Who are the 2% students and how do we design test items and assessments that provide greater access for them? Results from a pilot study with Georgia students

Karin Hess, Patricia McDivitt, & Melissa Fincher, 2008

Introduction

State departments of education are now provided an opportunity to administer assessments designed to maximize the use of the 2% proxy flexibility rule for purposes of meeting Adequate Yearly Progress (AYP). This assessment option is allowed through new regulations regarding No Child left behind (NCLB) Act, which provides for the development and implementation of an alternate assessment based on modified academic achievement standards (proficiency or performance standards) for students. This paper describes the methods and tools for Georgia's empirically-based approach to consider design issues for the development of the "2%" alternate assessment based on modified academic achievement standards. Extensive analyses defining the student population; analyses using both innovative quantitative and judgmental approaches for determining the types of items most likely to "work;" item revision workshops with teachers to create more accessible test items; and item piloting provided the framework for the project.

Documentation of the general procedures and analysis tools used, including definitions of the terms "revised items" and "enhanced items," distinguishing between chunked reading passages and segmented passages, and various "scaffolding strategies" are clarified and illustrated with examples.

Purpose and Goals for the Project

Georgia's Enhanced Assessment Grant (EAG) Goals:

1. To understand who the lowest performing students are
2. To understand the achievement of this population of students
3. To evaluate Georgia's current assessments in light of what is learned about these students
4. To pilot and evaluate additional access methods through revised and enhanced test items

Why take this approach? Georgia took a long view – and a systemic approach believing that in order to build a technically sound assessment, it is imperative to understand who the target population is and what the students in this target population know and can do. Therefore, several methods were used to better understand this population of students and their specific learning needs *before* selecting items for an operational assessment. The goal was to target the assessment toward the students' strengths and provide scaffolding to help them improve upon their weaknesses.

Defining the population of students and their potential barriers

As described in the NCLB Modified Academic Achievement Standards Non-Regulatory Guidance (Draft) document (2007), an alternate assessment based on modified academic achievement standards is for all students who are covered under the Individuals with Disabilities Education Act (IDEA), whose disability has prevented them from achieving grade-level proficiency on the general education assessment and who will not reach

grade-level achievement in the same timeframe as other students, even with significant instructional interventions. Furthermore, this sub-group of students does not meet the participation criteria as defined by a given state's alternate assessment for students with significant cognitive disabilities. This sub-group (2% of the population) typically comprises a population of students who have a wide range of diverse skills and abilities, but who must have an IEP. In addition to having an IEP, they must also be receiving instruction in grade-level general education content. Some of these students may be able to take the state's general education assessment for one content area, but they may need to be assessed on the alternate assessment based on modified academic achievement standards in another content area. For example, a student in grade 5 is succeeding in mathematics and may take the general education assessment; however, in reading and language arts that same student may need to be assessed on the alternate assessment.

Since there is likely more than 2% of the students who meet this definition, each state must clearly articulate the targeted population. For the alternate assessment based on modified achievement standards, some states are targeting students whose performance is just above those students with significant cognitive disabilities. Other states are targeting students with IEPs whose performance puts them much closer to their grade-level peers. Understanding the targeted population should have a major influence on the development of the assessment. As a result, the first step was to define the population and provide a rationale for that definition. For example, some states are only developing alternate assessments based on modified academic achievement standards beginning at grade 5, because they are including in their definition that a student must have scored at the lowest performance level on the general education assessment for two years in a row. These persistently low performers may be encountering barriers to demonstrating proficiency on the general assessment, even with appropriate accommodations. For these states, the alternate assessment based on modified academic achievement standards may be developed with the goal of removing these barriers (Perie, Hess, & Gong, 2008).

In defining the population, the state of Georgia mined its assessment data (from 2004, 2005, and 2006) to first identify and describe the population of students who were "persistently low performing" – those who scored at the lowest performance level on the general assessment three years in a row. While there were no clear patterns among this population, in terms of specific skill deficits, some general descriptors did emerge: approximately 40%-55% were students on IEPs; there were more African American students (60-65%); more students on free and reduced lunch (75-80%); more students with mild intellectual abilities (20-30%); and more males in this group (60-65%).

In addition to the 3-year assessment data review, Georgia teachers were also asked to describe the population from their perspective –as learners in their classrooms. This took place at the training sessions for item revisions. The training sessions with Georgia educators began with asking teachers three open-ended questions. The questions asked were as follows:

- Who are these students? How would you describe them - when they read or do mathematics in your classroom?
- What are their greatest struggles in reading or mathematics?

- What instructional strategies seem to best support their engagement and learning in reading or mathematics?

Prior to reviewing the results of the data mining, Georgia educators described persistently low performing students as passive learners, who were hesitant to take risks; they exhibited meta-cognitive deficits, such as the inability to generalize skills and concepts to new situations or problems, failing to make connections, having difficulty changing topics, and unable to readily access and apply strategies. The committee also mentioned that these students tend to use limited vocabulary and prior knowledge; have poor decoding, fluency, and comprehension skills; and demonstrate poor number sense. Strategies employed in the classroom to assist students included guided practice; previewing words and questions; grouping, chunking or summarizing information; utilizing visual tools such as number lines, place value charts, manipulatives, graphic organizers, and multiple representations.

After identifying students who scored at the lowest performance level on the general assessment three years in a row, the state then examined the test items at two grade levels in reading and mathematics from two perspectives:

1) a technical review looked at items that this population of students was able to answer correctly and seemed correlated with overall performance and other items that did not appear to function well for these students; and
2) content experts simultaneously examined the same items to determine potential barriers to demonstrating proficiency on these test items.

Findings of the technical review were then triangulated in order to select items, and design and conduct a pilot study in February 2008, using "revised" and/or "enhanced" items designed to increase accessibility for the 2% population. Understanding the performance of this population on these items should help to both guide how to revise the items appropriately in the future and how to determine what "proficient" performance might mean for this population. (See Appendix D for a sample template for compiling these analyses.)

Applying a research base for revising and enhancing test items and test formats

Prior to any item revision work, a review of the literature, including empirical research on assessment formats for students with learning disabilities (e.g., Dickson, Simmons, & Kame'enui, 1995; Universal Design), factors that can influence overall reading comprehension (e.g., RAND Study Group, 2002) and text clarity (Hess, 2008), and mathematics learning (e.g., National Mathematics Advisory Panel, 2008) was used to identify potential strategies to guide teacher committees in revising or enhancing test items and considering alternative test formats. The purpose of the literature review was to identify potential factors and/or ideas that related to some of the characteristics of the target population – students with learning disabilities in reading and students with mild intellectual disabilities. The following table lists a brief summary of factors identified as having some potential for increasing accessibility for the target population. Educators and test developers working to maximize readability and accessibility of test items for this

population should examine the role of text complexity and text structure, test formats, and item constructs including some of the ideas below.

Using a Research Base for Revising & Enhancing Test Items and Test Formats Hess, McDivitt, & Fincher, 2008	
Factors to consider	What the research literature says…
Use of pronouns	Sentences with excessive pronoun use may cause a student to lose track of the main point of reference in an item.
Use of negatives	Difficulty of text may vary due to "complex Boolean expressions." Such expressions are challenging because "the respondent needs to keep track of different options and possibilities." In the case of negative expressions, an unnecessarily high cognitive loading may be added to items that employ negatives within items (e.g., "Which of the following is not a reason why the captain wanted to turn the ship around?").
Vocabulary load	There is an inverse correlation between the level of vocabulary in text and its readability. (For purposes of the study, unless the construct of an item was to test a particular vocabulary skill, it seemed reasonable that vocabulary demands could be reduced as a principle of access.) Vocabulary load can be a significant factor in accessibility of test items (Johnstone, C., Liu, K., Altman, J., & Thurlow, M. (2007).
Non-construct subject area language (specialized vocabulary)	Each content area has a specialized vocabulary of its own (e.g., characterization, denouement, alliteration). When these words are part of the intended construct, it is appropriate to include them; however, when these terms are extraneous to the intended construct, they may introduce "nuisance variables" (Haladyna & Downing, 2004).
3 "Tiers" of Vocabulary	Beck, McKeown & Kucan (2002) identify 3 tiers of a literate individual's vocabulary. Tier 1 includes most basic words (e.g., clock, baby) that rarely require direct instruction. Tier 2 includes high frequency words used by most "mature language users" found across most domains (e.g., absurd, fortunate, anxious). These words often have multiple or nuanced meanings - intended meaning that directly impacts verbal functioning and comprehension. Tier 3 words include low-frequency, domain-specific words (e.g., land forms, classifications of plants). These words are learned for a specific purpose or need.
Complex sentences, dense text	Sentences with "dense clauses" are sentences that "pack too many constituents or idea units (i.e., propositions) within a single clause"…and/or with "dense noun phrases" as sentences with "too many adjectives and adverbs modifying the head noun." Items with either one of the above sentence types may be problematic to reading comprehension (RAND, p. 96).
Text Structure	There are text structures that are easier to understand and use to organize and recall information (sequence, chronology, enumeration/description, definition, compare-contrast). More complex text structures require understanding of entire texts, such as cause-effect, problem-solution, and proposition-support (Hess, 2008).
Graphic organizers	Graphic organizers are effective instructional supports used during a lesson to assist students in understanding such things as the text structure (e.g., story map for narrative, timeline for chronology, Venn diagram for compare-contrast) (Schumm, 2006). The layout or format of a table, graph, or graphic organizer visually organizes information for conceptual understanding (Robb, Richek, & Spandel, 2002).
Chunking text	"Chunking" of texts is an effective instructional support used to assist students in conceptualizing ideas presented in longer texts, reducing the demand on working memory. Research related to the use of chunking has focused on chunking parts of sentences (phrases), rather than on chunking full texts. Overall comprehension may not be significantly improved by dividing sentences into smaller parts; however, students with reading disabilities have identified a preference for shorter chunks while reading (Dickson, Simmons, & Kame'enui, 1995).

Using a Research Base for Revising & Enhancing Test Items and Test Formats (continued)	
Legible Text (factors that affect a less threatening visual presentation of text)	
Segmenting text	"Segmenting" text is a term generally used as it applies to breaking text into subparts to reduce the length and visual appearance of the reading load. Segmented texts group segments with their corresponding test items. Recent studies have examined the impact of segmenting on students' non-cognitive domains such as anxiety, fatigue, frustration, and motivation as well as on their comprehension. Results suggest segmented text may be more accessible to students with disabilities; building in short "breaks" for the reader may also reduce the need for added accommodations (Abedi, Leo, Kao, 2008).
Spacing (the amount of space between each character)	Letters that are too close together are difficult for partially sighted readers. Spacing needs to be wide between both letters and words (Gaster & Clark, 1995). And we assume also for numbers and symbols.
Typeface (characters, punctuation, and symbols that share a common design)	Italic is far less legible and is read considerably more slowly than regular lower case (Worden, 1991). Boldface is more visible than lower case if a change from the norm is needed (Hartley, 1985).
Text Justification (text is either flush with left or right margins – justified – or staggered/ragged – unjustified)	Staggered right margins are easier to see and scan than uniform or block style right justified margins (Arditi, 1999; Grise et al., 1982; Menlove & Hammond, 1998). Justified text is more difficult to read than unjustified text – especially for poor readers (Gregory & Poulton, 1970; Zachrisson, 1965). Justified text is also more disruptive for good readers (Muncer, Gorman, Gorman, & Bibel, 1986). A flush left/ragged right margin is the most effective format for text memory. (Thompson, 1991). Unjustified text may be easier for poorer readers to understand because the uneven eye movements created in justified text can interrupt reading (Gregory & Poulton, 1970; Hartley, 1985; Muncer, Gorman, Gorman, & Bibel, 1986; Schriver, 1997). Justified lines require the distances between words to be varied. In very narrow columns, not only are there extra wide spaces between words, but also between letters within the words (Gregory & Poulton, 1970).
Line Length (length of the line of text; the distance between the left and right margin)	Lines that are too long make readers weary and may also cause difficulty in locating the beginning of the next line, causing readers to lose their place (Schriver, 1997; Tinker, 1963).
Blank Space (Space on a page that is not occupied by text or graphics)	Blank space around paragraphs and between columns of type helps increase legibility (Smith & McCombs, 1971) A general rule is to allow text to occupy only about half of a page (Tinker, 1963). Too many test items per page can make items difficult to read.

Cognitive analysis of items

One of the most important principles to uphold in the study was to ensure that whatever revisions and/or enhancements made to the items would provide for the widest possible range of students with disabilities to be afforded access to and evaluated on the grade-level content as defined by the state's content curriculum standards. In order to achieve this goal, the first step in the process involved a thorough review of the content curriculum standards, the existing test content blueprint, and the item specifications by two independent content expert reviewers. The independent reviewers first analyzed the content curriculum standards for the strands of content underlying each of the standards, including the constructs to be measured by the standards for grade 5 and grade 8 reading and mathematics. The independent reviewers then analyzed the grade 5 and grade 8 reading and mathematics test content blueprints for range and balance of items in each grade-level reading and mathematics assessment. They also reviewed the item specifications for depth of knowledge (e.g., recall, application of skills and concepts, and strategic thinking or analysis), range of difficulty, and conceptual and cognitive complexity of the items as required by the content curriculum standards.

During the review of the content curriculum standards, the test content blueprint, and the item specifications, the independent content expert reviewers also reviewed existing items on the grade 5 and grade 8 reading and mathematics Spring 2006 operational assessment for optimal accessibility, including cognitive complexity. Items were examined specifically for the mental processes students might use in order to respond correctly in a given question. To guide this review process, the independent reviewers used three levels of depth of knowledge (i.e., recall, basic application of skills and concepts, and strategic thinking or analysis) as defined by Norman Webb (2002; 2006). In addition, the *Bloom's Taxonomy of Educational Objectives I: Cognitive Domain* (Bloom, 1956; [Revised] Anderson, et al., 2007), the most widely used scheme for labeling levels of cognitive processes, was also used. Using this taxonomy in conjunction with Webb's depth-of-knowledge levels, the independent content expert reviewers classified each item into one of three cognitive levels as follows:

<u>Recall & Reproduction</u>

The independent reviewers determined that recall questions were those written to measure students' ability to recall isolated facts, concepts, principles, processes, and routine procedures. The independent reviewers determined that a student answering a recall question would most likely either know the answer or would not know the answer; that is, to answer correctly the question, the students would not need to solve a problem or figure out the correct answer, they would simply recall from memory the answer. When reviewing the Spring 2006 grade 5 and grade 8 reading and mathematics assessments, items requiring recall of information were coded Level 1 by the independent reviewers.

<u>Application of Basic Skills and Concepts</u>

The independent reviewers determined that application questions were written to measure students' simple interpretations or limited applications of data or information. Questions written at this level would typically require some problem-solving skills and would

include the engagement of some mental processing beyond simple memory or recall. Application items require students to make some decisions as to how to approach the question or to solve the problem and would, therefore, require more than one mental step or mental process. When reviewing the Spring 2006 grade 5 and grade 8 reading and mathematics assessments, items requiring application of skills and concepts were coded Level 2 by the independent reviewers.

Strategic Thinking or Analysis

The independent reviewers determined that strategic thinking or analysis questions were written primarily to measure students' skills involving evaluation of data and problem solving. Responding to these questions would require application of good judgment and problem-solving skills. These questions involve higher cognitive processes, with multiple mental steps or mental processes required to answer the question or to solve the problem. When reviewing the Spring 2006 grade 5 and grade 8 reading and mathematics assessments, items requiring strategic thinking or analysis were coded Level 3 by the independent reviewers.

In examining the cognitive level of each item, the independent reviewers independently asked questions, including, but not limited to, the following:

- What skills and concepts are assessed in each item, and how closely does each item measure what it is intended to measure?
- Does the item focus on conceptual understanding, fact-based content, or processes/skills?
- What is the depth of skill or reasoning required by the item (e.g., locate information vs. drawing conclusions; computation vs. applying a concept in a new context)?
- What is the vocabulary load of each item, and what is the overall readability of each item?
- How complex is the thinking required for each item (e.g., use of figurative language, simplicity of language)?
- How many steps or mental processes are needed to solve the problem or to figure out the answer?
- Is a chart, table, graphic, or other artwork used to support a given item, and, if so, how complex is the chart, table, graphic, or other artwork to interpret?
- For each reading passage, with accompanying items, what are the genre, the text structure, and the length of the passage?

In addition to the review of each grade 5 and grade 8 reading and mathematics item, the independent content expert reviewers made preliminary recommendations concerning which of the items might best lend themselves to revisions and/or enhancements designed to allow for more accessibility for the defined 2% of the population of the students. In this review, the independent reviewers examined the existing Spring 2006 operational items to see if the items would lend themselves well to revision and/or enhancement while retaining the construct of the standard being measured. The reviewers examined each existing item for optimal accessibility, including, but not limited to, the following:

- Using simpler/more familiar language while retaining the construct of the grade-level standard measured;
- Simplifying graphics, tables, charts, artwork that might be too complicated or confusing;
- Reformatting and/or adjusting the layout of a given item and/or passage with items (e.g., adding more space between paragraphs of a reading passage, dividing passages into segments followed by items related to a specific segment or "chunking" of a given item or a passage);
- Underlining or bold facing key words/phrases or symbols;
- Numbering paragraphs within a passage, if students needed to refer to a specific paragraph;
- Reordering items for passages (e.g., with accompanying items ordered from literal meanings to more interpretive or analytic, more sequentially ordered, etc.) for greater accessibility;
- Adding short introductions to a given passage, if the passage needed additional background information, thereby providing support or a purpose for reading to the reader;
- Adding and/or enlarging graphics, tables, charts, artwork, etc., including graphic organizers that provide conceptual support;
- Adding definitions and/or simpler words set off by parentheses; and
- Adding helpful hints or points to remember.

Each independent content expert reviewer's results were carefully tabulated, and the results were compared and discussed. The content reviewers also considered the quantitative review of test items. Joint qualitative and quantitative discussions led to decisions as to which items might best lend themselves to revisions and/or enhancements by committees of educators. Thirty items for each grade (grade 5 and grade 8) and content area (reading and mathematics) were selected and prepared for review by committees of educators for potential revisions and/or enhancements, including scaffolding strategies.

.

What do we mean by enhancements to items using scaffolding strategies?

To clarify any confusion about what the term "scaffolded" means for a paper/pencil test item (from the perspective of this study), scaffolding is an approach to enhancing items. While "scaffolded instruction" provided the framework for the revisions to the items, not every item could be revised using scaffolding strategies. (See also Appendices B and C for content-specific examples of items that could be revised using scaffolding.)

There are several forms that scaffolding can take (Dickson, Simmons, & Kame'enui, 1995). Scaffolding can come from teacher or peers, content, task (such as breaking it down into manageable parts), and materials. The purpose is to provide support during learning in order to gradually remove the support when learning becomes solidified and/or the learner becomes more independent. This is why it's often referred to as "scaffolded instruction." Examples are as follows:

- Teacher/peer: more support with introduction of new concepts, tasks, strategies; support is slowly removed over time; peer scaffolding would be something peers reading together or solving a problem. Guided think alouds are another example of teacher/peer scaffolding.
- Content: simpler versions of the (essence) content/concepts are introduced before more challenging (deeper or broader) ones are introduced
- Task: simpler tasks or complex tasks broken into smaller steps are introduced before more challenging ones or new applications are expected
- Materials: use of graphic organizers, study guides, visual cues, predictable patterns in texts, etc.

Use of scaffolding is generally considered an effective teaching practice for all students, and there is a great deal of research to support the use of scaffolding for struggling learners. This is why Georgia considered revising test items by providing enhancements that apply scaffolding strategies for the 2% test items.

In a testing situation, tasks or materials seem to be the most viable approaches if the goal is not to change the content assessed (e.g., reading assessments still use grade level text and ask students to identify characters, setting, or main idea). Below is an example of how scaffolding might be applied in a testing situation using a reading task for the 2% population compared to the general assessment**.**

> Typically, general education students might read a 2-page text passage and answer both literal and inferential questions about the passage. Test questions may or may not be "ordered" to parallel how the information is presented in the passage.
>
> Scaffolding for the 2% population might look something like this: Students are given the same reading passage as the general education population at that grade level, but the passage is "segmented" – split up into parts of approximately the same length:. While the same questions used for general education might be used for the 2% population, those that apply directly to each segment would appear right after that section of the text and begin with literal (text-based) questions first. Questions for the 2% would follow an order that parallels how information generally appears in the passage. Inferential questions, such as author's purpose or theme, would come at the end, after the entire passage (all segments) had been read and literal questions have been answered. All constructs assessed in the general education assessment are also assessed in the 2% assessment. (See also Appendix F for a sample of segmenting a passage.)

Something discovered when attempting to "chunk" passages: The original intent of chunking was to divide reading passages into meaningful subsections for purposes of providing a type of organizational scaffold that would assist conceptual understanding for readers, thus reducing the load on students' working memory. However, it was discovered that chunking a given passage in this way did not always work well. In other words, some passages did not lend themselves to being divided into *conceptually*

meaningful subparts, such as an introduction that clearly presents the controlling idea of the piece, a body that provides key ideas with supporting evidence, and clear a conclusion that brings the controlling idea full circle. Passages that did not generally lend themselves to being broken into meaningful subparts were literary passages, such as poems, plays, and diary entries, and expository texts without clearly established text structures, meaning expository texts with one clear text structure and identifiable signal words and semantic cues (Hess, 2008). This was due in part to the fact that the passages used for the Georgia pilot were not originally written (or chosen) for their ability to be divided in this way. Two alternatives to chunking a passage into conceptual parts were (1) segmenting passages into parts of approximately the same length; or (2) creating/composing new passages with chunking in mind.

As a result of this discovery during the preparation phase of the study, the desire to use existing Georgia reading passages, and time constraints for conducting the pilot study within the Spring 2006 school year, segmenting was chosen as a viable strategy to include for making texts more accessible in the reading assessments. This decision was supported by research that students with learning disabilities prefer passages presented in this format (Dickson, Simmons, & Kame'enui, 1995) and that segmenting builds in test breaks for the reader. Recent research suggests that segmented text may be more accessible to students with disabilities (SD) and does not affect performance of non-SD students; therefore, it does not alter the reading construct. Since segmenting passages improves the reliability of the reading assessment without altering the construct, states can apply this feature into their assessments (Abedi, Leo, Kao, 2008).

A lesson learned from this experience is that in order to "chunk" a given passage in a meaningful/conceptual way, or to break passages into segments so that items can be grouped with the segments, test developers need to know that the initial passage with questions must lend itself to these types of scaffolding. This should be considered early-on, before any passages or items are actually written. Also, one needs to carefully examine the state's initial content test blueprint early in the process to ensure alignment. In this way, accessibility can be woven into the fabric of the modified alternate assessment. States should not assume this can be done with any existing general education reading passage.

Other strategies to enhancement items

Another example of how items might be enhanced or revised is to simplify the question stem or distracters, so that excessive vocabulary or multiple subordinate clauses are eliminated. While the goal was to keep the original construct as assessed in the general education assessment, the content experts found that sometimes revisions of this kind did result in reducing the cognitive complexity even if the items remained aligned to the grade level content standards.

Working with Teacher Committees

One significant component of the project included the involvement of educators. This step in the project served to enhance stakeholders' understanding of the link between assessment and classroom teaching and learning specifically as it related to students who

consistently, across the years, were and are struggling to meet expectations. Using the items selected by the independent expert content reviewers from the Spring 2006 operational assessments as potential candidates for revisions and/or enhancements, a committee of educators was convened. The committee consisted of approximately 13 educators for each content area (26 educators total). For the most part, committees were comprised of classroom teachers, with some curriculum experts, administrators, measurement experts, and educators with special expertise in understanding the diverse needs of the student population struggling to meet expectations. Additionally, staff from the Georgia Department of Education, Divisions of Assessment Research and Development, Special Education Services, and Academic Standards, were on hand to assist the committees and answer questions or concerns.

Committees met on separate days - by content area - and worked for a day and half each to review and revise items. The committee members were divided into two groups by grade level (5 and 8). Each group reviewed selected items for mathematics or for reading. The committee members reviewed each selected item with a particular emphasis on the specific construct assessed and cognitive complexity or depth of knowledge of the item for the targeted population. The actual performance of each item for the general population and/or the percentage of students who correctly answered the item were not provided; however, committees were provided with the frequency of each incorrect response for items, which was helpful in thinking about how students might be interpreting the item.

During the review of each item, the committee of educators was asked to provide some recommendation as to the estimated difficulty of the existing item, without revisions and/or enhancements, for the targeted population of students. The purpose of this review, using each committee member's recommendation as to the estimated difficulty of each item, helped guide what revisions and/or enhancements to a given item might, in fact, improve accessibility for the given population of students. Specifically, the committee members:

- Reviewed the purpose and goals of the alternate assessment, based upon modified academic achievement standards.
- Defined the population of students and achieved a common understanding of the academic characteristics of students who consistently struggle to meet expectations.
- Reviewed the general education content curriculum standards or what students should know and be able to do at a given grade level.
- Received training in understanding depth of knowledge, range of difficulty, and conceptual and cognitive complexity of each item for the targeted population.
- Estimated student performance on each item and determined the content probed by each item.

- Reviewed graphics, visual supports, etc. and suggested revisions and/or enhancements to visual aspects in order to remove any challenges that might impact students who consistently struggle to meet expectations.

- Reviewed each mathematics item and each reading passage with items, and revised each item to include revisions and/or enhancements designed to provide for improved accessibility of the item for students without compromising the constructs of the material being tested.

- Made recommendations for future item development.

The meeting began with training committee members on how to review the selected items for the purpose of providing revisions, and/or enhancements designed to enable each item to be more accessible. Information concerning depth of knowledge, cognitive complexity of items, principles of universal design, etc. was provided for reference. The training also included a discussion as to what types of revisions and/or enhancements to items could be made with the goal of maintaining the integrity of the construct.

After the initial training, the committee members were divided into groups according to grade level. Committee members were first asked to review each item independently to determine what, if any, revisions and/or enhancements might be made to the item. Group discussion regarding each item then followed. The group discussion focused upon analyzing the thought processes students might follow in order to find the correct answer to the item; what aspects of the item might be confusing or might have barriers preventing struggling students from answer the item correctly; and what, if any, revisions to the item and/or enhancements to the item might be made in order to improve its accessibility to students who are consistently not meeting expectations. Possible revisions and/or enhancements included, but were not limited to, the following:

- Reformatting, including enlarging font, adding extra space, and bolding certain words,
- Reordering items and reordering passages,
- Adding footnotes and helpful hints,
- Adding graphics, charts, tables, and artwork, including graphic organizers,
- Simplifying the language, and
- Providing definitions for certain words.

Following the review meeting, a summary report was prepared capturing the committees' revisions and/or enhancements to each item. Recommendations for future item development, with the goal of items that will be all-inclusive in design without compromising the constructs of the material being tested, were also provided.

Based upon the committees' review, items were revised accordingly and placed on field test forms. A discussion of the field test and the subsequent results of the field test are provided in the sections below.

Use of curriculum implementation surveys
An added component of this research study was to customize a survey for each grade level and content area assessed, in order to gather data from Georgia educators as to (1) the degree of implementation of content standards taught to the 2% population and (2) teacher expectations for this population. The model used for the curriculum implementation survey was an adaptation of Andrew Porter's Survey of Enacted Curriculum (SEC). Porter's SEC format was modified to reflect the specific content standards and objectives for Georgia educators teaching mathematics and reading at grades 5 and 8.

For this survey, teachers were asked to estimate the approximate instructional time spent for each strand and objective during the school year: from no time - to slight emphasis (1-10 lessons/year) - to moderate emphasis (11-20 lessons/year) - to sustained emphasis (21+ lessons/year) - to systematic instruction (almost every day). Teachers were then asked to identify the depth or breadth of their expectations for each content objective for the 2% population. Surveys were posted online near the end of the school of year. Given the time of year, educator participation was not robust, but it was adequate to begin to analyze how instructional practice and teacher expectations might also be impacting the 2% population. (See Appendix E for the sample survey format for mathematics.)

Appendices

A. Bibliography of Resources
B. Training Examples for Mathematics
C. Training Examples for Reading
D. Sample Summary Template for Item Selection - Mathematics
E. Sample Curriculum survey (separate PDF file)
F. Sample reading passages without and with segmenting and enhancements (separate Word file)

Appendix A

Bibliography

Abedi, J., Leo, S., Kao, J. (2008). "Segmented text study." Paper presented at the 2008 CCSSO National Conference on Student Assessment, Orlando, FL.

Anderson, L. & Krathwohl, Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Rths, J., & Wittrock, M. (Eds.) (2007). *A taxonomy for learning. teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Beck, I., McKeon, M. & Kucan, L. (2002). *Bringing words to life: robust vocabulary instruction*. NY: Guilford Press.

Dickson, Simmons, & Kame'enui. (1995). *Text organization: Curricular and instructional implications for diverse learners*. Eugene: National Center to Improve the Tools of Educators, University of Oregon.

Hess, K. (2004) "Applying Webb's depth-of-knowledge levels in reading." [online] available: www.nciea.org

Hess, K. & Biggam, S. (2004). "A discussion of text complexity, grades K-high school," published by NH, RI, and VT Departments of Education as part of the New England Common Assessment Program (NECAP) Grade Level Expectations for Reading (www.necompact.org or www.nciea.org).

Hess, K. (2008) "Teaching and assessing understanding of text structures across grades." [online] available: www.nciea.org

Johnstone, C., Liu, K., Altman, J., & Thurlow, M. (2007). "Student think aloud reflections on comprehensible and readable assessment items: Perspectives on what does and does not make an item readable" (Technical Report 48). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. [online] available: http://cehd.umn.edu/nceo/OnlinePubs/Tech48/TechReport48.pdf

Perie, M., Hess, K., & Gong, B. (2008). "Writing performance level descriptors: Applying lessons learned from the general assessment to the 1% and 2% assessments," www.nciea.org

Petit, M. & Hess, K. (2003, 2006) "Applying Webb's depth-of-knowledge levels and NAEP levels of complexity in mathematics." [online] available: www.nciea.org

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: author.

Robb, L., Richek, M., & Spandel, V. (2002). *Reader's Handbook: A Student Guide for Reading and Learning*. Wilmington, MA: Great Source Education Group/Houghton-Mifflin Co.

Schumm, J. S., (Ed) (2006). *Reading assessment and instruction for all learners.* NY: Guilford Press.

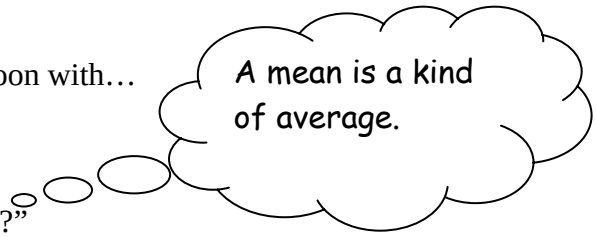
National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the national mathematics advisory panel.* US Department of Education, Washington, DC. [online] available: www.ed.gov/MathPanel

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Appendix B – Training Examples

Some ways to revise or enhance mathematics test questions…

1. Add a Hint: Use visual cue, such as a thought balloon with…



Definition or Synonym
Short or Simple Example
Procedural Prompt: "What is the rule when you see…?"
Provide Reminder of the Correct Formula

2. Provide a scaffolding enhancement that helps to organize information:
(a) procedurally in order to solve or test solutions; or
(b) conceptually to understand how information is interrelated.

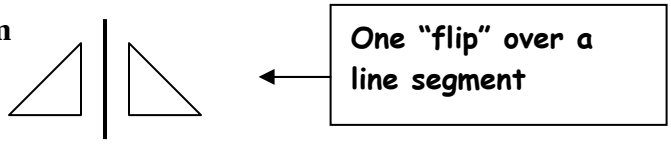
Add a T-chart or graphic organizer to help find and organize information – should be customized for the problem, *not a generic table with too many boxes*

Monday	Tuesday	Wednesday	Thursday

Add sub-questions or steps to break up/think through multi-step problems before solving

1. Circle the question you need to answer
2. What key words in the question tell you what to do? (e.g., "all together" means to add)
3. Underline information needed to answer the question
4. Is there extra information you do not need?

Add a graphic to illustrate a term



Add a familiar graphic organizer *to focus/assist conceptual understanding,* such as a blank hundreds chart or number line

0	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

0

3. Provide Representations for Testing Solutions

Blank Graphical/Grid

Blank Table of Values – customize for problem

x	***y***

4. Reduce Language Demand/Vocabulary Load
Eliminate a sentence, a phrase, or extraneous language.

Before	**After?**
Kathy's parents are remodeling her bedroom and she can have new wall-to-wall carpeting. The room is 20 feet long and 15 feet wide. How large will the carpet have to be in square feet to fit Kathy's room?	

Substitute another word that means the same thing WITHOUT changing the construct. Substitutions should be words in common use.

Stem
Distracters

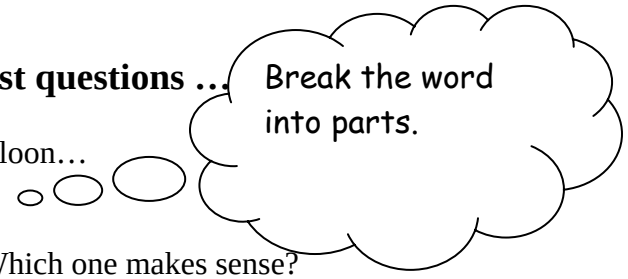
Add organizational text features

Separate problem context from question with white space
Use bullets with key ideas, instead of full sentences

5. **General Format & Order of Questions – consider:**

Visual discrimination issues (e.g., should a place value item have a "6" and a "9")
Spacing of /between text and graphics and responses and question/stem
Order of items (e.g., group all measurement items together, then all algebra items, etc.)
Order of difficulty of items (e.g., don't place all difficult items at the end of the test; order the math strands so that the beginning and ending strand are less demanding than the middle strands)
Text/font size
Use consistent item format and scaffolding structures throughout (e.g., use the same graphic organizer for the same conceptual idea in different items)

Appendix C – Training Examples
Some ways to revise or enhance reading test questions …



1. Add a Hint: Use visual cue, such as a thought balloon…
- Definition/synonym
- Short example
- Prompt: Try each meaning in the sentence. Which one makes sense?

2. Provide scaffolding for conceptually organizing information (interrelationships, time sequences, etc.):

Add a T-chart or graphic organizer to help find and organize information – should be customized for the problem, *not a generic table with too many boxes*

Character Name	Character Trait

Add sub-questions or steps to break up multi-step problems

5. Circle the question you need to answer
6. Underline key words in the question (e.g., who, how, where, etc.)
7. Locate information in the text needed to answer the question

3. Reduce Language Demand/Vocabulary Load without changing the construct

Substitute another (more familiar) word that means the same thing WITHOUT changing the construct.

Shortened stem, changed 2 distracters using more common synonyms	
In paragraph 9, what does the word genuine mean? A. real B. pretend C. content D. anxious	In paragraph 9, what does genuine mean? A. real B. pretend C. happy D. worried

Simplify Stem, use bold (or underlining) for key words	
The main point the author is making in this passage is about the A. hardships of ocean travel in the nineteenth century. B. struggles of the early immigrants entering America. C. many opportunities to make money in America. D. effect of immigration on European countries.	This passage is **mostly** about A. hardships of ocean travel in the nineteenth century. B. struggles of the early immigrants entering America. C. many opportunities to make money in America. D. effect of immigration on European countries.

Simplify Stem and make distracters complete sentences	
The United States eventually reduced the number of immigrants allowed to enter the country because A. the United States already had too many people. B. the immigrants were taking away jobs from American workers. C. the immigrants had too many hardships to face in America. D. the country that the immigrants came from was angry about their leaving.	Why did the United States reduce the number of immigrants? A. The United States already had too many people. B. The immigrants were taking away jobs from American workers**.** C. Immigrants had too many hardships to face in America. D. The country that the immigrants came from was angry about their leaving.

Simplify Stem & Distracters	
Reread the lines beginning with "I admired" (line 45) and ending with "aching jaw" (line 64). The speaker **most** admires the fish because she thinks it A. has escaped from being caught fishermen. B. is strong and intelligent. C. has strange movements. D. has the speckled barnacles.	The speaker **most** admires the fish (lines 45-64) because of its A. ability to escape. B. human-like intelligence. C. strange movements. D. speckled barnacles.

4. **General Format of Presentation of Questions or Passage**

Spacing (e.g., between paragraphs, more white space with dense text)
Order of questions (e.g., literal to inferential; logically following the presentation of information in passage)
Use underlining, bold, or CAPITAL letters of text to call attention to key words or to assist in locating vocabulary in the text. Do not *italicize* key words in text.
Scaffolding of Passages:

- Meaningful "chunking" of passage when possible (e.g., when there is a clear text structure with signal words) with corresponding questions – determine where to break up the text and place the questions immediately after that section; or
- Segmenting of passage into parts of approximate length with corresponding questions – determine appropriate length for segments and place the questions immediately after that section (See Appendix E for a sample of segmenting example.)
- Inferential questions at the end, after literal questions have been asked.

Appendix D - Sample Summary Template for Item Selection

Summary for All Reviewers

Content Area _____ **Grade Level __**

Explanation Key
P values are for general population
SELECTED ITEM = at least 2 reviewers (qualitative and/or qualitative) flagged item for potential revisions

Explanation of Qualitative Review Categories:

- **Content Strand/Concept Tested**—in addition to the specific mathematics strand/standard tested, what concept is actually targeted for assessment?
- **Wordiness**—supports understanding/question (+), neutral (0), or makes question confusing (-) (e.g., context gets in the way)
- **Vocabulary**—appropriate for grade level, above nominal grade level, or below nominal grade level; vocabulary used in mathematics should be one grade level below. Tier 1==very simple common words/need no direct instruction; Tier 2=non-content specific, but terms used across disciplines and require instruction; Tier 3=content specific terms that require direct instruction (See Beck, McKeon, & Kucan, 2002)
- **Cognitive Demand** –Webb levels 1, 2, or 3 (See specific DOK examples in Karin Hess' articles, applying Webb levels in reading and mathematics)
- **Use of Graphics**—supports understanding (+), neutral (0), or detracts understanding (-)
- **Spacing**—does the layout of the item help (+) or hinder (-) deciphering the item?
- **Negative use**—does the item use negative phrasing, and does this make the item more confusing than necessary? Add this to the comment box
- **General item quality/clarity**—what is your overall impression of this item [good (+), neutral (0), needs improvement -)]?
- **"other comments"** – might be related to statistical analyses
- **Candidate for Revision?**—Mark if the item should be selected for review by the teacher committee
- **Source of Challenge** (Webb, 2002): "The Source of Challenge criterion is only used to identify items where the major cognitive demand is inadvertently placed and is other than the targeted content, concept, skill and application. Excessive reading demands, cultural bias, or specialized knowledge could be reasons for an item to have a source of challenge problem. Such item characteristics may cause some students not to answer an assessment item or answer an assessment item incorrectly even though they have the content knowledge, understanding, and skills being assessed. Items with an appropriate source of challenge level will differentiate between those students who have the content knowledge and understanding the assessment item intends to measure from those students who do not have this knowledge."

Sample Summary Item Review Template for All Reviewers Test Math Grade Level _

Example of Summary of Quantitative (statistical) and Qualitative (content) Reviews & Item Selection for Item Revision or Enhancement											
Item #	GA Content Tested		Items Selected for pilot	STATISTICAL REVIEW Candidates?		Cognitive Demand/ DOK	Graphics use & comment	Potential Source of Challenge?	CONTENT REVIEW Candidate for Revision?		Other comments about items	P value (all students)
	Strand	Concept		Item has "potential"	"problem-atic"	1-2-3	+/0/-		REVIEW #1	REVIEW #2		
	NSN	Prime factors exponents		X		2	0				Multi-step	.7
	PRA	Absolute value				1	0					.8
	SP	Choose best display of data	X			2/3?	0	Multiple terms reading load	X Need to simplify	X confusing item Multi items better; add a graphic?	Multi-step Vocabulary terms (Tier 3)	.31
	NSN	Order fractions/ common denom	X			2	0		X Needs spacing	X better spacing between fractions	Multi-step more words than needed	.47
	PS	Multiply, add, estimate		X		2	0				Multi-step	.76
	PRA	Select equation for solving problem	X			2	0	Reading load solutions with equations	X confusing wordy simplify	X confusing item-what's being tested?	Multi-step – ask for solution, not equation too much reading	.34
	GM	2 flips	X			1	- Visually confusing?	Visual discrimina-tion?	X Awkward stem graphic	X use 2 sentences in stem		.6
	SP	Data table; mean & mode	X		X	1 (mode)			X Simplify	X Could be two items	Don't need both mean & mode	.54

Appendix E - Sample Georgia On-Line Curriculum Survey – PDF file not attached

Appendix F – Sample of segmenting of Passages – Word file not attached