to be hard because ideas in the passage must be reworded and reordered to map the question to the location of the information needed to correctly answer it (Craik & Lockhart, 1972).

Similar research by Sheehan and Ginther (2001) examined the relationships between task features and item difficulties for Main Idea Reading Comprehension (RC) questions from the Test of English as a Foreign Language (TOEFL-2000). Sheehan and Ginther modeled item difficulties in terms of *Activation Processes* by which an individual selects a response alternative. They described a memory-type model of processing, in which the examinee selects the response that is most highly activated in the individuals' mind. First, grossly incorrect distractors are eliminated during early global falsification. The remaining model of item difficulty was defined by item and passage features that define two intermediate structures: activation of the key (the correct answer) and activation of the remaining distractors (the incorrect choices). The activation of the response option is similar to activation of nodes in memory theory; the element with the highest level of activation is most likely to be selected. In the context of multiple choice questions, the response alternative that is most highly activated is selected as the correct answer. Questions with high key activation and low distractor activation should be easy because the key is far more likely to be selected than any of the incorrect responses.

Sheehan and Ginther (2001) found three types of item and passage feature effects to be critically important for defining activation in Main Idea questions: Location Effects, Correspondence Effects, and Elaboration of Information. *Location Effects* refer to the location within the text of relevant information for answering a particular question. Kintsch (1998) suggested that as comprehension proceeds while reading a text, the location of information in mental representation, the representational text, is related to the location of the information in the text itself. Therefore, information closely positioned in the text is more easily found in and retrieved from memory because it is stored in relatively close proximity. The location of relevant information within the text was found to be related to comprehension item difficulty. Furthermore, Sheehan and Ginther (2001) found similar results to those of Freedle and Kostin (1993) such that information found earlier in a passage was more easily accessed than information found later in a passage.

The second activation-related effect, *Correspondence Effects*, refers to the lexical and semantic similarity between the response option and the text, or what Freedle and Kostin (1993) might call lexical overlap. The operationalization of correspondence in Sheehan and Ginther's model included literal and synonymous processing, comprehension of difficult vocabulary, understanding of metaphorical word use, proficiency at generalizations, and ability to generate inferences to bridge text. This definition, based on Mosenthal and Kirsch (1991), is a more elaborate characterization than earlier lexical models of correspondence such as that suggested in Drum et al. (1981; see above). Activation of the key based on correspondence is a decision process and is affected by the similarity of the information presented in the key and the information in the text. When the wording is similar, less processing is required to correctly map the key to the text.

Sheehan and Ginther's (2001) third effect, *Elaboration of Information*, represents the extent to which the topic of the question is discussed within the passage itself, and to what level of

detail it is described. Memory research (e.g., Craik & Lockhart, 1972) suggests a strong positive relationship between the amount and nature of information elaboration and the level of activation of information for later recall. When highly elaborated information in the text appears in one of the response options, this option becomes highly activated. As with the location and correspondence effects, high key activation contributes to item easiness, and high distractor activation contributes to item difficulty.

Implications for Reading Comprehension Assessment Design

Across the range of cognitive models and IDM studies for passage-based reading comprehension test items, several consistent findings emerge with implications for enhanced assessment design. First, it is clear that the linguistic features of all aspects of the test questions—the passage, the questions, and the response options—affect students' response processes. This is not surprising, nor undesirable. Second, the relationship between the question and the passage affects student processing. This suggests that two items with identical linguistic characteristics—including reading difficulty—but that differ with respect to the relationship between the question and the passage could be processed by students quite differently. Finally, the structure of the item itself, specifically the phrasing of the response options in relation to the question stem and the relevant portions of the passage, significantly impacts student problem solving. Thus, if items are to be modified in terms of cognitive load and resources, then attention must be paid to items in their entirety, not just linguistic structure and not only the text in the reading passage.

Cognitive Interviews: A Tool to Access Student Cogntivie Processes in Test Taking

Cognitive interviews are being used more frequently in assessment development in order to better understand the cognitive processes used by students with disabilities when taking a reading assessment (Almond et al., 2009; Ferrara et al., 2004; Winter, Kopriva, Chen, & Emick, 2006). The data gathered about student thinking processes contribute to understanding the processing requirements of different items. Researchers do not consider cognitive interviews as a single research method (Ericsson & Simon, 1993), but rather a variety of techniques of self-report, self-observation, and self-revelation that can be used in combination or on their own to study the thoughts and cognitive processes of individuals in a variety of settings or tasks (Alavi, 2005).

Such cognitive interviews are increasingly being used during item writing to gather empirical information on how students interpret items and whether items perform as intended (Ferrara et al., 2004; Leighton, 2004). The interviews can point out where students misinterpret items or can explain unpredictable response patterns long before items function invalidly in pilot or operational administration (Capraro & Joffrion, 2006; Gorin, 2006; Winter et al., 2006).

According to Alavi (2005), there has been much debate surrounding the use of cognitive interviews dating back to beginning of the 20th century when researchers first successfully defended the credible use of introspective reports for investigation of mental processes. Later cognitive

interviews became less accepted as behaviorism gained popularity but eventually were reintroduced into research by contemporary cognitive psychologists (Ericsson & Simon, 1993). Even so, conflicting views continue to persist regarding their suitability for studying cognitive processes (Pressley & Afflerbach, 1995; Nisbett & Wilson, 1977). Ericsson and Simon (1993) warn that inaccurate verbal accounts may not only be due to the participants' inability to access their mental processes, but also to "inadequate procedures for eliciting verbal reports or requesting information that could not be provided even if thoughts were accessible" (p. 45). Cohen (1994, as cited in Alavi, 2005) provides a more exhaustive list of issues and limitations including: inaccessibility of cognitive processes (Seliger, 1983); mismatch between the subjects' verbal response and their natural thought processes (Ericsson & Simon, 1993); conversion of introspection into retrospection (Boring, 1953); an intrusive effect of verbal protocol and the possible distortion or the process or the task the subjects are asked to do (Mann, 1982); variety of verbal protocols according to the type of instructions given, the types of material used in collecting protocols, and the nature of the data analysis (Olson, Duffy & Mack, 1984); alteration of the original thought processes if respondents do a task in a target language and report on it in their native language or another language (Faerch & Kasper, 1987); and, the impact of subjects' characteristics (e.g., their verbal skills) on verbal protocol (Olson et al. 1984)

Despite these challenges, researchers continue to recommend the use of cognitive interviews "for the purpose of detailed examination of the information to which people attend while performing tasks" (Alavi, 2005, p. 3) and have acknowledged its potential to provide information to test hypotheses and models of behavior (Ransdell, 1995). *Think aloud* protocols are one type of verbal report that is commonly used to examine test items. With this technique, subjects report their thinking as they do a task. According to Pressley and Afflerbach (1995) this approach has at least three advantages: (1) verbal reports can provide data on cognitive processes and reader responses that otherwise could be investigated only indirectly; (2) verbal reports can provide access to the reasoning processes underlying sophisticated cognition, response, and decision making; and (3) verbal reports allow for the analysis of affective processes of reading in addition to (or in relation to) cognitive processes.

According to Ericsson and Simon (1993) this approach focuses on two constructs in information processing theory: long term memory, and short term memory. Long term memory is vast in capacity and includes our procedural (how to do things) and factual knowledge. Short term memory, on the other hand, is extremely limited in capacity and is typically used to refer to the information currently held in consciousness that is derived from external stimulation and long term memory. Think aloud protocols generally target information held in short term memory, whose current contents can be quickly accessed and reported, but it is also often possible to report what was recently held in short term memory as some of the contents of short term memory are converted to long term before they exit short term awareness. This allows individuals to recollect what they were thinking about a short time ago and can be accessed using a *retrospective*

think aloud, a process by which participants describe their thinking processes shortly after the event (such as completing a reading test and then describing the thought processes during the test). However, the validity of these recollections will decrease over time and the quality will depend on retrieval cues.

Another area of contention is the effectiveness of different retrieval cues or prompts. Prompts may be open ended or can direct participants to provide very specific types of information. Given open ended direction, participants might feel compelled to report any and all information that they can access in short term memory, whether or not it is relevant to the primary interest of the researcher (Ericsson & Simon, 1993). Ericsson and Simon also caution that some questions do not stimulate accurate verbal reports. Specifically, people often cannot accurately answer "why" questions regarding the motivation for their behaviors. Individuals can usually produce some answer to this type of question, but may try to produce a logical answer or one to fill in a gap in their thinking, instead of the exact contents of their thoughts. Instead, Ericsson and Simon (1993) recommend that prompts used in think aloud protocols aim to elicit the exact contents of short term memory. The farther removed verbal reports are from the exact contents of short term memory, the less valid they may be considered. For example, under this guideline, a report of "5-2-9-3-9-6-7" is a more credible response than "I am thinking of my telephone number" because the latter is not the exact content of what was held in short term memory. Ericsson and Simon (1993) argue that subjects can report immediate and final products of problem solving with greater accuracy, much more certainly than the processes per se and that it is the job of the researcher to infer cognitive process from these reports.

Ericsson and Simon (1993) provide several methodological recommendations to enhance the validity of think aloud data. As mentioned earlier, ideally, self reports should reflect exactly what is being thought by subjects. This can be accomplished through clear direction to participants that they should not attempt to make self reports more coherent. Researchers, not participants, make inferences about processes used to complete task/items during data analysis. This guideline can be particularly helpful for researchers concerned that participants may find the task of thinking aloud overwhelming or confusing. The goal should be to encourage participants to simply repeat the thoughts aloud, rather than to provide interpretative descriptions of their thought processes. As supplemental information, researchers may also want to ask participants to point to, highlight, underline or otherwise indicate what portion of test materials they are using, rather than say it aloud, but they recommend that subjects be discouraged from self reporting why they are carrying out a process as such explanations may heighten awareness of the effects of processing and affect subsequent processing.

Another important point for researchers to keep in mind as they design and use "think aloud" protocols is that as people learn and become more familiar with procedures, their processing becomes more automatic. Fully automatic processes will be more difficult to self report as subjects

may be unaware of them. Therefore, protocol analysis will be much more sensitive to processes that have not yet become automatic, which have remained under the participants' conscious control. Researchers using think aloud protocols to examine test items may want to consider the exposure students have had to academic material, test materials, and test taking strategies. Students who are more familiar with content and test materials. or have had more opportunities to employ test taking strategies may find it difficult to report all their thought processes as they work through questions. Researchers may also want to consider the difficulty of items being investigated. Easy items may be processed automatically and consequently students may have difficulty reproducing their thoughts as they work.

Ericsson and Simon (1993) also conclude that thinking aloud is a natural enough process that lengthy training is not required for adults to be able to carry it out. They provide little guidance, however, about who should be better able to self report and who would be disadvantaged. Specifically, they mention very little about the extent to which verbal protocols can be used with younger children and students with disabilities. More recent research from Johnstone, Bottsford-Miller, and Thompson (2006) found that most fourth-grade and eighth-grade students with disabilities were able to verbalize while thinking aloud. Ericsson and Simon (1993) do note that as a task proceeds, people sometimes forget to think aloud and recommended the use of reminders or gentle prompts to continue to think aloud if a participant is silent for a length of time.

Exploratory Phase

This chapter describes the triangulation of the cognitive modeling and cognitive interviews, and how the two studies contributed to understanding more about both item characteristics for a 2% assessment and the characteristics of the students who could be eligible for such an assessment. The study was conducted in two stages. In the first stage, the cognitive modeling and the cognitive interviews sought to identify the specific barriers to accessing the assessment, and provided specific recommendations for item manipulations. The first phase addressed these questions:

1. What item features of the assessment predict difficulty and cognitive processing that could be manipulated to reduce cognitive load without reducing construct-relevant processes?

2. What cognitive barriers do students with disabilities describe when taking a high school reading assessment?

Exploratory Phase Research Design

The research design was divided into three parts: the first part examined the original assessment to identify possible item modifications; the second part included a revision of 34 items; and the third part examined the revised items. This paper focuses on the first and third parts of

the research study. In the first part, cognitive modeling and cognitive interviews were used to identify cognitive barriers.

Cognitive Modeling

Sample

Student responses were selected from the state populations based on inclusion criteria intended to match the target population. The target population for this study was defined as students with a documented disability in the state, district, or school record who scored no more than 1 standard error above the state's cut score defining proficiency. The target population data included scored item responses from a total of 5236 respondents' data (1408 from Maine, 399 from Montana, and 3429 from the New England Common Assessment Program—NECAP).

Data

The data analyzed included scored item responses (i.e., 0 or 1) to 81 multiple-choice items selected from the 3 state assessment programs (Montana 21 items, Maine 48 items, and NECAP 12 items). The results presented here pertain only to the data from the target population[1] of students in the five states.

Attribute List

A total of 47 item attributes were coded for analysis, each belonging to one of four general categories: *Key and Distractor—Linguistic* attributes, *Key and Distractor—Reasoning* attributes, *Necessary Information* attributes, and *Latent Semantic Analysis (LSA)* attributes. The Key and Distractor—Linguistic attributes included variables associated with the length and vocabulary level of the keyed response and the four distractors for each item. The Key and Distractor—Reasoning attributes described features of items associated with confirmation of the correct response, falsifiability of the distractors, level of transformation of the key relative to the text, reasoning level of the key, and plausibility of the distractors. Finally, LSA attributes measured the lexical correspondence between the keyed response and the distractors with the relevant portions of text for each questions (i.e., Necessary Information).

Attribute Coding

To begin, the necessary information (NI) required from the passages for each question was identified by four members of the project advisory panel with expertise in ELA assessment and theory. IDM research on text-based reading comprehension items has shown that the features of the entire passage associated with a question are of less relevance than characteristics of the subset of the text containing the information needed to answer the question (Sheehan & Ginther, 2001). The NI could consist of as little as one sentence of text, and up to as much as the entire text (1 or 2 passages).

[1] Target population refers to the portion of the student population selected for the study as potential candidates for future administration of enhanced assessments.

The Key and Distractor–Linguistic attributes were coded using natural language processing (NLP) tools available in Microsoft Word 2003. The Key and Distractor–Reasoning attributes were coded to describe reasoning and higher level cognitive processes associated with the interaction between the keyed response and distractors, and the NI. These attributes were scored by human raters based on the coding rules developed in previous reading-comprehension modeling literature (Embretson & Wetzel, 1987; Gorin & Embretson, 2006, Sheehan & Ginther, 2001). For example, for the Paraphrase Level of the Key, Level 1 was defined as Verbatim, Level 2 was defined as Verbatim Transposed, Level 3 was defined as Paraphrased, and Level 4 was Paraphrased Transposed. For the necessary information attributes, the portions of text designated as NI by the ELA experts were coded in terms of linguistic characteristics using two NLP tools, Coh-metrix 2.0 (Graesser, McNamara, Louwerse, & Cai, 2004) and the Microsoft Word tools, the same tools used to code the Key and Distractor–Linguistic attributes. Finally, the LSA attributes were generated based on the cosines of the distance between the vectors of the response options and the NI in latent semantic space. The semantic distance between the response options and the NI is assumed to represent the correspondence between the two. According to activation theories of cognition, information that is more similar (i.e., more closely located in semantic space) should be more highly activated. The response option that is most highly activated is the one most likely to be selected. Sheehen, Kostin, and Persky (2006) showed that when keyed responses are more highly activated due to high correspondence with NI, items will be easier; whereas, when distractors are more highly activated than the key, the item is more difficult.

Analysis

The purpose of these analyses was to examine the utility of the preliminary list of item features/attributes that have variability across items, and explain significant amounts of variability in item difficulty. Descriptive statistics of the attributes, including means, standard deviations, and bivariate correlations between attributes and item difficulty for the target population were calculated. Next, scored item responses were modeled via the linear logistic latent trait model (LLTM; Fischer, 1973) separately for each state's data.

The LLTM incorporates content information into the calculation of probabilities of a correct response to an item. Explanatory item response models like the LLTM have been successfully applied to reading comprehension (Gorin, 2005), abstract reasoning (Embretson, 1998), spatial reasoning (Embretson & Gorin, 2001), and mathematical reasoning (Embretson & Daniel, 2008) to identify cognitive processes associated with student response behaviors and item difficulty estimates. Examination of the attributes included in an item model can provide information regarding the nature of the trait measured by the item itself. On the basis of the weights for each attribute, modifications to items could be considered that should alter the cognitive processing in prescribed ways. That is, test developers can control the contribution of specific cognitive skills by manipulating empirically validated characteristics of test questions.

Separate LLTM models were fit for each state's data. Overall model fit was determined based on the correlation between the LLTM predicted difficulty estimates and the Rasch based estimates. Multiple *R*s are reported for each model. Further, the impact of each item attribute was examined in terms of a *t*-statistic. Attributes with significant *t*-statistics were assumed to have a significant impact on item difficulty.

Exploratory Cognitive Interviews

Sample

The target population for the cognitive interviews was students with disabilities at the high school level who took the regular state reading assessment (with or without accommodations) and who did not reach proficiency on that assessment. They also had to be able to communicate verbally. For the cognitive interviews, the sample described may have included students who may be higher-performing than the AA-MAS population, but who face similar cognitive barriers, and who also may be better able to articulate those barriers. Thus, the pool provides particularly useful data in the cognitive interviews.

Once the pool of students was determined, schools were selected by convenience using researcher and state contacts. Within the schools, all eligible students were invited to be part of the study. In most of the schools in the sample, 80-100% of the eligible students participated. This resulted in a final pool of 27 students. Of those, one student withdrew after the practice session, and a second student did not have sufficient verbal skills to participate, leaving a final sample size of 25 students. As shown in Table 1, the largest group of students have learning disabilities.

Table 1. Identified Disabilities on Student IEPs

Disability Types	Number of Students*
Autism	0
Emotional Disturbance	1
Cognitive Disability	3
Traumatic Brain Injury	1
ADD/ADHD	6
Specific Learning Disability	12
Other Health Impaired	3
Missing	3

*Adds up to more than total students because of multiple identifications.

The students had received different amounts of special education services (Table 2), ranging from less than five hours weekly (n=11) to more than 20 hours weekly (n=5).

Table 2. Total Weekly Hours in Special Education Services

Hours	Number of Students
Less than 5 hours weekly	11
5-9 hours weekly	3
10-14 hours weekly	5
15-19 hours weekly	3
more than 20 hours weekly	3

All of the students had taken the regular state assessment, some with accommodations. None had reached proficiency. Table 3 divides the scores (all below proficient) into three sub-categories.

Table 3. State Assessment Proficiency Categories

Performance Level	Number of Students
Substantially below proficient*	6
Mid-range	12
Almost proficient	6
Missing	1

*While the state proficiency categories have two levels for students performing below proficient, we have created three levels to show the number of students at the lowest range.

According to the students' teachers (Table 4), 13 were reading below grade level and 7 were reading at grade level (five missing).

Table 4. Teacher Estimation of Student Reading Ability

Reading Ability	Number of students
No, not reading on grade level	13
Yes, reading on grade level	7
Missing	5

Data Collection

The project research team chose to use all released reading passages and items from the three state systems in the first round of cognitive interviews, resulting in a total of 12 passages and 81 items. Eight forms were developed with passages in different orders, and students were given one of the forms, with the goal of having at least nine students answer each item. Students were interviewed twice, completing two passages the first day and two the second day. They worked in each session for 45 minutes, and stopped when the time was up, whether or not they completed the assessment. Thus, for many of the longer passages, fewer students attempted the later questions.

Four researchers conducted the interviews. Researchers were trained together and observed each other conduct interviews, then conducted them separately. An interview protocol, adapted from King and Laitusus (2008), provided detailed instructions about conducting the cognitive interview, including how to conduct practice sessions with the students, appropriate probing questions to use, and follow-up retrospective questions for each item and each passage. The protocol designed for the cognitive interviews was tailored to identify the processing requirements of each item. Based on the literature, a number of decisions were made about the interview process:

- Both concurrent and retrospective prompting were used for each passage and item. Students were instructed to describe all their thinking processes, and encouraged with prompts such as "what are you thinking right now?" They were also asked questions after each item and passage to elicit more information about their cognitive processing.

- Despite the recommendation to not ask students "why" they chose a certain answer, researchers decided to ask the students why they had chosen each answer. For many of the students, their disability interfered with their meta-cognitive processing, and they found it difficult to spontaneously describe their thinking processes. Asking them why helped them to unpack those processes. We decided that the benefit of increased information outweighed the risk of providing students with "scaffolding" that would have an impact on later items.

Additional data on each student included written researcher notes, transcripts, IEPs, and teacher interviews. Students' reading teachers were interviewed to obtain information about each student's access to the assessed curriculum, instructional supports provided, and the teacher's perspective about the student's cognitive development in reading comprehension. Information was also collected on students' course sequences and levels, grades, and accommodations. The instructional and performance data helped to interpret the extent to which student-item interactions are strictly item-related or confounded by instructional factors.

In the protocol for the exploratory interviews, students were given an overview of the objectives of the study along with the chance to opt out. Once verbal permission was obtained for recording, the interview commenced with a warm-up reading exercise. During this exercise the students were introduced to the think-aloud process, which the interviewer demonstrated and the student then practiced. Once it was established that the students understood the process, the interview began. For each passage the student was given the opportunity to answer the items in any way that he/she chose, whether reading the passage through or going directly to the items, or a combination of the two. They could also request that the passage be read aloud. They were reminded to verbalize all of their thoughts as they completed the items, and interviewers used prompts as necessary for reticent students.

Analysis

Four people participated in coding the exploratory interviews in two stages. First, all passages and items were coded to identify patterns in cognitive errors. An initial coding scheme included reading patterns (full passage then items, skimming, etc.), ability to identify important information, demonstration of understanding, recall, decoding, and answer strategies (executive processes) (Pressley & Afflerbach, 1995). From this coding, the most common cognitive challenges were identified and the passages and items most amenable to manipulations were chosen. Second, the four passages chosen for further manipulation were then coded in greater detail, using the same codes, as well as codes that emerged as recommended manipulations were identified. Two broad categories of cognitive barriers were identified: *linguistic barriers* that included the structure of the stem (open vs. closed question format), vocabulary, order of answer options, attractive distractors, and lack of question clarity. *Formatting barriers* included lack of visual links between each item and its corresponding passage text and physical distance of the item from the corresponding passage text.

Exploratory Phase Findings

Cognitive Modeling Findings

Recall that the list of scored item attributes was generated based on the existing literature regarding reading comprehension items and previous item difficulty modeling studies of similar items. Due to differences across state testing programs, for some attributes it was the case that scored attributes were present in only a small number of items. Further, it could also be the case that some scored attributes were only present in items coming from one of the three state testing systems. For example, when coding the attribute regarding *Number of Passages Associated with an Item*, only items from the Maine assessment were ever coded as being associated with more than one passage. Such attributes were not eliminated at this point in the study, regardless of frequency, unless they were not present in any item. Consideration of the number of items including a particular attribute was considered further when item attributes were entered into regression and LLTM analysis to explain variability in item difficulty.

Of most interest are the correlations between the item-attribute scores and the item difficulty for the target population. Table 5 shows these bivariate correlations for items from all assessments and their associated p-values. As can be seen, the attributes most strongly related to item difficulty were (a) the various *measures* of vocabulary difficulty, (b) the length of the key, (c) the relationship between the wording of the NI and the response options (i.e., verbatim versus paraphrased), and (d) attributes related to higher order thinking processes such as hypothesizing and inferencing. The attributes with the highest correlations were then entered into the LLTM analyses.

Table 5. Bivariate Correlations Between Attribute Codes and Item Difficulty for the Target Population, Including Items from all Testing Systems

Attribute	Correlation	*Sig.*
Key in Numeric Form	-0.09	0.41
Number of Words in the Key	-0.18	0.11
Word Frequency Index of the Key	0.08	0.48
LSA Cosine for the Key where NI is Highest	0.16	0.16
Number of Distractors with LSA Cosines Lower than Key (by 0.10)	0.04	0.75
LSA Cosine Value for the Key with the NI	0.10	0.40
LSA Cosine of Highest Distractor with the NI	0.00	0.97
LSA Cosine of Highest Distractor with the Key	0.02	0.85
Sequence (Order) of Item on Operational Test	-0.20	0.08
Number of Passages	-0.20	0.07
Separated/Scattered NI	0.08	0.50
Number of Words in Passage	0.14	0.22
Number of Characters in Passage	0.14	0.20
Number of Words in Pre-Passage Material	0.01	0.93
Number of Characters in Pre-Passage Material	0.03	0.79
Title Included in Passage	0.71	0.00
Figure or Picture in Passage	0.38	0.00
Gunning Fog Index of Passage	-0.58	0.00
Flesch Reading Ease Score	0.47	0.00
Flesch-Kincaid Reading Grade Level	-0.53	0.00
Flesch Reading Ease Score of the NI	0.29	0.01
Flesch-Kincaid Reading Grade Level of the NI	-0.28	0.01
Type of Text—Genre	-0.01	0.93
Text is Persuasive	0.01	0.94
Inference is Required	-0.03	0.77
Item asks about the purpose of author's use of style, passage structure, or specific words.	-0.12	0.30
Item asks the examinee about a hypothetical or to suppose something.	0.26	0.02
Item asks the examinee to evaluate something in terms of a summary of the text.	0.03	0.81
Item asks about the author's main purpose or the passage's main idea.	0.05	0.63
Item requires knowledge of vocabulary with little context from the sentence.	0.23	0.04
Item requires paraphrasing of direct information in the passage.	-0.19	0.09
Item requires verbatim or close to verbatim information.	0.05	0.64
Item requires examinee to order information in a different sequence from that presented.	0.24	0.03
Item requires knowledge of vocabulary with good context from sentence.	0.08	0.49
Key requires knowledge of challenging vocabulary.	-0.15	0.18
Item requires the use of a figure or illustration.	0.03	0.80

Table 5. Bivariate Correlations Between Attribute Codes and Item Difficulty for the Target Population, Including Items from all Testing Systems (continued)

Item requires the examinee to understand the gist or tone of the passage, without referring to specific text.	0.04	0.71
Item is structured to include the word "EXCEPT" for the multiple choice format.	-0.08	0.50
Item asks about knowledge of text structure or language tools.	-0.13	0.24
Item uses or compares information from two passages.	-0.17	0.14
Item asks examinee to infer missing information.	-0.07	0.51
Item requires examinee to suppress information from other parts of the text in order to answer correctly.	0.08	0.47
Item requires examinee to select information from the text to support an argument.	-0.02	0.89
Length of Passage (Long vs. Short)	0.19	0.21
Higher order thinking Item	0.18	0.10

Three separate LLTM models were developed, one for each state, including the strongest predictors identified with the correlational analysis. In terms of overall model fit, the multiple R values between LLTM estimated difficulty and 1-PL estimated difficulty for the three states were quite high: .90 for Montana, .94 for NECAP, and .80 for Maine. Though the significant predictors differed slightly across each assessment system, the goal was to identify common attributes that accounted for significant amounts of variance in item difficulty across all three tests. This would allow a set of three possible modifications for the enhanced assessment to be made consistently across all items, regardless of the test from which they were derived. Based on LLTM parameter estimates, several candidate attributes emerged: (1) the vocabulary level of the NI; (2) the length of the key; (3) the vocabulary level of the key; (4) the LSA of the key—that is, the degree of lexical similarity between the key and the NI; and (5) whether the NI is located all in one place versus being spread out throughout the text.

The high proportions of explained variance in Rasch item difficulties by the IDM model features are impressive relative to previous IDM studies of similar items with general populations of students. The nature of the significant model predictors is of particular interest. Interestingly, the same variables that determine item difficulty for general education students appear to be relevant for the target population, suggesting that the cognitive processes applied by the two populations of students are similar—response decision processes and location of necessary information needed to answer test questions.

In terms of implications for item modifications, two general types of modifications were identified by the results: (1) changes to the linguistic properties of the items, key, and distractors; and (2) grouping/location of the NI in the text relative to the question. The expectation is that manipulations of these features of items, specifically reductions in the linguistic complexity and overlap of text, items, key, and distractors could decrease the difficulty of items. The question

that cannot be addressed by any of the presented data is whether such manipulations (a) change the construct so significantly that items are no longer at grade level, which would violate the NCLB legislation governing the assessment design, and (b) are merely principles of universal test design. Regarding the second point, it is not clear whether reducing the effect of decision processes that are sensitive to problem solving strategies (i.e., matching strategies) constitute construct irrelevant variance for all test takers. If so, then item redesign to increase the impact of text representation processes (i.e., encoding and coherence) could be worthwhile for all assessments, including those designed for general education students. Empirical examination of the effects of the recommended item manipulations for general education *and* special education students should provide some insight into these issues.

Exploratory Cognitive Interview Findings

Using transcripts from the exploratory interviews, item-level coding was completed for all 34 items in the four passages selected for manipulation. Item summaries were developed that provided details about each item, including specific words or structures that were challenging, response patterns that emerged, and cognitive processes that were identified by the students. We describe below four specific aspects of items that were identified to be addressed in the subsequent item manipulations.

Vocabulary

Complex vocabulary presented a challenge to students in those items that specifically measured vocabulary, as well as in many other items. In some cases, students latch onto a complex word that they do not understand, assume that it must be important, and choose an answer based on that word, even without knowing its meaning, especially if they feel that they can eliminate other choices with more accessible vocabulary. In other cases, students immediately discard any answer choice that includes a vocabulary word they do not understand. In either case, the complex vocabulary presents a barrier to students and makes it more difficult for them to identify the correct answer. Even without changing vocabulary within the passages themselves, the cognitive interview analysis suggested that simplifying complex vocabulary words in the items could provide greater access to students, and allow them to focus on the question's meaning, rather than to employ a blanket strategy of either choosing or rejecting all answers with complex vocabulary words.

Structure of Stem

There were some items that students found difficult to understand because of the structure of the question. One item was phrased in the negative: "According to this article, all of these factors contributed to the decline of totem pole carving **except** the…" Five of the eight students who attempted this item misunderstood the question and looked for a factor that did contribute to the decline rather than a factor that did not. Other items were written as open-ended sen-

tences rather than as questions, and some students commented that they did not understand what the question was asking, because it did not have a question mark.

Links Between Distractors and Passage

For a number of items, students were 'tricked' into choosing an incorrect answer because they found an idea in the answer that was related to content in the passage. Even if they could access the vocabulary, and understood what the question was asking, they were literally distracted by a word or phrase in an incorrect answer that seemed more closely linked to the passage than the correct answer.

Formatting and Scaffolding

As noted earlier, when designing the cognitive interviews for this study, the researchers considered research about cognitive interviews and think alouds, and in particular the limited research on using cognitive interviews with students with disabilities. Previous research noted the difficulty of having students with disabilities grasp the meta-cognitive aspects of thinking aloud, and the potential that they would not speak out loud their thinking processes because they do not recognize, for example, reading itself as a thinking process. Thus, the researchers deliberately chose to use the cognitive interview model rather than the cognitive lab model. The major difference between the two is that in cognitive interviews, the researchers engage more directly with the students during their think aloud. In contrast, cognitive labs attempt to remove the researcher as much as possible, creating a "laboratory" environment rather than an interchange between researcher and student. Cognitive interviews included questions such as "Why did you choose 'a' rather than any of the other choices?" For a number of students, when they were asked this question, it prompted a new thinking process. They may have chosen "a" without even considering the other options, and when asked the question, they return to the item and consider those options. The question itself served as a scaffold, to prompt the student to engage again with the item. In a number of instances, the resulting cognitive processing led the student to better understand the meaning of the item and to choose a different answer. This method resulted in a number of examples where the probing questions themselves became sources of scaffolding for the student, allowing them to engage differently with the item. The analysis of the scaffolding process raised the possibility that students might benefit from smaller portions of each passage with questions in between. This could address short-term memory issues and also serve to scaffold knowledge for students, as well as minimizing the physical distance between the necessary information in the passage and the item. The conversations between the student and the interviewer helped to scaffold the distance between the two, leading to the hypothesis that placing items closer to the necessary information might provide a natural scaffold without having to develop new items.

The scaffolding finding led to two other considerations which were not included in the current manipulations, but which could be considered in the future: first, the cognitive interview proto-

col itself can be adapted for instructional purposes so that teachers can provide the scaffolding questions for students; second, scaffolding items could be inserted in the assessment. Because this study specifically chose only to manipulate existing items and not create new items, it was not appropriate to create new items for this study.

Item Modifications Based on Cognitive Modeling and Cognitive Interviews

To complete the first part of the study, the results from the item difficulty modeling based on coding and analysis of cognitive features of items and difficulty parameter estimates, the cognitive interviews to determine the underlying processes involved in students' encoding and responding to the test items, as well as an analysis of performance data were triangulated to identify specific manipulations of assessment items that might improve access for students with disabilities. The results of these analyses were synthesized by content experts and researchers to develop specifications for item and test alterations and to identify those reading passages and items most amenable to manipulations. Two types of item manipulations were selected that were hypothesized to affect the cognitive load required of students when solving the items: linguistic modifications and formatting modifications. The desired effect was a reduction in the cognitive load of item processing, specifically reducing the processing requirements of non-construct related effects such as decision processes related to the multiple-choice item format or surface level features of items and passages. Face-to-face meetings were held with researchers and item developers, and each item was discussed in detail. Consensus was reached on the specific cognitive processes used in each item, and categories of specific manipulations that could improve access by lowering the cognitive load for those processes were identified and considered. Item developers then took all the information and produced the final manipulations: 34 items from four reading passages.

Confirmatory Phase

The effects of linguistic and formatting manipulations to passages and items were examined individually as well as in combination. The following research questions were addressed:

1. What impact, if any, does reduced cognitive load due to formatting have on item difficulty compared to item difficulty in the original construction?

2. What impact, if any, does reduced cognitive load due to linguistic modifications (LSA) have on item difficulty compared to item difficulty in the original construction?

3. What impact, if any, does reduced cognitive load due to formatting and LSA have on difficulty compared to item difficulty in the original construction?

The confirmatory cognitive interviews used the same manipulated items to address two research questions:

1. Do item modifications minimize the identified cognitive barriers for students with disabilities?

2. Looking at the characteristics of the target population of study (all students with disabilities who did not reach proficiency on their high school reading assessment), how can we develop criteria for identifying students eligible for AA-MAS?

Confirmatory Phase Research Design

Pilot Study

Sample

The sample of students participating in the pilot study was selected based on the following criteria: (a) the student must have an identified disability, (b) the student must have participated in the previous year's general assessment with or without accommodations, and (c) the student must have performed no higher than 1 standard error above the "proficient" cut score on the general state standardized assessment the previous year.

A total of 1,063 test booklets were distributed to a sample of students enrolled in schools across the five states participating in the three participating state testing systems. Of these booklets, data for several students were considered "missing" if none of the items were completed or if all of the items associated with a particular passage were omitted. This resulted in a final sample size of 1,051 student response strings.

Design and Materials

A total of four passages with 34 associated multiple-choice items were selected for the pilot study. These passages and associated items were selected in order to represent items and passages from each of the state testing systems (to the extent possible) and to maximize the effect size for the potential manipulation of items from the general assessment. The titles of the four passages included a) History of Blues (13 items), b) Totem Pole Carvers (11 items), c) Wrappings (5 items), and d) Taking the "Bait" Out of Rebates (5 items)[2]. Sixteen forms were created by fully crossing the item manipulations and passage order with the four passages.

[2] The passage titles will be referred to as Blues (History of Blues), Totem (Totem Pole Carvers), Wrappings (Wrappings), and Rebates (Taking the "Bait" Out of Rebates).

Experimental Manipulations

For each category of manipulations one or more design features could have been manipulated depending on the initial structure and content of the item. In all cases, the text in the passages remained unchanged in terms of wording. That is, no changes to the passages were made for the linguistic manipulations. The category of LSA manipulations included the following possible changes to item design: (1) simplification of the language and vocabulary of the stem; (2) simplification of the language and vocabulary of the key; (3) simplification of the language and vocabulary of the distractors; (4) closing the stem to change its structure from cloze to a formal question; (5) shortening or lengthening of the stem to make the question more clear; and (6) restructuring of the stem to make the question more clear. For the formatting manipulations, passages were altered in one of three ways: (1) sections of text were separated into multiple "chunks" of text, each of which was followed by a subset of the test questions pertaining only to that specific section of text; (2) specific words in the text were bolded to highlight or emphasize a word relevant to answering a question; or (3) pre-passage or post-passage text was added or modified.

Analysis

Though a series of analyses were conducted to examine the effects of item manipulations on item difficulty, total scores, and internal consistency, we focus our presentation here on the central issue of changes in item difficulty for the 2% population. For the analysis of item difficulty, we begin by presenting the individual item *P*-values followed by the summary descriptive statistics for *P*-values by passage. We then present results from two-way repeated measures ANOVAs to test the effects of item manipulations on item difficulty. For each ANOVA, the *P*-values of all items under each manipulation are compared to the *P*-values for the No Manipulation condition.

Next, the effects of several specific item manipulations (e.g., closing the stem of a question, bolding words) on item difficulty are tested with two-way repeated measures ANOVAs. The repeated measures factor is the repeated administration of items in the No Manipulation condition and then again in the manipulation condition. The between subjects factor compares items that involved a specific change as part of the manipulation (e.g., changing the vocabulary of the key) to items that did not receive that particular modification. It is hoped that this finer-grained analysis of item manipulations may reveal the causes of differences observed for the broader categories of manipulations (e.g., linguistic manipulations and formatting changes).

Confirmatory Cognitive Interviews

Sample

Students participating in the cognitive interviews had the following characteristics: an identified disability; did not reach proficiency on the high school reading assessment; and participated in the ART 2% pilot study. A total of 32 students participated in the confirmatory cognitive interviews. The exploratory and cognitive interview samples involved different students. The

selection criteria were almost identical for both sets of interviews, with the added criteria for the confirmatory interviews that the students had participated in the pilot assessment.

Data Collection

The confirmatory cognitive interview protocol used the 4 passages and 34 manipulated items. The interviews used the items that had both LSA and formatting changes (the no manipulation, LSA only, and formatting only versions were not used because of sample size limitations). Four forms were developed, with the passages in different orders, and randomly administered to the participating cognitive interview students. Between 9 and 20 students answered each item in the confirmatory interviews. The students typically sat for one 90-minute session in order to disrupt their schedules as little as possible, although a few chose to sit for two 45-minute sessions. This round did not include teacher interviews or video recording. As with the exploratory interviews, interviewers collected school transcripts, test scores, and IEP information on all students.

Similar to the first round, the interview involved a warm-up exercise to introduce the student to the think-aloud process, but the activity was modified: students were given a coloring activity and asked to describe all of the steps they took as they colored. They did this concurrently with the interviewer, which allowed a more informal interchange and modeling of the think aloud process.

Analysis

In the confirmatory interviews, the coding process began with the same set of codes used for the first round, which allowed a comparison between rounds. The defined coding categories for the first round were reading patterns, identifying important information, understanding item/distractors/key ideas, recall (positive and negative), decoding errors, executive processes (i.e. strategies for answering item), and student evaluation of items and distractors. In the confirmatory interviews, codes were also developed for each of the linguistic and formatting barriers identified in the first round and described above. The coded interviews were analyzed through two lenses: first, looking at all responses for particular items, and second, looking at student profiles to better understand student cognitive characteristics.

Confirmatory Phase Findings

Pilot Study

As can be seen in Table 6, the mean item difficulty was lowest (i.e., most difficult) for the items from the Maine assessment, followed by the NECAP and Montana items. It should be noted that across all assessment systems, the average percentage of target students passing an item never

reached 50%. Further, the item difficulties varied less for the Maine assessment versus the other two testing systems suggesting that the items were similarly difficult on this test.

Table 6. Item Descriptive Statistics, Including Mean Item Difficulty and Standard Deviations

Descriptive Statistics				
	State	**Mean**	**SD**	**N**
Item Difficulty for Target Students	Maine (Blues)	0.24	0.07	48
	Montana (Totem)	0.41	0.10	21
	NECAP (Wrappings and Rebate)	0.40	0.10	12
	Total	0.30	0.12	81

The average item difficulty for items associated with each passage was calculated under each condition. Table 7 shows the means and standard deviations of the *P*-values by passage and condition. On average, the items associated with the Blues passage (Maine) were more difficult than the mean difficulty of items with any other passage, regardless of manipulation. Though this is the most obvious difference, it is not of primary interest to the current study. The passage effect is an artifact of the difference in test content across the different state testing systems.

Table 7. Means and Standard Deviations of Item Difficulty by Condition and Passage

	Item Difficulty for No Manipulation	Item Difficulty for Formatting	Item Difficulty for LSA	Item Difficulty for Formatting and LSA
Blues (Maine)	.27(.08)	.29(.09)	.35(.13)	.36(.11)
Totem (Montana)	.48(.14)	.56(.14)	.53(.13)	.58(.13)
Wrappings (NECAP)	.51(.10)	.50(.23)	.52(.24)	.53(.20)
Rebate (NECAP)	.48(.04)	.50(.20)	.57(.18)	.60(.16)

Of more interest is the change in mean *P*-value that is observed across the different manipulation conditions. For each passage, the mean *P*-value of the items under the LSA and the LSA & Formatting conditions are consistently higher (i.e., easier) than in almost any other conditions, specifically as compared to the No Manipulation condition. To test this effect statistically, the results of several ANOVAs and ANCOVAs were interpreted.

Next, repeated measures ANOVAs testing the effect of manipulation and passage on item *P*-values comparing each manipulation separately to the No Manipulation condition were conducted. The manipulation effect in each ANOVA was statistically significant (see Table 8). The smallest, but still significant, manipulation effect was observed when comparing the Formatting Change condition to the No Manipulation condition, $F(1, 30) = 6.11$, $p = .02$, partial $\eta^2 = .17$. For the LSA condition, the LSA manipulation accounted for 29% of the variance in item difficulty, $F(1, 30) = 12.43$, $p = .001$, partial $\eta^2 = .29$. The manipulation effect was even stronger when comparing the average item difficulties under the LSA & Formatting condition to the No Manipulation

condition, $F(1, 30) = 21.55$, $p < .001$, partial $\eta^2 = .42$. Not surprisingly, the effect of passage on item difficulty was significant under all conditions, which likely resulted from the more difficult items associated with the Blues passage as compared to any of the other three passages.

Table 8. Repeated Measures ANOVA Results Comparing each Manipulation Separately to the No Manipulation Condition

	Two-tailed *p*-values		
	Manipulation Effect	**Passage Effect**	**Interaction Effect**
Formatting	0.02	<.001	0.04
LSA	0.001	<.001	0.38
Formatting & LSA	<.001	<.001	0.30

Finally, in order to understand the effect of specific manipulations to the items, seven different specific design feature manipulations were examined. Six LSA design features included *closing the stem, simplifying the language or vocabulary of the stem, simplifying the language or vocabulary of the options, changing the order of the options, editing attractive distractors*, and *expanding the stem*. One formatting design feature was examined—*bolding a word or sentence*. Under the Formatting condition, the interaction between the repeated measures factor (Formatting) and the between subjects (item) factor (Items with Bolding) was significant, $F(1, 32) = 10.11$, $p = .003$, partial $\eta^2 = .24$. In both the LSA and LSA & Formatting analyses, the interaction effects were only significant when the item design feature included closing the stem, $F_{LSA}(1, 32) = 7.21$, $p = .01$, partial $\eta^2 = .18$, $F_{LSA\ \&\ Formatting}(1, 32) = 9.77$, $p = .004$, partial $\eta^2 = .23$. Other significant effects included the main effects of the design feature changes associated with modifications to the response options. Analyses of the LSA manipulation condition resulted in significant main effects when linguistic or vocabulary simplifications were made to the response options $F(1, 32) = 10.02$, $p = .03$, partial $\eta^2 = .24$, or when attractive distractors with linguistic overlap to irrelevant portions of the text were changed, $F(1, 32) = 16.68$, $p < .001$, partial $\eta^2 = .34$. Under the LSA & Formatting change conditions, the same main effects were significant ($F_{Simplify\ Option\ Language}(1, 32) = 15.95$, $p < .001$, partial $\eta^2 = .33$; $F_{Edit\ Attractive\ Options}(1, 32) = 27.64$, $p < .001$, partial $\eta^2 = .46$), in addition to the effect of Bolding, $F(1, 32) = 26.73$, $p < .001$, partial $\eta^2 = .46$.

Confirmatory Cognitive Interviews

The cognitive interview analysis focused on identifying the ways in which students successfully or unsuccessfully answered the items, in particular looking at whether the students faced the same cognitive barriers as with the un-manipulated items from the first round (two different student populations were used for each round, so there was not a direct comparison of student responses). This section provides a detailed analysis of the 11 items from the Totem passage. For each of the items, we compared the number of correct responses in Round 1 vs. Round 2, we identified the specific manipulations for each item, and examined Round 2 student responses

in light of those manipulations. Four of the 11 items showed no clear evidence of changing student access to the item (two of these were vocabulary items). One item appeared to be more difficult than the original item. Six items showed some evidence of increasing access by reducing the cognitive load of the item: three items made the distractors less similar to the passage, two provided overall more clarity, and one changed from negative to positive question structure. Table 9 provides information about each of the Totem items based on the cognitive interviews. The first two columns show the number of students who answered the item correctly out of the total students who attempted the item. The next column lists the specific manipulations that were done to each item, and the final column provides a brief summary of the findings from the confirmatory cognitive interviews.

Table 9. Evidence of Impact of Manipulations for Totem Items

	Number of Students Answering Correctly/Total Attempting Item		Manipulations	Overall Evidence of Impact of Manipulations
	Round 1	Round 2		
Totem_1	5/8	12/18	Simplified vocabulary in distractors Option order changed Edited attractive distractors	No clear indications of overall impact.
Totem_2	4/6	17/18	Edited attractive distractors	Vocabulary item that might not benefit from manipulations—the cognitive load is not that complex and there were not many possible manipulations.
Totem_3	6/8	13/17	Simplified vocabulary in distractors Edited attractive distractors Bolding	Not clear if there are any changes in cognitive load from round 1 to round 2.
Totem_4	4/8	15/16	Simplified language/vocabulary in stem Simplified language/vocabulary-options Edited attractive distractors Expanded stem Added bolding	This may be an item where the manipulations did reduce the barriers by making the distractors less similar to the passage allowing students to demonstrate their understanding of the text.
Totem_5 (photo)	3/6	13/16	Key moved from A to B Minor changes to stem Edited distractor that was close to key Item moved from #10 to #5	The round 2 version seems to be clearer and allows students to demonstrate inference skills more than the round 1 version.

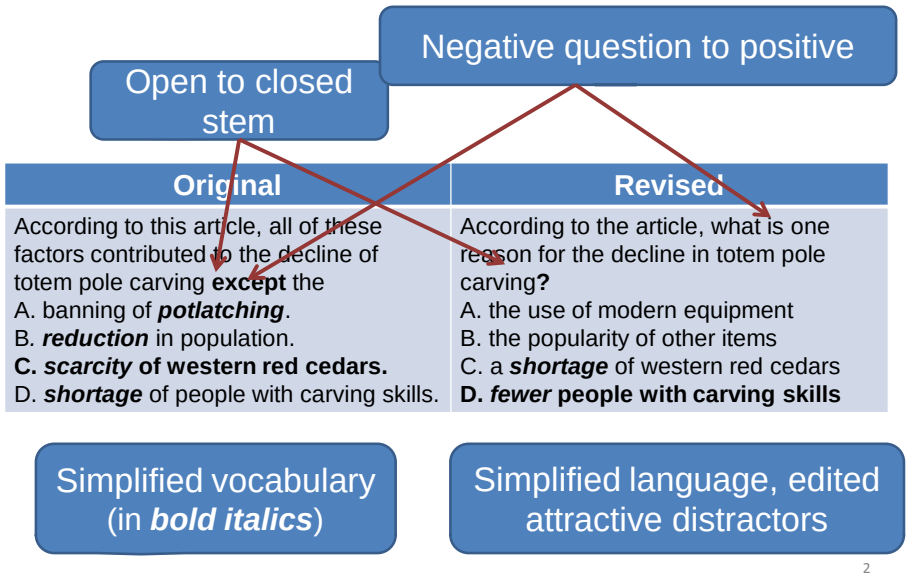
Table 9. Evidence of Impact of Manipulations for Totem Items (continued)

Totem_6 (#5 in round 1)	2/6	11/14	Changed open to closed stem Changed negative question to positive Changed vocabulary in distractors Simplified language Edited attractive distractors	The change from negative to positive seems to have increased access to this item.
Totem_7 (#6 in round 1)	2/7	3/12	Simplified language/vocabulary-options Option order changed Edited attractive distractors	Because the largest cognitive challenge was an essential component of the question (thread of continuity) it was difficult to reduce the cognitive load on this item in ways that could improve access. Eliminating some of the more attractive distractors may have allowed those students who did understand the passage and item to use their inference skills more successfully.
Totem_8 (#7 in round 1)	2/5	10/13	Edited stem Edited attractive distractors Simplified language/vocabulary-options Option order changed	Students were much more successful with this item than with the original item in round 1. In round 1 they did incorrect matching, were unable to find the important information in the passage, or misunderstood the distractors and the key.
Totem_9 (#8 in round 1)	6/7	10/13	Closed stem Edited attractive distractors	Vocabulary items might not be the best items for manipulation.
Totem_10 (#9 in round 1)	5/7	3/11	Closed stem Simplified language/vocabulary-options Edited attractive distractors	This item seems to be more difficult than the original one, based on the first round of cognitive interviews.
Totem_11 (#10 in round 1)	4/7	7/13	Option order changed Simplified language/vocabulary-options Edited attractive distractors	This item changed all of the distractors, and does seem to have chosen distractors that are more clearly fact than opinion. It seems to have increased access to the item.

Figure 1 provides an example of the revisions for Totem_6. Items had anywhere from one to five changes; this item has four. All four of the changes are in the LSA category; changing from an open to a closed stem question, making the question positive rather than negative, simplifying the vocabulary, and simplifying the overall language. Of the six students who attempted the original item in the exploratory cognitive interviews, two students answered correctly. In contrast, 11 of 14 students answered the revised item correctly in the confirmatory interviews. One student was even able to use inference with the manipulated item, a skill not seen in very many of the interviews: “The answer’s not like in it, but you can maybe read it carefully, you

can actually see what they're talking about, with fewer people with the skills, because the carvers were dying and weren't passing on their knowledge, and skills and stuff."

Figure 1. Original and Revised Item Totem_6



*Note: the correct response is bolded ('c' in the original, 'd' in the revised version.

Findings were similar for the other three passages. There were clearly items where student responses in the second round were similar to responses in the first round, indicating that the changes to the item had not had an impact on the item's cognitive load in a way that contributed to increasing student access to the item (such as items #2 and #9 in "Totem"). Those items that underwent a more significant change in cognitive load, such as the change in wording from negative to positive in Totem #6, showed a greater change in student access to the item from the exploratory to the confirmatory interviews. Finally, for a small number of items, it appears that the heaviest cognitive load is directly related to the construct being measured, such as Totem #7, which requires that students understand a concept from the passage that cannot be changed through item manipulation. For these items, the manipulations did not improve access for students. Looking across all the items in the study, there were no striking differences between the first and second rounds of the cognitive interviews, and often there was no clear evidence that students in the second round were using any of the changes in the item to increase their access to the essential content.

The Students

As previously described, students participating in the cognitive interviews had the following characteristics: an identified disability; did not reach proficiency on the high school reading assessment; and participated in the ART 2% pilot study. The 32 students who participated met each of the three criteria, but even so, they were quite diverse in terms of their test scores, their IEPs, and their performance on the pilot test and cognitive interviews. An examination of their state reading assessment scores, their performance on the ART 2% pilot, their scores on the items in the cognitive interviews, their IEPs, as well as their comments during the cognitive interview, demonstrated that there is great diversity within this broad population that, under the federal regulations, could potentially be eligible for the AA-MAS. The analysis of the students focused on identifying the characteristics which could indicate whether or not the student could benefit from an AA-MAS. However, it soon became clear that it would be very difficult to identify specific cognitive characteristics that could be used as indicators of eligibility. Instead, five different categories of students emerged, with one or more students fitting in each category (some students fit in more than one):

- *Students who are almost proficient:* students who demonstrate that they are performing close to proficiency, for whom the current assessment provides a valid measure of what they know and can do, and for whom an AA-MAS would not provide a more valid measure.

- *Students who lack opportunity to learn:* students who do not seem to have had sufficient access to the curriculum to challenge them at their level and provide an opportunity to learn to their full potential.

- *Students who are far below grade level:* students whose reading and comprehension level is too far below grade level to be able to access any kind of grade level assessment, even if modified like the AA-MAS.

- *Students who do not demonstrate their cognitive processing:* students who provided little or no evidence of their cognitive processes.

- *Potentially eligible students:* students for whom the manipulated assessment or the cognitive interview format provided greater access.

While these categories were developed based on all 32 cognitive interviews, this section provides specific information about eight of those students, describing how they exemplify the category in which they were placed. Table 10 provides a brief summary of each of the eight students. The names of all students have been changed, and they are referred to by pseudonyms in this chapter.

Table 10. Focus on Student Characteristics

	Pseudonym	Disability	Least Restrictive Environment	Opportunity to Learn (assigned ELA class)	State Score Level	Pilot Score (of 34)	Pilot % Correct	Cognitive Interview Score	Cognitive Interview % Correct
Almost proficient	Stewart	SLD	In general education 80% or more	Standard/ Honors	Almost proficient	13	38%	16/18	89%
	Alice	SLD	In general education 80% or more	Standard	Almost proficient	9	26%	9/19	47%
Lacks opportunity to learn	Richard	SLD	In general education <40%	Special education	Almost proficient	15	44%	9/16	56%
	Tom	ADD/ ADHD	In general education 40-79%	Special education	Very Low	21	62%	9/18	50%
Far below grade level	Emma	SLD	In general education 40-79%	Special education	Very Low	11	32%	1/9	11%
Does not demonstrate cognitive processing	David	ADD/ ADHD	In general education 80% or more	Special education	Mid-level	19	56%	13/18	72%
	Emma	SLD	In general education 40-79%	Special education	Very Low	11	32%	1/9	11%
	Isla	SLD	In general education 40-79%	Standard	Very Low	12	35%	9/19	47%
Potentially eligible for AA-MAS	David	ADD/ ADHD	In general education 80% or more	Special education	Mid-level	19	56%	13/18	72%
	Oliver	SLD	In general education 80% or more	Standard	Mid-level	6	18%	14/20	70%
	Tom	ADD/ ADHD	In general education 40-79%	Special education	Very Low	21	62%	9/18	50%

Almost Proficient

Students who are almost proficient demonstrate relatively high levels of cognitive processing, and, while they have not reached proficiency on the reading assessment, are demonstrating academic success in other areas, and there is evidence that they are very close to proficient. Stewart (all names are pseudonyms) is a good example of a student who demonstrated high levels of cognitive abilities and awareness of his cognitive processes. He has a specific learning disability that affects his writing abilities. He is in regular education at least 80% of the time, with just over four hours a week spent in special education. He scored "almost proficient" on his state assessment test. Stewart was originally placed in a high-level English class until his writing disability made it difficult for him to complete his work, and he completed AP U.S. History. His almost proficient score and his articulateness in the cognitive interview indicate that he does not need major changes to an assessment to demonstrate his proficiency – the assessment may be an accurate measure of his almost proficient reading level.

Lacks Opportunity to Learn

While all students must be given sufficient opportunity to learn by having access to a grade-level curriculum, that is not always the case for every student, particularly students with disabilities. Among the eight profiled students, two in particular demonstrated large gaps between the different assessment scores and the reading level of their classes. Tom, who was described in the table above as a student who may benefit from an AA-MAS because of his strong scores on the cognitive interview in comparison to the state assessment, is one of those students. His low performance on the state assessment matches the low level of reading classes (special education or basic), but does not match the passion with which he described graphic novels during his interview, nor his success during the cognitive interview. Richard is another student with discrepancies that may indicate a need to focus on his opportunity to learn rather than on the assessment. Richard's specific learning disability interferes with his ability to communicate his level of comprehension. He scored within the Almost Proficient range on his state reading assessment, got 44% correct on the pilot assessment, and answered more than half of the questions correctly during the cognitive interview. Despite this relatively strong performance in three different assessment settings, Richard has been placed in almost all special education classes and has a very low GPA. His record indicates that while an AA-MAS may help him, he may actually benefit more from a more challenging level of instruction that recognizes his ability to demonstrate almost proficient levels on assessments.

Far Below Grade Level

Emma is a good example of a student who participated in the cognitive interviews for whom there was no evidence that any of the item manipulations, or the cognitive interview process, were helpful to her. While she "attempted" 10 questions, she skipped six of them because she did not understand them at all, and of the four she did answer, she only got one correct. Her reading

and comprehension level is clearly far below grade level, and no assessment that attempts to measure grade level skills, even in a modified format, will be accessible to her.

Does Not Demonstrate Cognitive Processing

Not surprisingly, for many of the students with disabilities who participated in the cognitive interviews, the interview process itself presented a challenge. Even students who may benefit from the AA-MAS, such as David and Isla, found it very difficult to articulate their thinking processes during the interview. David did not communicate well throughout the interview, making it difficult to understand his cognitive processing and inference patterns. He often explained his train of thought by saying "It makes sense," or even "I don't know how I answered it." He did not respond to multiple prompting attempts and was generally disengaged throughout the entire process. He was easily annoyed when asked to explain his reasoning. Overall, it was difficult to determine much about David's cognitive processes or his knowledge, skills, and abilities from the interview, scores, and IEP information. Thus, there is very little concrete information about what aspects of the manipulated items, or of the cognitive interview itself, provided a benefit to him. Emma is another student who was unable to describe her cognitive processing. She is a twelfth-grader with a specific learning disability that mostly affects her behavior and emotional state in the classroom. Her IEP describes her as caring, friendly, and thoughtful, with a good sense of humor, though she faces severe emotional issues that disturb her concentration and engagement with her schoolwork. Throughout the cognitive interview Emma was easily distracted, easily frustrated, and a very slow reader. She also performed at the lowest level on the state assessment, corroborating her general low proficiency level. Her interview was difficult to analyze due to a lack of output, making her a good example of a student who does poorly with no clear explanation of the cognitive processes that are challenging for her, and thus no clear indication of what kinds of changes to assessment items would be helpful for her.

Potentially Eligible for AA-MAS

A number of students performed better on the cognitive interview than on the state assessment or provided evidence that one or more of the item changes provided them with access to the item. David is an eleventh-grade student who has difficulty completing his work. He is currently in a pullout resource English class that covers the same material as eleventh-grade English, but with increased time and modifications for assignments. Although he did not respond well to the cognitive interview process and the one-on-one setting did not appear to be beneficial for him, he demonstrated a higher level of reading comprehension on the cognitive interview (72% correct) and the pilot (56%), than would be expected given that he has been placed in special education classes for English, and his state assessment score was in the mid-level (below proficient) range. The changes to the assessment items may have provided greater access, although he was unable to articulate how. Similarly, Isla's state assessment score in reading was categorized as very low, but her cognitive interview and pilot scores are better than would be

expected given the state assessment score. It may be that some aspects of the pilot or interview eliminated some barriers for her.

Oliver is a 12th grader whose primary learning difficulties are in the areas of language arts, written expression, and grammar. He is motivated to learn and is employed part-time in a vocational setting. According to his IEP, Oliver has difficulty organizing his thoughts, daily work, quizzes, and tests; he also struggles with comprehension and reads slowly. He scored mid-level on the statewide reading assessment and according to his IEP is currently reading below grade level. At times during the cognitive interview he was quite articulate and communicated his level of understanding, using higher-order thinking and without needing to say "it fits" or "it makes sense" to substantiate his response. In contrast to his relatively strong performance on the cognitive interviews (13 out of 20), he scored at the mid-level below proficient on his state assessment and got only 18% correct on the pilot assessment. At the end of the session, he told the interviewer that he found that being asked "why" made him really think about his choice of answer and figure out his reasoning.

Tom had one of the largest differences between his state assessment score (very low), the cognitive interviews (62%), and the pilot (50%). In the interview, Tom described passions that demonstrate literacy skills but not in accepted school-based contexts (i.e. graphic novels), and throughout high school he has been placed in special education or very basic English classes. His variation in performance indicates that there may be a gap between what Tom knows and can do and his state assessment score. Given his higher scores on the pilot and relative success with the cognitive interview, Tom could be a candidate for an AA-MAS.

In summary, none of these categories is a clear demonstration of suitability for an AA-MAS. The in-depth profiles provide information about the different kinds of students with disabilities who are not reaching proficiency on the state reading assessment, but they do not help to create a checklist of specific characteristics for AA-MAS eligibility. The students in the cognitive interviews, even though they were a very small sample, presented such a diversity of abilities and disabilities that it was not possible to generalize to the assessment eligibility.

Discussion

The item pilot study was designed to test the effects of specific item manipulations derived from cognitive modeling and cognitive interviews on student performance on a manipulated version of a general education assessment of reading comprehension. After a large scale pilot test of 20 forms of the manipulated items across five states participating in three state testing systems, several general conclusions were made. The most relevant conclusion to be made regarding the manipulations of items intended to reduce cognitive load is that linguistic modifications

to items, either in conjunction with or without modifications to item formatting, had the most positive effects on student scores and item difficulty.

Regarding conclusions about specific manipulations of design features (e.g., simplifying language or vocabulary in the stem, bolding) the distribution of these changes across items limits our ability to conduct meaningful analysis in most cases. Aside from closing the stem of items, all of the other manipulations were either present for as few as 1 or 2 items or as many as all of the items. Thus, sufficiently large sets of items to allow for group comparisons were not available. The exception is the closing of stem design feature manipulation. Items for which stems were originally open-ended that were then closed to form a question as part of the LSA or Formatting & LSA manipulation became easier for the population that took the assessment. It is hypothesized that by framing the item as a true question the students better understand what is being asked of them. This reduces cognitive requirements associated with interpretation of the question and allows students to focus on using the text to answer the question based on their text comprehension. Some support for this hypothesis was provided from the cognitive interviews, but further testing of this hypothesis with carefully designed studies could offer more persuasive evidence.

What did seem to make a difference in the effect of manipulations was the specific passage under investigation. It was clear that items from some passages were (a) more easily modified by the item developers, and (b) more sensitive to modifications in terms of changes in item difficulty. If states pursue the approach to enhanced assessment design that modifies existing general assessments, they should consider the possible limitations carefully. Some passages and items can be better modified to suit the needs of these students. Clearly this approach will be limited by the passages and items initially selected for the general assessment—a process that is not likely to incorporate considerations regarding item modification for the non-general student population. Finally, given that the 2% student population may require more cognitive effort to complete traditional assessment item formats (e.g., multiple choice items) than the general population, shorter testing times with fewer items over multiple testing sessions may benefit these students.

While this study focused on item manipulations that could benefit the eligible 2% target population, and did not include students without disabilities in its sample, we found that some of the manipulations may follow principles of universal design for reading test items, and could be considered for all students:

- The linguistic manipulations intended to reduce the effects of non-construct relevant variance may constitute general good testing practices. Making a question clearer is more helpful to all students. This may seem obvious, but language or ambiguity regarding what the question is actually asking should not get in the way of the skill that is being tested.

- It is important for students to understand what is being asked of them. Students respond better to a question with a question mark rather than being asked to complete sentences.

- Keeping in mind grade-level appropriate vocabulary, vocabulary and syntax can be made more accessible to all students.

- The correct answer should be clear and defensible. Distractors should not be too close to the correct answer.

The study results also provided more general implications for assessment development:

- Assessment developers must acknowledge the difficulty of assessing all students with a single assessment and address that difficulty in assessment design.

- Human contact is critical for some students to demonstrate their knowledge and skills.

- Subsequent reviews by content experts as well as an internal and an external alignment study suggested that item manipulations did not significantly alter the reading comprehension construct (or grade-level standard) intending to be assessed. This is important because it supports claims that items can be modified in ways that enable student access without violating federal guidelines requiring states' AA-MAS to measure grade-level content.

- Cognitive modeling of assessment data can help to provide direction for writing or modifying items for use with specific populations. The use of empirically based data on sources of item difficulty provide principled approaches to cognitively-based item design for AA-MAS.

The student profiles demonstrate the heterogeneity of the population in the sample, and the difficulty of identifying specific criteria for a population appropriate for the AA-MAS. The categories of students identified—(1) students who are almost proficient and do not need an AA-MAS; (2) students who lack opportunity to learn; (3) students who are far below grade level; (4) students who do not demonstrate their cognitive processing; and (5) students potentially eligible for an AA-MAS—demonstrate that students with disabilities who do not reach proficiency on the state assessment are a heterogeneous group with many different needs that may not be able to be met by a single assessment, with or without manipulations. Some of the students found that the cognitive interview process itself increased their access to the assessment; think-aloud can be a successful teaching strategy, particularly for students who find meta-cognitive processing to be challenging. One-on-one contact helps this process. The study confirmed that there are students who are not being measured well in the current assessment system, and that it is difficult to develop specific criteria to identify those students, in part because for many students the issue may lie more with their access to the curriculum itself and their opportunity to learn rather than the design of the assessment system.

Conclusions

As described here, the specific manipulations used for this study were not found to be a sufficiently effective way to improve access to the reading comprehension test for the target population. Modifying existing tests may be too constrained for this population. Instead, different item types or assessment strategies may be needed. At the item level, future studies could unbundle the item manipulation strategies and determine which students benefit from specific manipulations. More research is needed to understand how specific manipulations improve access for specific types of disabilities. At the assessment level, the study results suggest that a single paper and pencil assessment, even with item-level manipulations, does not provide all students the opportunity to show what they know and can do. Other alternatives should be considered, among them adaptive testing or using more than one assessment for accountability. As growth models grow in popularity, research is needed on the expectations of growth for this population, and a recognition that the item bank may expand beyond grade level. At the instructional level, content standards matter. There needs to be careful evaluation of the constructs being measured in the assessment and care needs to be taken that assessment limitations do not have an impact on what is taught. Finally, at the student level, the study identified five categories of students who initially met the study criteria of having a disability and not reaching proficiency on the state reading assessment, but only one of those categories would benefit from an AA-MAS, and there was no set of objective criteria that could differentiate those students from the others. In addition, the categories indicate that much more is needed than just an AA-MAS; some students need increased access to high-quality instruction and the opportunity to learn, some need an assessment that will provide greater discrimination at their lower academic level. For all of the students, the traditional single assessment may be a poor measure, and other options deserve further scrutiny.

References

Alavi, S. M. (2005). On the adequacy of verbal protocols in examining an underlying construct of a test. *Studies in Educational Evaluation, 31*(1), 1-26.

Alderson, J. C. (1990). Testing reading comprehension skills: Part 2. Getting students to talk about taking a reading test (A pilot study). *Reading in a Foreign Language, 6,* 425 – 438.

Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (2009). *White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS).* Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Anderson, R. C. (1982). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42,* 145-170.

Bechard, S., & Godin, K. (2007). Identifying and describing students in the gap in large-scale assessment systems. In New England Compact (Ed.), *Reaching students in the gap: A study of assessment gaps, students in those gaps, and assessment alternatives to lessen the gap*. Newton, MA: Education Development Center, Inc.

Boring, E. G. (1953). A history of introspection. *Psychological Bulletin, 50,* 169-189.

Capraro, M. M., & Joffrion, H. (2006). Algebraic equations: Can middle-school students meaningfully translate from words to mathematical symbols? *Reading Psychology, 27*(2/3), 147-164.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16,* 486-514.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50,* (3), 328-344.

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343 – 368.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension. *Applied Psychological Measurement, 11*(2), 175-193.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Revised edition)*. Cambridge, MA: MIT Press.

Faerch, C., & Kasper, G. (Eds) (1987). *Introspection in second language research.* Clevedon, UK: Multilingual Matters.

Ferrara, S., Duncan, T. G., Freed, R., Vélez-Paschke, A., McGivern, J., Mushlin, S., Mattessich, A., Rogers, A., & Westphalen, K. (2004). *Examining test score validity by examining item construct validity: Preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment*. Paper presented at the Annual Meeting of the American Educational Research Association.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Freedle, R., & Kostin, I. (1992). *The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: Main ideas, inferences, and explicit statements* (ETS Research Report RR 91 – 59). Princeton, NJ: ETS.

Freedle, R., & Kostin, I. (1993).The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10*, 133 – 170.

Gorin, J. S. (2005). Manipulation of processing difficulty on reading comprehension test questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*, 351-373.

Gorin, J. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, Winter 2006*, 21-35.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 3*(5), 394 – 411.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers. 36*(2), 193 – 202.

Graves, M. (1986). Vocabulary learning and instruction. In E. Rothkopf (Ed.), *Review of Research in Education* (vol 13, pp. 49-89). Washington, D.C.: American Educational Research Association.

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.

King, T. C., & Laitusis, C. C. (2008). *Sample cognitive interview protocol*. Princeton, NJ: Educational Testing Service.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163-182. doi: 10.1037/0033-295X.95.2.163

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist. 49*, 294-303.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Oxford: Cambridge.

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5*, 257-274.

Kintsch, W., & vanDijk, A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Landauer, T. K. (1998). Learning and representing verbal meaning: The latent semantic analysis theory. *Current Directions in Psychological Science, 7*(5), 161-164.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 359-384.

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*(4), 6-15.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3-16.

Mann, S. J. (1982). Verbal report as data: A focus on retrospection. In S. Dingwall & S.J. Mann (Eds.), *Methods and problems in doing applied linguistic research* (pp. 87-104). Lancaster, UK: University of Lancaster, Department of Linguistic and Modern Languages.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Olson, G., Duffy, S. A., & Mack, R. I. (1984). Thinking-out aloud as a method for studying real time comprehension processes. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 253-286). Hillsdale, NJ: Erlbaum.

Parker, C. E., & Saxon, S. (2007). "They come to the test and there is nothing to fold:" Teacher views of large scale assessments and classroom context. In New England Compact (Ed.), *Reaching students in the gap: A study of assessment gaps, students in those gaps, and assessment alternatives to lessen the gap*. Newton, MA: Education Development Center, Inc.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Earlbaum.

Ransdell, S. (1995). Generating thinking-aloud protocols: Impact on the narrative writing of college students. *The American Journal of Psychology*, *108*(1), 89-98.

Seliger, H. W. (1983). The language learner as linguist: Of metaphor and realities. *Applied Linguistics, 4,* 179-191.

Sheehan, K. M., & Ginther, A. (2001). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section.* Paper presented at the 2000 Annual Meeting of the National Council of Measurement in Education.

Sheehan, K., Kostin, I., & Persky, H. (2006). *Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying performance on the NAEP grade 8 reading assessment.* Paper presented at the 2006 Annual Meeting of the National Council of Measurement in Education.

Title I—Improving the Academic Achievement of the Disadvantaged; Individuals With Disabilities Education Act (IDEA). Final rule. 72 Fed. Reg. 17748–17781, pts. 200 and 300 (2007, April 9).

Winter, P. C., Kopriva, R. J., Chen, C. S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning & Individual Differences, 16*(4), 267-276.