Achieving Accurate Results for Diverse Learners with Access-Enhanced Items: Summary of Results from VAELL Grant

Rebecca Kopriva, University of Maryland
Phoebe Winter, Independent Consultant
Chen Su Chen, Post Doctorate Student

This research was supported by a Title VI Enhanced Assessment Instruments Grant and a Title I Enhanced Assessment Grant Application Number S368A030005. The contents of this article reflect the views of the researchers and do not represent the policy of the Department of Education, nor is endorsement by the federal government implied.

Abstract

New strategies for test development, including access-based item development, are critical to the valid, reliable, and accurate assessment of some students. Access-based item development minimizes some language challenges while providing compensatory avenues to access meaning, problem solving, and demonstration of solutions. Having accurate results for all students, including English Language Learners (ELLs) and students with language-based disabilities, is essential to ensure the accountability of educational systems, determine how to best meet students' educational needs, and track student progress over time. The ultimate goal is to design either sufficiently edited or parallel forms that provide comparable scores for students who access assessments in ways that are most appropriate for them.

The access-based item format produces test items specifically designed to increase access for ELLs and students with certain language-based disabilities. The access-based item represents a carefully crafted variation of a traditionally written item or an item that has used construction techniques of the type discussed here. Recent research in the area of test accommodations for these students suggests that these types of formats may be effective for increasing access to test content (e.g., Abedi, Courtney, & Leon, 2003; Kopriva, 2000). The results discussed here provide some promising evidence that using appropriate item templates plus focusing on specific structural and contextual factors impacts item access while retaining the integrity of the target objectives and the comparability of scores across forms.

Achieving Accurate Results for Diverse Learners: Access-Based Item Development

With the implementation of the two most recent reauthorizations of Title I of the Elementary and Secondary Education Act and the resulting emphasis on appropriately assessing all students, state educational agencies, test developers, and research and advocacy organizations have actively looked for new ways to produce more valid test results for students who are English language learners (ELLs) or who otherwise have difficulty accessing the content of an assessment because of language or literacy challenges. Several authors (for instance Kopriva & Lara, 1997; Solano-Flores & Barber, 2001; Solano-Flores & Trumbull, 2003) argue that current large scale academic testing practices are insufficient and in fundamental ways problematic for English language learners. Cognitive psychologists (see Chudowsky & Pellegrino, 2003; Pellegrino, Baxter, & Glaser, 1999) argue that current practices are out of date with current theories of thinking and learning, and that foundational underpinnings of today's assessments are significantly flawed. Donovan, Bransford, & Pellegrino (1999), Heath (1983, 1989), among others, emphasize that cognitive processes associated with making inferences about students' academic abilities are influenced by linguistics, language acquisition, dialect, culture, prior experiences, and current setting, emphasizing that theories associated with these fields have not been properly integrated into large-scale assessment practices.

There is a growing body of research attempting to determine the proper use of accommodations for ELLs in large-scale tests. Unfortunately, the current results are inconclusive due, in part, to the complexity of issues surrounding appropriate accommodations for particular students (Kopriva & Mislevy, 2005; Sireci, Li, & Scarpati, 2003), the varying levels of construct analysis in the research, and how the methodology of the accommodation research was conducted (e.g., the use of blanket vs. individualized accommodations and the study of individual vs. packages of accommodations; Tindal & Ketterlin-Geller, 2004). Recent research associated with the Selection Taxonomy for English Language Learner Accommodations (STELLA), which is a computerized taxonomy for assigning accommodations for individual students, supports the use of appropriately assigned packages of accommodations for ELLs (Kopriva, Cho, & Carr, 2006). Recognizing the limitations of after-item-development accommodations and potential effects on validity of score interpretation, some researchers have turned their attention to creating tests that are accessible to a larger proportion of the tested population.

As the testing industry continues to embrace new developments, this research should encourage the field to increasingly design new items and tests from the ground up that embrace features such as those discussed below. Frameworks, such as Mislevy's Evidence Centered Design (1999), Embretson's work focusing on differential needs of students (e.g. 1998), research that takes advantage of using technology to vary testing approaches and goals (Bejar et al., 2003; Samuelsen & Kopriva, 2004) and emerging work that focuses on the identification and measurement of learning progression in large scale tests promises significant paradigm shifts in measurement (e.g. Popham, Pelligrino, Berliner, Flick, & Kopriva, 2006). As this information is being integrated into the testing culture, however, it appears that most of the work in the near future will continue to focus on re-engineering existing assessment systems. This includes adapting existing items, writing items that are more access-based but still interchangeable with the types of items used in today's assessments, and modifying review procedures and data collections. This article is being written to encourage new approaches but also to provide guidance in re-engineering existing assessments. In this way, adaptations can be somewhat responsive to today's students, while not inhibiting the development of new generations of assessments more geared to ongoing learning and the more complex differential cognitive approaches and challenges evident in the U.S.'s diverse population of students.

This article describes an approach to creating item frameworks that supports the development of access-based assessments, with items *that measure the same targets* as those on standard test forms. The structure for the frameworks derives from research in the fields of language and cognition as well as subject-related learning. After an introduction to the concept of comparable item frameworks, information about the process of developing access-based items will be provided. Some empirical research will be presented to lend support for the idea of incorporating access-based item development into large-scale testing systems.

Item Structures

The idea of creating structures or frameworks for item development is not new (e.g., See Haladyna & Shindoll, 1989). Even before the publication of Haladyna and Shindoll's article, some test developers used rules for item development that placed strict constraints on item stems and response options. More recently, research in developing parallel item structures has focused along general two lines: (1) systematically varying item features to create items with the same specific psychometric properties and (2) creating items that measure the same precise targets within a domain.

First, Bejar, Lawless, Morley, Wagner, Bennett, and Revuelta (2003) have developed and tried out procedures for creating item models that allow for computer generation of quantitative items for adaptive testing. The goal of their research is develop items that are interchangeable (isomorphic) in terms of content covered and psychometric properties. Content experts developed models that would supply content variability but equivalent difficulty. From the examples shown in their article (2003) and an earlier chapter (Bejar, 2002), their item generation models allowed for rule-based variations in the specific numbers or variables used in the problem, with the rules used to maintain aspects that could affect difficulty (e.g., numbers are constrained to a specific range; the ratio of one number to another is maintained), and the item structure, format, and context the same across variations within the item model.

Using a cognitive design system approach, Embretson (1998) developed abstract reasoning test items. The items were generated based on item structures that specified theory-relevant aspects of the item. The resulting assessment had acceptable psychometric properties and demonstrated that successfully using cognitive theory to generate items provides strong evidence of construct validity. Carrying the idea of using cognitive theory to the achievement-testing sphere, Enright, Morely, & Sheehan (2002) used construct-driven item development to develop mathematics problems. Using problem-solving theory as a foundation, the authors systematically varied three item features in mathematics problems in the areas of rate or probability to determine item parameters were affected. The authors were able to explain much of the difficulty of the items, particularly the rate items, based on values of the item features. For rating items only, the item features predictably affected discrimination and guessing. The results of this study indicate that construct-driven item development, at least in mathematics problem solving, has promise, but that we need better information about the constructs assessed and how they are manifest in items.

Researchers have also attempted to use less constrained item models to generate items that measure the same knowledge and skills. In application of Haladyna’s model to performance assessment, Solano-Flores, Jovanovic, Shavelson, & Bachman, (1999) created task shells for generating science assessments. The goal was to develop comparable assessments by controlling for level of inquiry, science concepts, and format. While the shells generated tasks that were similar in appearance, their psychometric properties varied and the targets of the tasks were not interchangeable. In another analysis of study results, Stecher at al. (2000) concluded that we do not have enough understanding of how these performance tasks work from a cognitive perspective to vary features in a predictable way. More research into the student-task interaction through activities such as think-alouds and deeper cognitive analysis of task requirements should improve the comparability of such tasks.

Perhaps the most general framework for item and task development where psychometric qualities are allowed to vary is Mislevy et al's (2003) Principled Assessment Design for Inquiry (PADI) model. The model links cognitive psychology and research in science learning as the basis for developing frameworks for assessment design patterns. These comprehensive design patterns are intended to provide a structure for assessment developers to produce tasks that support desired inferences about student knowledge and skills by clearly connecting the inferences to types of evidence needed to support them and types of situations that are likely to invoke construct-relevant behaviors. The limit to this work is that targets are identified at a fairly high grain size, which leaves open the possibility that rigorous target equivalence may still be elusive.

In a more limited sphere, ongoing work that investigates the development of interchangeable structures that address access issues has begun. Kopriva & Mislevy (2005) and Solano-Flores and associates (1999, 2001, & 2003) have begun to systematically address this issue. Kopriva's work shares many of the features of construct-driven line of structure-based inquiry through researching how to build items where item-specific targets are appropriately defined and constrained. In that way, non-target access barriers over like items can be minimized through collecting appropriate information about student factors and subsequently assigning proper options. Solano-Flores argues that variation in proficiency across items and across the four domains (reading, writing, speaking, and listening), two languages (L1 and English), and sometimes dialect impacts how items and forms are structured and the number of items that may be necessary to use for students with specific profiles. Linguistic levels vary by context, language instruction, and various cultural indicators. This research, as well as those from other researchers, will be embedded in the discussions that follow.

The Centrality of the Targeted Construct

Traditionally, most measurement experts argued that, for similar inferences to be assumed across students taking different forms, testing conditions should be standardized for all test takers, and forms should be parallel in terms of their psychometric properties. Constraining testing by using the same set of testing conditions over students was considered to be an equality issue and fundamental to being able to robustly generalize score meaning. Parallel forms typically meant that a similar balance of item coverage should also be apparent across forms. As such, form results could be equated and placed on a common scale to produce standard scores across forms. Preceded by Messick's arguments related to test validity and invalidity (see 1989), improvements regarding how to consider and build parallel forms and items have been suggested over the last 15 years or so. Cognition experts (e.g. Pellegrino et al., 1999; Resnick & Resnick, 1992) and some psychometricans (for instance, Haertal & Wiley, 1993; Linn, 1993 ; Mislevy et al., 2003; Shepard, 2000; Wiley & Haertal, 1996) have stressed the importance of explicitly developing construct-driven tests and forms, where depth as well as breadth is required, content domains are clearly specified, and at times item intent might be explicitly defined at some grain size level, especially for constructed response items. Recently, creating computer generated items using identified algorithms has heightened the demand for item rules-based templates for forced choice response items (e.g. multiple choice), as well as constructed response items although the grain-size issue remains of concern (Forte & Popham, 2006).

It became clear that, while the field had spent considerable time improving the explicit descriptions of test-level constructs, how item level targets were to be defined had not been thought through well enough. This was especially apparent for researchers interested in improving access to test materials for special populations because, as studies investigated variations in testing conditions, the issue of comparability of results over conditions rested squarely on the documentation of construct invariance in items and tests (Bielinski, Thurlow, Callender, & Bolt, 2001; Kopriva, 1996; Kopriva & Lowery, 1994; Solano-Flores et al., 1999; Tindal, Health, Hollenbeck, & Harniss, 1998). Solano-Flores and Shavelson (1997) and Solano-Flores and others (1999) proposed building item shells, identifying item targets at a rather broad level and outlining key dimensions of item construction to be considered in order to improve access for ELLs. Kopriva (1996, 2000) and Kopriva and Martin (1998) argued that identifying the proper grain size for the item targets, and explicitly identifying non-target elements in items and tests, were key in building access-based items and tests for assessments in general. Recently Poplam and others in the state of Wyoming have attempted to address the grain-size issue in Wyoming's new state testing program (Popham et al., 2006).

Defining the targets too broadly led to items measuring different aspects of constructs; defining the targets too narrowly meant that no variation in conditions or item language could occur. Further, following the lead of access researchers by identifying what aspects of items were *not* relevant to the target, allowed item writers to construct interchangeable items that addressed the target features more precisely. One of the foci of a large study, convened in 2002, was to specifically develop item templates that would address the grain size of item targets and what item and test elements can vary (Kopriva & Winter, 2003). The procedures, while applicable to item writing in general, are particularly important for populations with language, literacy or attention challenges, including English language learners, who have only limited access to standard forms and materials. Once targets have been properly set, specific guidelines can be applied to build access-based items and forms.

Access Defined

Item and test access allows a student to properly approach, reach and understand the relevant content in items and the problem solving requirements associated with the content. Once the requirements are properly understood, access is also making available to the student the proper resources and tools to solve the problems and the availability of proper information exchange avenues that allow the student to communicate their answers so that they are properly understood by the scorers or identified scoring mechanism. Thus, the essential points of access during the student-item interaction are at the *apprehension* stage of problem solving, the *activity* stage of finding a solution to the problem identified in the item, and the *explanation* stage where the student effectively communicates the solution (Winter, Kopriva, Chen, & Emick, in press).

Access for each of these stages needs to be maximized throughout the presentation materials. Presentation materials or accommodations include access-based items and forms. They also include other tools and relevant resources. Examples of tools are word lists, mathematics manipulatives, blocks, or scientific supplies; examples of resources are video clips, a dynamic problem interactive environment surrounding the presentation of the items, or information links related to non-targeted prior knowledge needed in an item or series of items. Access-based materials provide broadened opportunities for some students to *apprehend* or understand the problems they are expected to solve by improving the language load in the text materials or providing compensatory information or stimuli to support the item requirements. Some options, such as contextual surrounds or providing appropriate tools, are designed to improve access by allowing students to engage effectively in problem solving *activities* where they would otherwise be meaningfully barred. Carefully designed items and judicious use of supporting tools and resources can be presented in such a way that increases the number of possible *explanation* avenues available to students to communicate their solutions. Finally, research (see Scireci et al., 2003 for a review) suggests that, for English language learners, administration and response accommodations that occur post hoc to test development are usually necessary but not sufficient without access-based presentation materials. Since many aspects that constrain administration and response conditions originate within the items and associated materials, attention to the choice of administration accommodations, as well as response options, must be considered when presentation materials are being developed and selected.

The Process of Developing Accessible Items

Completing Access-Based Item Templates

Access-based item templates identify both the measurement intent of an item and what the item measures that is not intended. Access-based item templates, when completed, provide information about each of the components that specify the conceptual parameters of an item. This information includes location within a hierarchy of definition that spell out how items substantively fit relative to the content standards, as well as precise information about the measurement targets of each item. Then, target-irrelevant knowledge and skills are identified that explain how test takers will communicate with and use item contents. The idea is that any parallel items should assess the same measurement intent but vary in terms of the active target-irrelevant characteristics. Identifying the irrelevant characteristics also allows personnel to assign individual tools, administration, and response accommodations to students at the item level (e.g., in computer-based testing) or at the test score level (e.g., in paper and pencil tests). Table3 is an example of a completed template. Since terminology for the template definitions being used here is not standardized in the field, one set of terms will be identified and explained as they are used. Readers are encouraged to adapt these definitions and terms to fit their own testing vocabulary as needed.

Template Components

General Construct Maps

Targeting what is and isn't intended in individual items is based on clearly specifying **construct maps** that progressively unpack broad-based constructs identified for testing and stipulated in the test specifications.

Figure 1: Construct Map of a Grade 3 Number Strand in Mathematics

Domain	Number
Standard	Knowledge of number operations
Indicator	Understand visual representations of operations

As Figure 1 illustrates, each item is assigned to a Domain, Standard, and Indicator, corresponding to the levels of a state's content standards or test publisher's domain definitions being used to develop the assessment. In this example, an item would be measuring the Domain "Number," the Standard "Operations," and the Indicator "Understand visual representations of operations." Once this assignment of an item to the general map is completed, specification of the core content in the item targets and so on can be undertaken.

Item Core Targets

To begin, items are specified by identifying information that becomes what we have defined as the core target. These cores identify the target at the correct grain size that we believe is necessary for access comparability. Outside of the core, items can change in *any* manner that is not specifically identified by the core. For instance, if not specified, type of item can change (e.g., multiple-choice to constructed response) or scaffolding within a type. As such, information in the core must be as complete as possible.

Any items that share a core are considered to be interchangeable from a construct-driven structural perspective. For coverage purposes (including issues of both breadth and complexity of skills), it is expected that more than one core will be identified in those Standards or Indicators that are being tested. Each core has three dimensions. Within each dimension, test developers are asked to consider the three stages of access. Table 1 explains each of these components.

Table 1: Target Core

i. Objective: targeted content knowledge and skills. Targets are defined within a specific Indicator and usually the same across more than one item (depending on how detailed test inferences are expected to be).

ii. Item-specific subject matter (ISSM): content knowledge and skills that define only an *instance* of the Objective. ISSM varies idiosyncratically by core and should include an explanation of the content as well as the targeted complexity of the item's skills.

iii. Item-specific constraints (ISC): Any additional constraints are specified here. These may be test level constraints that would be invariant over all items, such as not changing the item type or retaining the same numbers over mathematics items that share the same core. Other constraints may be core specific and may or may not be content-related. An example of a content-related constraint is requiring knowledge of prerequisite skills (such as multiplication or addition) when the target is asking students to compute an algebraic algorithm. If this is not specified as a constraint, then it is assumed that lack of prerequisite skills could be compensated for in a parallel item.

Decisions about constraints should ultimately be guided by research or clearly qualified within the test inferences. For instance, Bejar (2002) suggests that some variations have been found to be reasonably trustworthy in terms of not changing the measurement intent of the item. In cases of tests that use only one item type, inferences should be clearly constrained to reflect the range of type of information that this item type can produce about student knowledge and skills. Figure 2 and Table 2 provide an example of how an item fits within the targeted definition hierarchy.

Figure 2: Grade 3 Mathematics Item

Maria is going to spend her allowance on stuffed animals. Each stuffed animal costs $5. What is the largest number of stuffed animals she can buy if she has $28?

A. 6
B. 5
C. 4
D. 3

Table 2: Hierarchy of Targeted Item Information

Construct Maps:	**Domain**		Number
	Standard		Operations: Understand and use operations on numbers (addition, subtraction, multiplication, division)
	Indicator		Choose and use the appropriate operations to solve single-step or multi-step word problems
	Core:	**Objective**	Solve single- and multi-step word problems involving a single operation, including problems about money
		ISSM	Division, ignore remainder, or repeated addition, index number of addends; context – find largest possible number
		ISC	Multiple choice item type, use same numbers, present the unit cost before the total amount of money available.

Target-Irrelevant Knowledge and Skills

Irrelevant information in items is ALWAYS a part of testing because information about item requirements, use and item constraints must always be communicated back and forth between assessment and test taker. Therefore, the goal of developing access-based items isn't to eliminate all irrelevant or ancillary information in items. Rather, it is to 1) become increasingly cognizant of the irrelevant aspects of items and 2) deliberately develop items that use specific, non-target ancillary knowledge and skills in intentional ways to minimize the barriers to testing for students with particular challenges and needs. It stands to reason that, in order to accommodate all test takers so the integrity of the intended targets can be communicated properly, a limited number of item and form options, along with form supports, will probably be needed.

After the item's targeted construct knowledge and skills have been identified, the preferred non-target ancillary components should be specified. If already constructed items are being evaluated, they should be examined to determine which irrelevant factors are being used to communicate the target information. Kopriva and Winter (2003), Winter et al., (in press), and Mislevy et al., (2005) identified specific target-irrelevant components in items like those used on today's achievement tests by considering structural and contextual factors particularly salient for ELLs and some other students. Mann, Emick, Chen, & Kopriva (2006) investigated the use of a set of access-based multiple-choice and constructed response items that were part of a large-scale district-wide test which was mostly comprised of standard items. They found that the validity of the multiple-choice scores for ELLs was still less than for non-ELLs but there was evidence that the constructed response test scores in grade 5 were as valid for low ELLs and poor readers as they were for non ELLs and good readers, respectively. Siskind (2004) is investigating if access-based items which follow the type of guidance discussed here are useful not only for ELLs but also for some students with learning disabilities and hearing impaired students as well. In addition to English access-based forms, additional broad-based forms which attend to other irrelevant factors may be important to develop, including forms tailored to students with L1 literacy, and forms for students with some amount of literacy in both L1 and English (Kopriva, 2000). In each of these situations, some of the ancillary structural factors identified in the template will probably be completed at the form level (for all items), rather than at the individual item level. Issues of context, prior knowledge, and format still need to be considered for each item and will be briefly discussed below.

Table 3 provides an example of a completed template. This is an evaluation of the item found in Figure 2. The item is from a released item data-base and is not considered particularly access-based. However, it provides an illustration of the type of ancillary factors an item writer would want to address. The next section will provide examples of how the specification of irrelevant skills and knowledge that address language and contextual issues of ELLs can be translated into effective items for many students in this population.

Table 3: Completed Template

Core:	**Objective**	Solve single- and multi-step word problems involving a single operation, including problems about money
	ISSM	Division, ignore remainder, or repeated addition, index number of addends; context – find largest possible number
	ISC	Multiple choice item type, use same numbers, present the unit cost before the total amount of money available.
Target Irrelevant Knowledge and Skills	**Nouns**	'Stuffed animals'- vocabulary and double meaning 'Allowance'—vocabulary
	Context	Having an allowance is not a common experience for some students
	Verb	"is going to spend" is a complex tense

	Adjective	‘Largest’—double meaning
	Sentence Structure	Use of adverbial phrase.

Implications

Several implications stem from developing and using this type of access-based template. Two will be noted here. First, the issue of ancillary fit between test taker and item that was just discussed relates directly to the level of error due to mis-match that different types of assessment systems are willing to tolerate. In general, it seems reasonable that the more detailed the inferences, the less error of this type a test should be willing to allow. For instance, in situations where test results have high stake consequences, the level of ancillary fit should be more tightly controlled. Research aimed at identifying acceptable levels of mis-match should be undertaken for different types of tests and tests used for different purposes.

Second, from a measurement perspective, this type of item-structure template primarily addresses construct-equivalence across like items. That is, items produced from the access-based template are considered to be interchangeable for particular groups of students because the integrity of the target 1) has been clearly identified and 2) has subsequently not been disturbed by changes that minimize the effect of text based challenges salient for the specific groups. Research has begun that addresses whether one set of parameters can be fit for students who need and receive different items that attend appropriately to their text based challenges. It is hypothesized that, for students for whom today’s standard items adequately serve their communication and problem solving needs, parameter estimates based on these items will be retained when the students take like items from forms using access-based item templates.

However, for students for whom the standard items are problematic, it is expected that item difficulty will change over the two administrations when these students are given the proper access-based items and particular supplementary supports they need (Siskind, 2004). This will occur for students who have enough construct knowledge, whereby, when they are able to access the item requirements, they will demonstrate their targeted knowledge and skills more effectively than when they were blocked from the item content. Changes in item difficulty for these students mean that the parameter estimates of their ability under standard parameter setting conditions could be confounded by difficulty resulting from target-irrelevant variables rather than difficulty arising from their knowledge and skills regarding the content expectations of the items.

Building Access-Based Items

Malcolm (1991), among others, is insistent that multiple avenues must be used to provide access to the item for ELLs. The common purpose of the multiple avenues is to provide alternative compensatory support in addition to minimizing language and cultural challenges. As such, the avenues must be varied to respond to the divergent needs of this population. Multiple avenues need to involve the addition of item elements discussed below that frame and support context and support or replace text. They also involve the thoughtful use of language tools, home language forms, and administration and response options geared to the specific needs of individual students. For instance, Solano-Flores and others (2001; 2003) found that students utilized information from English and from L1 at different times, for different reasons, and to varying degrees in different items. Shaw (1997) emphasized the importance of using some type of activity to thoughtfully engage the students and supplement missing background knowledge. It is important that a variety of avenues be evident in each item. Findings from a cognitive lab investigation suggested that both compensatory and reductive mechanisms in items were important contributors to apprehension of item requirements (Winter et al., in press). Further, results suggested that compensatory strategies that were not germane to a particular student did not appear to negatively impact his or her performance. This finding suggests that more general items could be constructed that would be useful to students with various access needs.

Item Development Factors

The purpose of this section is to *briefly* highlight considerations item writers need to address. For a more in-depth look, including relevant research and samples see Kopriva (in preparation). Although the contextual and structural categories are artificial, they provide a conceptual framework for developing access-based items. Context and culture pervade an adequate understanding of choices that are made about language, text and item supports for various students, and attending to access in various items and rubrics is a learned skill that uses the range of assistance provided here in different ways and for somewhat different purposes. A clear distinction between what item writers should and shouldn't do and when is not possible. Given this caveat, hopefully the table can provide some illumination of the types of activities which are required in order to build access-based items. Following this section an example of how to apply the item template and consider the item aspects will be provided.

Table 4: Structural and Contextual Factors

Contextual Factors		
Culturally Broad Experiences	Cultural expectations seem to have an impact on how a student understands the requirements of an item. These cultural expectations become especially problematic when a student's experiences or home culture values are distinctly diverse from those typically experienced by the mainstream population in the U.S. (Kopriva, 2000).	1. Prior knowledge that assumes mainstream U.S. experiences 2. Expectations that assume a common U.S. value system
Clear and Explicit	In classrooms, ELL experts know that it is not enough to assume that ELLs will understand test	1. Direct, explicit explanation of item requirements

Expectations	expectations and approaches familiar to those in the mainstream U.S. population. For large-scale tests that measure content over diverse types of students, clarity in expectations relative to all aspects of the items and tests need to be explicit and clearly stated (Farr & Trumbull, 1997; Kopriva, in preparation).	
Prior Learning Expectations	Two types of prior learning experiences are at issue: (1) is the pre-requisite knowledge related to target content that is required for an examination of more complex skills, or pre-requisite content knowledge and skills at older grade levels where knowledge builds on a foundation developed in the earlier grades, and (2) is the use of non-targeted content as context in items, especially items that measure processes such as reading comprehension or science inquiry skills (Kopriva, in preparation).	1. Assumptions of prior learning required for complex skills 2. Use of non-target content as context in items
Structural Factors		
Simple Language Structures	The issue of language organization is particularly salient for ELLs, because text in their home language is almost assuredly structured differently than English text. The basic presentation of text for ELLs involves a conscious and complex re-appropriation of structural conventions explicitly or implicitly learned as part of their home language experiences (Abedi & Lord, 2001; Johnstone, 2003).	1. Use of simple sentences 2. Use of similar paragraph organization 3. Use of present tense and active voice 4. Minimizing use of rephrasing
Vocabulary	The vocabulary in all items, directions, and supplemental test materials of both academic and social English must be considered when developing access-based items (Farr & Trumbull, 1997; Kopriva, 2000).	1. Use of familiar language 2. Limit use of substitute words 3. Careful use of multi-meaning words
Effective Visuals	ELLs are both learning the language and learning to read at approximately the same time. Visual cues provide help for ELLs as they struggle to learn English and become literate. However, not all graphics are equally beneficial; thus care must be taken when using this type of support (Filippatou & Pumphrey, 1996; Winter et al., in press).	1. Use of relevant visuals 2. Use of an effective format 3. Use of illustrations to mirror text 4. Use of illustrations to replace text 5. Use of first person visuals 6. Use of visuals to organize events in time 7. Use of visuals to clarify textual meaning
Effective Item Format	Formatting print material to focus the reader, clarify purpose, and otherwise effectively move the reader through the information is central to the work of print media professionals. Emerging work suggests that attending to and clarifying item	1. Separating key ideas 2. Clearly identify item questions 3. Use of titles 4. Use of mixing symbols and

	formats does play a part in making items more accessible for this population (Kopriva & Mislevy, 2005; Winter et al., 2004).	text 5. Use of examples 6 Highlighting key words or phrases 7. Use of boxes or lines

Text Amount	The guiding principles in making decisions about text amount is to retain the content complexity of the intended target while providing enough information in a non-textured form as possible (Kopriva & Lowrey, 1994; Wong-Fillmore & Snow, 2000).	1. Retain complexity of target while using non-interfering contextual cues
Text Support	Text supports are identified as text-based aids that occur over and above the item text and provides supplementary support (Abedi, 2001; Hipolio, 2006).	1. Bilingual glossaries or dictionary 2. Monolingual glossaries 3. Picture-word dictionaries 4. Side-by-side forms (a.k.a. dual language test booklet)
Content-Based Resources	Resources that provide additional information during the testing experience to fill in gaps or provide a common referent that minimizes additional textual load	Providing prior learning experiences information or primary sources via 1. Primary source documents 2. Prior experience information
Activities	Activities, such as a brief (15 minute) interactive discussion or an activity prior to the period of testing that provides context for all students and ELL experts suggest is another important compensatory support mechanism (Monore, 2004).	1. Brief interactive discussion 2. Brief collection of data
Maniptulatives	Content tools are objects which are item or content area specific and that can be used to aid the student in understanding the intent of a particular item, solve the problem or express their solution. For ELLs these tools provide interactive compensatory opportunities to cross over the minimum threshold for understanding key item elements or being able to navigate problems (Kopriva & Mislevy, 2005)	1. Concrete materials 2. Computer simulations with drag and click options and/or graphic/drawing opportunities
Impact of Home Language	Some item-drafting issues for ELLs reflect the influence of students' native language. For more frequently spoken languages and for those particularly prevalent in certain areas, some actions can be taken to minimize misunderstandings (Kopriva, in preparation).	1. Use of cognates 2. Reduce use of linguistically confusing words 3. Consistency with symbol use 4. Reviews of text by those familiar with the culture and language

An Access-Based Example

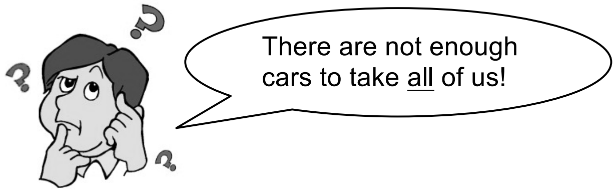
Consider the following fourth grade mathematics item. This item is a released item and a typical example of items that are used today on many achievement assessments across the country.

> At Jefferson Midlands Middle School, the sixth grade students and their teacher are planning a field trip to the state capital at the end of the year. In the morning they will visit the state legislature, and in the afternoon they will go to the zoo.
>
> There are 33 students in sixth grade. Five parents and two teachers will be coming with the students on the trip. Each of the adults has a car that can hold four students. One of the teachers says: "There are not enough cars to take all of us!" Do you agree with the teacher? Explain your answer.

This item was revised to be more accessible for English language learners, as shown.

- 33 students are going on a class trip. ← 1
- 5 parents and 2 teachers are going with the students.
- Each adult has a car. Each car takes 4 students.

A student says: ← 4



2

3

Is the student right? (circle one) **Yes** **No**

Symbols for "Explain 5

"Tactile Support" 6

Within the item there would be symbols to indicate directions or tools. For example, the explain symbol is introduced before the test and is common across all tests this state/district uses. It means students need to provide an answer and they can do so using words, algorithms, pictures, or other diagrams. A symbol for tactile support is introduced before the test and is common across all tests this state/district uses. It means that there is a tool available that students can tactilely manipulate to help them solve the problem.

1. Information that is not needed to set the context for the problem has been eliminated, reducing the amount of text.
2. Plain language principles have been applied to the item to reduce the semantic and syntactic complexity of the item. The sentences are shorter and straightforward, using present tense and active voice and reducing the use of prepositional phrases and dependent clauses. A visual is used to illustrate the item. Note that numerals have been used consistently throughout. The translation between a verbal and symbolic representation of a number was considered construct-irrelevant mathematics.
3. The formatting has been arranged to provide maximum access to the problem requirements. Each complete piece of information is presented separately, since, for this item, selecting the appropriate information from among relevant and irrelevant pieces of information was not part of the measurement target. The question is clearly separated from the rest of the text, and the two-stage character of the item, answering the question and explaining the response, is evident.
4. While both the base and the variation assume students are familiar with class trips, which may not be the case in all schools, potential cultural schooling bias has been reduced in the variation by having a student's statement the focus of the question. In some cultures, children are not used to questioning teacher judgments and decisions[1].
5. Students are given options for how they represent their response.
6. Students are allowed to use counters to help them represent and solve the problem. The targeted content knowledge and skills do not preclude allowing various methods of representation or solution, as noted in the ISSM. The manipulatives provide students who are ELLs a way to represent the text that may help them understand the problem situation.

Procedural Considerations

An iterative process for developing access-based items at the state or publisher level is being developed in conjunction with the South Carolina Department of Education. Because high quality access-based item development goes beyond traditional single session trainings, the iterative process appears to include an initial multiple-day training in which content item specialists from the agency or publisher interact with training personnel fluent in assessment, content, and access techniques. The training involves collaborative scaffolding in several content areas and over grades. It also includes formative feedback at many stages in the item writing process over the course of several months. This iterative process, which was first piloted during the VAELL (Valid Assessment of English Language Learners) project with university staff, students, and consultants included several rounds of review focused on the applying the principles of access and ensuring comparable alignment between versions of the item.

Current field testing on the process for four content areas (science, math, language arts, and social studies) and six grade levels (3-8) conducted with the South Carolina item writing staff and university researchers resulted in several insights. First, it has become evident in higher grades, more prior knowledge is assumed and is needed to access the grade-level content. As such, this needs to be attended to in addressing needs of students whose experiences may not have included the learning of key information necessary to access the target requirements. It is also more difficult to minimize language by using graphics in older grades, as the concepts were more complex and abstract than in the younger grades. In order to mitigate these factors consideration is being given to using computer-based simulations or movable options that allow for maximum interaction. Finally, items for the older graders are more likely to include multi-step directions. Item writers must diagnose out how to explicitly display each item requirement in a clear and coherent manner.

Within each content area certain item factors appear to be particularly salient. For example, across math content, it is important to separate key ideas, particularly for multi-step math items, using scaffolding formats (e.g., bulleting information), and removing distracting construct irrelevant information while retaining contextual support. For science, there seems to be a systematic need to make graphs/charts more readable. For younger grades, it is more possible to use first person visuals, mirror text visually or through concrete resources, or provide context. Test items in the older grades, on the other hand, present greater challenges with specific scientific language that can’t be mirrored or glossed. Social studies items for the younger grades contain more factual information, allowing for a reduction in low frequency words, mirroring of the text, and reduction in the range of directive language used. Social studies items for the older grades seem to be more abstract; thus it is important to focus on minimizing the directive language and using cognates. Social studies also is able make use of picture distractors more often than other context areas, particularly in the younger grades. As expected, Language Arts presented many unique challenges, particularly related to maintaining the targeted construct. Overall, however, there appears to be abundant opportunity to mirror text, simplify language structures, and, on a more limited basis, incorporate graphics as scaffolding to text. The use of plain language throughout the item and accompanying passages is an important tool but requires that item writers be cognizant of the target and the language being tested. Within Language Arts, state content specialists find it also is important to identify and minimize the use of non-construct relevant idiomatic or metaphoric language in passages when these are *not* part of the construct being measured: in some cases, idiom and metaphor are explicitly measured and are in the SC standards. Finally, ELA item writers find that it is particularly challenging to maintain measurement of the intended meaning in poetry passages as the intent of poems requires knowledge of the literal meaning of key language before the figurative understanding of the poem can take place. It appears to be important to use pictures, oral or brief written glosses, or other means, in order to ensure that students are being tested on the intended meaning of the poem's items and not getting caught, unintentionally, in the local, concrete meaning of its construct-irrelevant words.

An Analysis of Multiple Choice Item Pairs

This study will examine data from selected multiple-choice items collected from the recently completed VAELL project (Kopriva & Mislevy, 2001). A future paper will focus on the standard and access-based constructed response items. This project developed several multiple-choice and constructed response access-based test items from standard released items. These access-based items were then administered as part of a large-scale district-wide test to a diverse district located near a large metropolitan city. Within a month before and after the district-wide test administration, standard items were administered by teachers in classrooms to a large subset of students in the district who also took the large-scale test. The current analyses will examine if and how specific item characteristics differ from standard released items vs. access-based items.

Clearly, the conditions of item administration differ for the 2 sets of items. It is believed that much of what large-scale access-based items and accommodations are trying to accomplish would naturally be addressed in a classroom setting. Therefore, it is expected that any differences seen between items will be muted at best. However this analysis is still considered useful, as it can provide clues about how students might differentially respond to item characteristics that are deliberately designed to differ within pairs.

Given this important caveat it is hypothesized that a level of pervasive change within conditions would be considered to be a function of the administration. On the other hand, it is hypothesized that change that appears to be unique to the item pairs could be influenced by this item writing procedures. Analyses of the item pairs will be conducted on results from all students tested. Independent variables of ELL and reading status among other characteristics will be included to determine how these variables differentially affect the items.

Methodology

Procedures

The sample consisted of 2508 third (n=1283) and fifth grade (n=1225) students from 21 schools. Prior to test administrations, teachers of participating students (n=148) completed a questionnaire concerning each student’s participation in educational services, learning strengths and challenges, use of strategies in mathematics problem solving, assessment experiences, English/language arts skills, and factors that are hypothesized to either support or inhibit student access in testing math content. One section of the teacher questionnaire was devoted to teacher ratings of how often students successfully demonstrated knowledge and skills of particular mathematics elements that would be appearing in the VAELL test items. Teachers were asked how often over the course of their general classroom performance each student met content standards on a three-point scale (Rarely, Sometimes, Almost Always) at a medium grain-size level (e.g., for third-graders, an example is "*This student can solve a word problem involving a solution requiring subtraction with regrouping* or *This student can explain a simple multiplication fact using numbers or drawings"*). This teacher construct-identification-rating-system has been applied to characterize a number of different tests, including typical large-scale assessments throughout the U.S. and abroad (Achieve, 2004; Houang, 2004; Schmidt et al., 2001; Valverde, 2005). The characterization of target indicators on these tests and the teacher questionnaire for this project is modeled to be consistent with the grain-size identification used in the TIMMS curriculum content identification system. The teacher responses to these questions were used as the independent criterion in this study.

Subsequently, two parallel versions of a mathematics subtest were completed, one with standard released items and one with access-based items. 11 multiple-choice items and 8 constructed response items were identified from released large-scale tests in accordance with the Maryland Voluntary State Standards and selected as covering content standards that had been taught in the district in the first 3 months of the school year prior to the administration of the test. Access-based items were developed as per the procedures discussed above and were embedded in the district's mathematics test with about 70 items overall in each grade. The district-wide tests were administered over about 3 days. The standard items were administered after the 3-month learning window within a month before and after the district-wide test, and were administered by the teachers in the classroom to the same students in the 21 schools. This classroom assessment was to be treated as a unit test and given within a period of time that was most conducive to the content or content reviews being measured on the subtest.

Research Design and Analysis

The basic conceptual model for the analyses, taken from the project proposal, has three parts. This model can be seen in Figure 3 below. The three parts are

- Target abilities
- Ancillary abilities
- Test item response

The test item response reflects both target and ancillary abilities. When researchers can independently measure these abilities, it was suggested that analysis can be useful to determine the impact of ability measures on student performance.

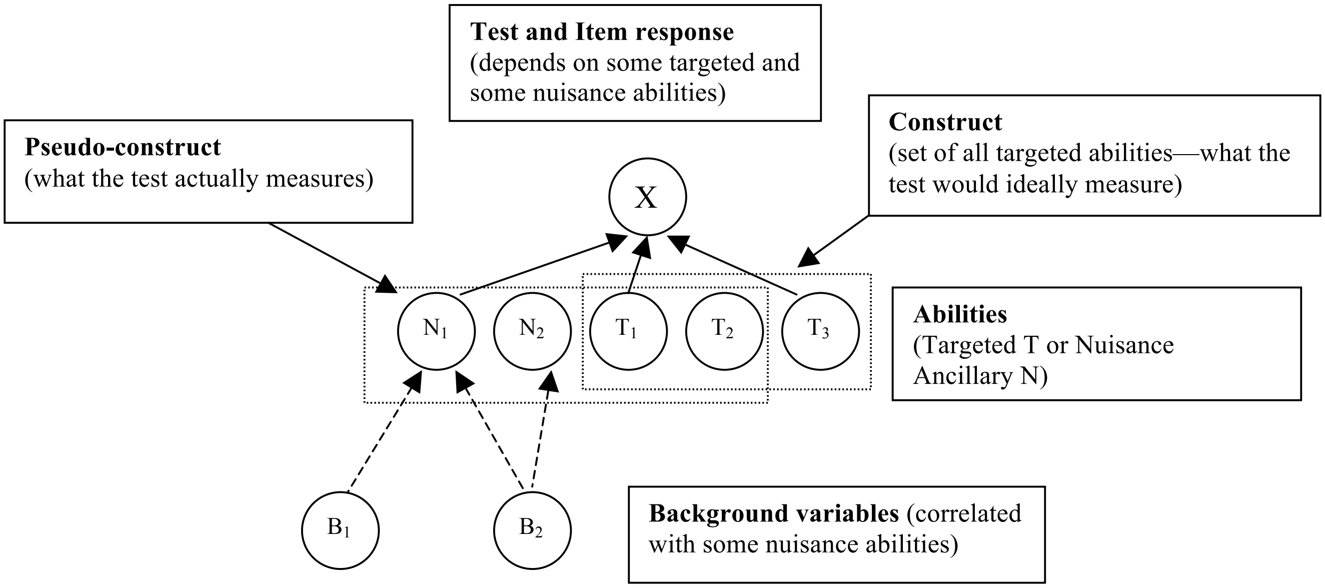

Test and Item response
(depends on some targeted and some nuisance abilities)
Pseudo-construct
(what the test actually measures)
Construct
(set of all targeted abilities—what the test would ideally measure)
X
N_1
N_2
T_1
T_2
T_3
Abilities
(Targeted T or Nuisance Ancillary N)
B_1
B_2
Background variables (correlated with some nuisance abilities)

Figure 3: Conceptual Model

For the purposes of this study, the mathematics knowledge and skills ratings associated with each item for each student is considered to be the estimate of the target criterion that the corresponding item is intending to measure. The ancillary skills used here were identified by the teacher for each student when the teachers completed the questionnaires. Four of the ancillary variables used in this study are reading level, testwiseness, item context, and psychosocial variables (frustration, anxiety, fatigue, distractibility, and motivation). The set of psychosocial variables were grouped by impact (of the five variables recorded, how many afflict the student). Further, a "gap" variable was also included in the analyses as an ancillary variable. This gap variable tracks how students who need and don't need accommodations perform across administrations. The notion here is that the gap should remain rather constant across administrations after a student's "true" mathematics abilities (as defined by the target ratings of the teacher), and the other ancillary variables are controlled for. Finally, the variable of English language learner status was included. This variable had five levels: Beginner ELL, intermediate ELL, advanced ELL, exited ELL or non-ELL (native English speakers).

Using the data obtained from the study, selected descriptive data were completed. Additionally, logistic binomial regressions were run on all items in order to analyze the relationship between the scores from each administration and the estimate of the students' target abilities from the criterion measure. This method was used because the estimated regression coefficients tend to be more stable and comparable across items. Dependent variables consist of dichotomous multiple choice item scores obtained under either the classroom or large-scale testing conditions for each student. First, only the criterion measure was regressed on the dependent variable scores. Subsequently, ancillary variables were added to each of the regressions in order to analyze the effects of the non-target variables. For each variable in each of the blocks, the slope of the betas was analyzed to determine significant difference from 0. For each item within each condition, the significance of the extent to which the target beta differs from 0 were analyzed with t tests.

It is argued that the coefficient (β) for the target evaluates how well the response measures students' true abilities as it provides a summative measure of discrimination of the target criterion relative to the item response for individual students. As discussed above, the estimate of the criterion is the teacher rating of each student's specific abilities with regards to particular knowledge or skills. On the other hand, coefficients for the ancillary abilities index the degree to which the ancillary variables distort the response away from the intended target meaning. As such, they index the components of invalidity in the response. These betas provide a level of comparison of the variables across item administrations, and the relative effect of the variables within administration. This index is the focus of this study rather than the traditional comparison of test scores. The beta was purposely chosen because the researchers for the study believe it provides information about the extent of "true score" clarity and impact on the item scores over students and within conditions.

Results

First, descriptive data will be briefly reported and then a detailed analysis of selected item pairs will summarize key findings. Unless otherwise indicated, the indicator that is being examined is the coefficient (β) or "beta" associated with the target and each ancillary variable. Betas range from 0 to +1 or -1. The target data for each item is coded 1-3 with 1 being little knowledge of the target construct. The dichotomous ancillary variables testwiseness, context, and gap are coded so that 0 indicates that the student is not fluent in the skills associated with the variable. For instance, a 0 for testwiseness says that the student lacks testwiseness skills. Reading data are on a continuum from 1-5 with 1 suggesting the student is consistently below grade level in their performance and 5 saying that the student reads above grade level. The psychosocial variable specifies whether or not the student exhibits frustration, anxiety, fatigue, distractibility, and motivation. This variable is reverse coded 0-4 so that 0 indicates that the student exhibits all of these behaviors. The ELL variable has five levels: beginner ELL, intermediate ELL, advanced ELL, exited ELL or non-ELL (native English speakers). Interpreting the following tables then, negative beta results suggest that students with less skill on the particular variable on the whole tend to score higher than those with more skill. As emphasized above, it is expected that real changes in impact caused by how items are constructed will probably be muted at best because of the difference in conditions within which these items were administered. For this reason, relatively small but distinct differences between standard and access-based items will be more closely analyzed as it is believed that the distinctions between the two may be actually be larger than they appear to be.

Descriptive data for Grade 3 follows. The box plot (Figure 4) summarizes the distributions of betas for each of the ancillary variables over students and over items. Testwiseness seems to be particularly influential although it covers a broad range in how it impacts items and student scores. In the access-based items, on the whole, the median effect is about -.2, indicating that it's effect favors students with limited skills. On the other hand, the median testwiseness beta in the standard items is about .1 which suggests a preference for students with fluent skills. Approximately 75% of the variance appears to be non-overlapping between the two administrations indicating a distinct effect overall.

The median of the gap variable (the factor that distinguishes between those who received accommodations and not) is close to zero over the access-based items, which is what is desired. Interestingly, the classroom administration of standard items favors the students needing accommodations, although this variable rivals testwiseness in its range. Our expectation that the classroom conditions have a substantial effect on student performance appears to be true, especially for students who need accommodations and even when the items are presented in standard form. Of note is the influence of the reading variable in both conditions, which singly has a higher impact than any of the others. Intriguing, though, is its impact in the large-scale access-based test where it looks like it acts more like a constant, as compared to how it functions in classrooms. The lack of range of influence for this variable in the large-scale administration (when certainly a range of reading skills is evident in the test takers) may suggest that it might be a proxy for other considerations. Psychosocial's influence in the standard items is also restricted in range, although it's median doesn't change much across administrations. The classroom impact may imply a baseline associated with test taking in general.

In reviewing the correlation table of the betas in Table 5 some interesting relationships emerge. The strongest relationships for the same variables across the two administrations are within the variables of reading, psychosocial, and ELL status, indicating a moderately positive effect across conditions. The lowest relationship for the same administration is context where the influence differs a great deal. Within each administration, a -.7 relationship between psychosocial and testwiseness suggests their inverse impact on the test scores. This means that, in the classroom and in the large-scale administration, testwiseness does not have much impact on scores for students whose scores are impacted by psychosocial concerns. For the standard items, a -.7 relationship between ELL and testwiseness implies that testwiseness does not have much impact on scores for students whose scores are impacted by ELL status. This latter relationship is weakened in the access-based items where the inverse relationship is low (-.3). Within the large-scale condition, there appears to be no relationship between ELL and reading (-

.01), implying that the impact of both of these variables operate independently of the other on access-based test scores. In the standard condition there is a small positive relationship (.2). Likewise, in the large-scale administration, no relationship exists between testwiseness and reading (-.01), while in the classroom condition the there is a small positive relationship (.2).

Figure 4: Grade 3 Box plots of Variable Item Betas

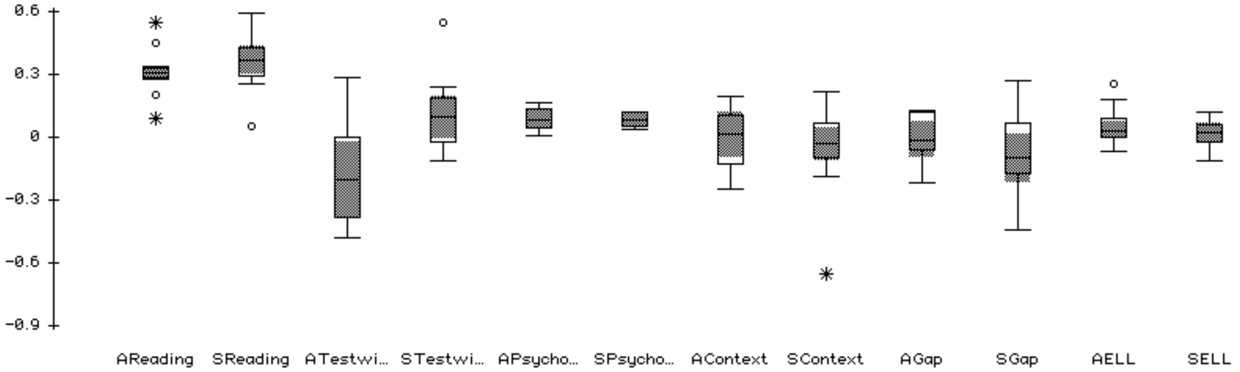


Table 5: Grade 3 Correlation of Variable Item Beta

	Standard Reading	Standard Testwise	Standard Psy/Social	Standard Context	Standard Gap	Standard ELL	Access Reading	Access Test	Access Psy/Social	Access Context	Access Gap	Access ELL
Standard Reading	1											
Standard Testwise	-0.218	1										
Standard Psy/social	0.408	-0.693	1									
Standard Context	0.463	-0.464	0.327	1								
Standard Gap	0.351	-0.195	0.256	0.086	1							
Standard ELL	0.177	-0.708	0.487	0.321	-0.128	1						
Access Reading	0.625	-0.186	0.218	0.579	0.213	0.088	1					
Access Testwise	-0.113	0.377	-0.694	0.055	-0.473	0.064	-0.007	1				
Access Psy/social	0.387	-0.297	0.62	0.371	0.034	0.141	0.384	-0.672	1			
Access Context	-0.158	-0.399	0.225	-0.032	0.647	0.007	-0.176	-0.299	-0.343	1		
Access Gap	0.172	-0.214	0.396	-0.152	0.262	0.335	-0.371	-0.278	0.022	0.041	1	
Access ELL	0.146	-0.513	0.761	0.102	-0.066	0.597	-0.009	-0.289	0.21	0.259	0.256	1

The box plot results for Grade 5 (Figure 5) are somewhat distinct from those found in Grade 3. Specifically, while the medians are not much different across the two administrations, the Context betas for the standard condition are spread across a broad range of items as compared to Context betas for the access-based items. Most of the standard items appear to either have little influence on student scores or favor those with contextual skills. The influence of this variable on the access-based items, on the other hand, more tightly hovers around the median and either slightly or more significantly favors those with limited skills. This differential impact across the item pairs are in contrast to Grade 3 results. For Grade 5, the gap variable is the factor, which most distinguishes the two sets of items. The access-based items appeared to most benefit students who do not need accommodations whereas the standard items administered in the classrooms, provided greater support to students who need accommodations. This finding is puzzling and not consistent with what was hypothesized. It may show the extent to which the classroom environment impacts and mitigates the intent of the access-based items for students who need accommodations, or it may reflect motivational distinctions between the two administrations. The betas for the ELL variable suggest that, although, those with higher English proficiency were favored, the distinction was narrowed for the access-based items as compared to their standard counterparts. Unlike Grade 3, the impact of testwiseness is similar across items for both sets. While Grade 5 reading had a greater range of influence across items than that found in Grade 3, across both item sets it still favored students with reading skills. The psychosocial factor, on the other hand, was the most constant across items for both the standard and the access-based sets and, in both cases, slightly favored those who did not present with psychosocial needs.

Examining the Grade 5 correlation table in Table 6 it appears that there are some strong relationships between the same variables across the two administrations, namely reading and testwiseness. But equally as important to note is the low relationship between several variables across the two administrations (i.e., psychosocial, context, gap and ELL). This suggests that relationship between variables across conditions is more complex than in Grade 3 and that the influence of particular variables differs significantly by condition. It also implies that, although, the box plot suggests that medians are often similar across items within a set, item pairs may not be similar. Within the standard administration there was a strong positive relationship (.73) between reading and context, which suggests that reading ability does have an impact on scores for students who scores are impacted by context. Interestingly, for the access versions, the relationship is weaker and inverse (-.485). This implies both that reading has less impact on scores for students who scores are impacted by context and that more reading skills are correlated with less contextual skills in item performance. Another strong positive relationship was noted within the standard administration between psychosocial factors and ELL status (.658). This relationship, albeit weaker (.446), is also seen within the access-based condition. Within the large-scale condition, there appears to be moderate relationships between ELL status and reading (.565), reading and psychosocial (.507), and testwiseness and context (-.571). Within the standard condition, there appears to be moderate relationships between testwiseness and accommodation status (gap) and between psychosocial and context.

Figure 5: Grade 5 Box plots of Variable Item Betas

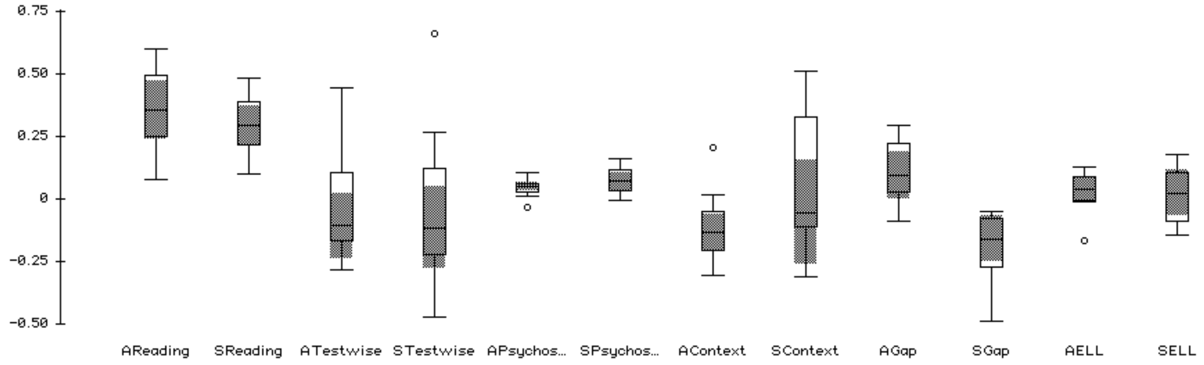


Table 6: Grade 5 Correlation of Variable Item Beta

	Standard Reading	Standard Testwise	Standard Psy/Social	Standard Context	Standard Gap	Standard ELL	Access Reading	Access Test	Access Psy/Social	Access Context	Access Gap	Access ELL
Standard Reading	1											
Standard Testwise	-0.218	1										
Standard Psy/social	0.156	-0.077	1									
Standard Context	0.73	-0.193	0.503	1								
Standard Gap	0.068	0.448	-0.267	-0.185	1							
Standard ELL	0.013	-0.146	0.658	0.408	-0.309	1						
Access Reading	0.651	0.199	-0.107	0.356	0.275	0.115	1					
Access Testwise	0.288	0.782	0.158	0.222	0.385	-0.206	0.304	1				
Access Psy/social	0.29	0.207	-0.048	-0.155	0.259	-0.161	0.507	0.363	1			
Access Context	-0.247	-0.503	-0.289	-0.114	0.091	-0.123	-0.485	-0.571	-0.279	1		
Access Gap	-0.375	-0.31	0.335	-0.283	-0.041	0.078	-0.433	-0.255	-0.094	0.15	1	
Access ELL	0.112	0.387	-0.139	-0.215	0.144	0.142	0.565	0.263	0.446	-0.703	-0.297	1

Analyses by Item Pairs

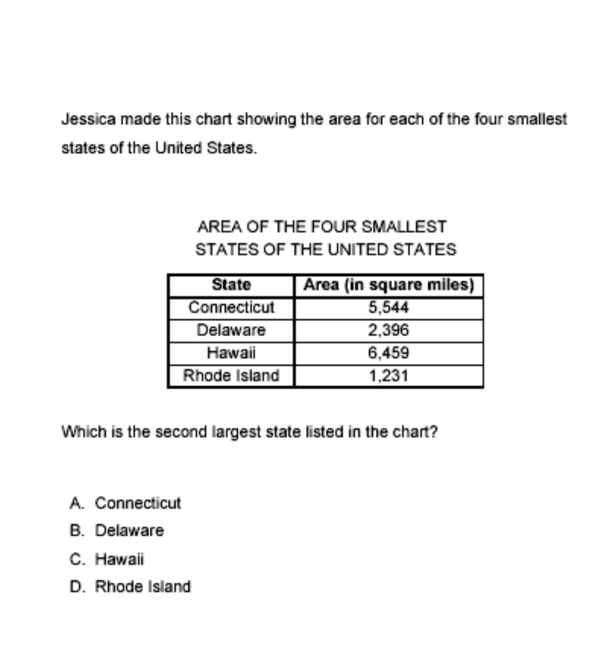
Below, selected item pairs will be presented for both grades 3 and 5. An analyses of each pair indicated that some changes occurred over all 11 in each grade. The pairs illustrated here (7 in third grade and 5 in fifth grade) were chosen to reflect the types of changes seen in ones not picked, or they were selected because they were unique in some way. Certainly, some of the issues discussed below will be artifacts of the individual items or of the particular population of students tested. However, the changes were distinct enough that the analysis appears to be worthwhile.

For each pair, the standard and access-based items will be shown, followed by a panel, which indicates the percent correct on each item for students with differing levels of mathematics ability as identified in the criterion. Then, the regression results of each item will be presented where the standard and access-based scores are the dependent variables. Initially, the impact of the independent variable of the target criterion rating will be displayed by itself (with the constant term). In grade 3 all 11 items were significant at p<.05 across both standard and access-based conditions. In grade five 10 standard items and 11 access-based items were significant at p<.05. Next, along with the constant, the betas of the target variable, and each of the 6 ancillary variables will be illustrated. In grade 3 all 11 items were significant at p<.05 for the standard items, but only 10 items were significant for the access-based items with VAELLN33 being the exception. In grade five, 8 standard items were significant at p<.05 but 11 items were significant for the access-based items.

Grade 3

Initially, regression analyses of all 11 of the items showed that the target criterion measure predicted mathematics achievement in the test scores for both the access-based items and standard items at a slope significantly different from 0. When the ancillary variables were regressed on the achievement data, the scores continued to discriminate the students' mathematics ability for 10 of the access-based items and 9 of the standard items. No variables were found to be pervasive across one administration condition versus the other. Reading was found to be significantly different from 0 in 10 of the 11 items for each of the conditions. This suggests that this variable continues to substantially define how students perform on items, whether they were administered in the classroom or in a large-scale setting and whether these contain more or less language. Results from 7 item pairs will be reported here follow by a brief explanation of the most salient findings. The remaining items not illustrated here present similar issues as those discussed below.

VAELL311 Standard VAELL311 Access-Based



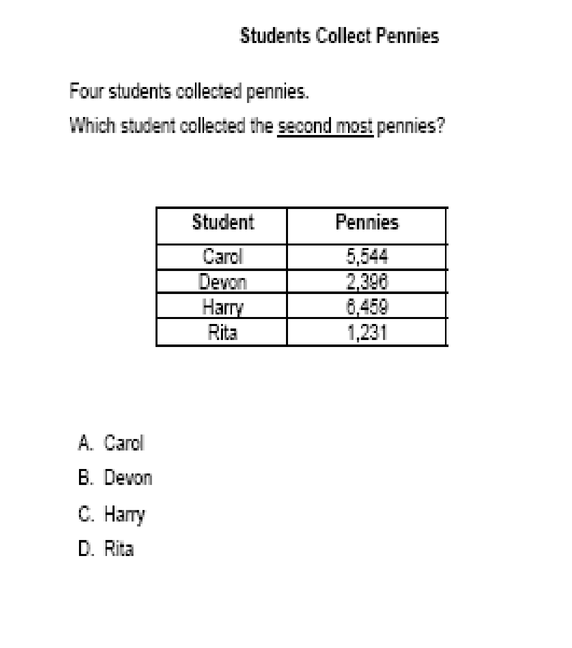
Jessica made this chart showing the area for each of the four smallest states of the United States.

AREA OF THE FOUR SMALLEST
STATES OF THE UNITED STATES

State	Area (in square miles)
Connecticut	5,544
Delaware	2,396
Hawaii	6,459
Rhode Island	1,231

Which is the second largest state listed in the chart?

A. Connecticut
B. Delaware
C. Hawaii
D. Rhode Island



Students Collect Pennies

Four students collected pennies.
Which student collected the second most pennies?

Student	Pennies
Carol	5,544
Devon	2,396
Harry	6,459
Rita	1,231

A. Carol
B. Devon
C. Harry
D. Rita

		VAELL311		Total
		0	1	
Math3.3b	Rarely	111	55	166
	Sometimes	234	216	450
	Almost Always	232	435	667
Total		577	706	1,283

		VAELL311		Total
		0	1	
Math3.3b	Rarely	109	57	166
	Sometimes	215	235	450
	Almost Always	191	476	667
Total		515	768	1,283

Standard						**Access-Based**				
Incremental χ^2	df	Coefficient	S.E.	Sig.	Variable	Incremental χ^2	df	Coefficient	S.E.	Sig.
118.88*	1	-1.255	0.159	**0.000**	Constant	141.309*	1	-1.232	0.160	**0.000**
		1.267	0.122	**0.000**	Target			1.406	0.126	**0.000**
40.911*	6	-2.083	0.278	**0.000**	Constant	70.43*	6	-2.422	0.278	**0.000**
		0.698	0.160	**0.000**	Target			0.700	0.165	**0.000**
		0.257	0.073	**0.000**	Reading			0.312	0.077	**0.000**
		-0.041	0.200	0.837	Testwiseness			-0.121	0.203	0.551
		0.122	0.045	**0.007**	Psychosocial			0.080	0.046	0.085
		0.204	0.174	0.241	Context			0.132	0.177	0.457
		-0.157	0.171	0.358	Gap			-0.032	0.101	0.752
		0.101	0.060	0.090	ELL			0.255	0.060	**0.000**

VAELL311

In this pair, as in most, reading is a prominent and significant influence across items. The psychosocial factors improved for the access-based item, however. This may be because of the perceived contextual load of standard item which used states (with long names); while the context was not central to the understanding of what was being measured, it may have produced anxiety before this was understood. On the other hand, the impact of ELL status is significant for the access-based. The reading level of "second most" is difficult for ELL students although it is unclear why it didn't have more of an influence in the standard (it tended towards significance, however, with a p=.09). The mathematics table shows the difference in correct response for students at different mathematics ability levels as defined in the target independent variable. For this pair it appears that most of the percent correct increase occurred for the students at the higher ability.

VAELL322 Standard

Ms. Kopriva is a band teacher. She had 9 old recorders. She bought 6 new recorders. Then 4 of the recorders had to be thrown away. Which number phrase can be used to find how many recorders were left?

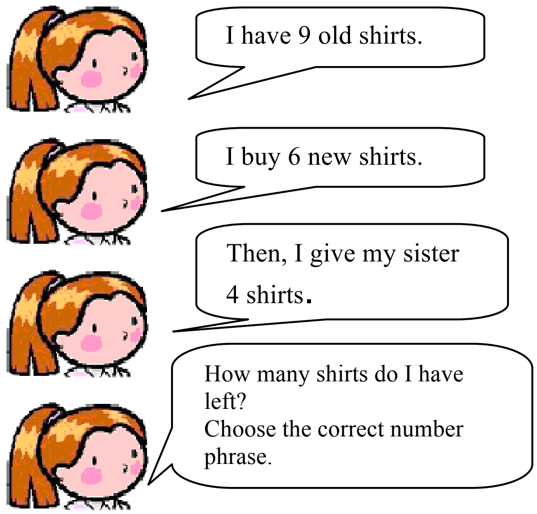
A. 9 – 4

B. 9 + 6 – 4

C. 6 – 4

D. 9 - 6 + 4

VAELL322 Access-Based


I have 9 old shirts.
I buy 6 new shirts.
Then, I give my sister 4 shirts.
How many shirts do I have left? Choose the correct number phrase.

A. 9 – 4

B. 9 + 6 – 4

C. 6 – 4

D. 9 - 6 + 4

		VAELL322		Total
		0	1	
	Rarely	96	72	168
Math3.1b	Sometimes	241	295	536
	Almost Always	179	400	579
Total		516	767	1,283

		VAELL322		Total
		0	1	
	Rarely	90	78	168
Math3.1b	Sometimes	255	281	536
	Almost Always	189	390	579
Total		534	749	1,283

Standard **Access-Based**

Incremental x^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental x^2	df	Coefficient	S.E.	Sig.
52.426*	1	-0.512	0.150	**0.001**	Constant	43.457*	1	-0.513	0.149	**0.001**
		0.828	0.116	**0.000**	Target			0.746	0.115	**0.000**
42.463*	6	-0.843	0.261	**0.001**	Constant	26.73*	6	-0.763	0.249	**0.002**
		0.178	0.157	0.255	Target			0.339	0.155	**0.029**
		0.391	0.074	**0.000**	Reading			0.314	0.071	**0.000**
		0.133	0.195	0.496	Testwiseness			0.037	0.194	0.850
		0.047	0.045	0.296	Psychosocial			0.010	0.045	0.816
		-0.014	0.172	0.937	Context			0.202	0.170	0.236
		0.074	0.170	0.665	Gap			-0.177	0.098	0.071
		-0.041	0.058	0.481	ELL			0.001	0.057	0.981

TX22

The reading variable continues to impact this pair; however, the target variable is not significant in the standard while it is in the access-based. It is not clear why the standard item is not predicted by the mathematics rating variable. Perhaps use of the specific band item is unfamiliar—recorders have several meanings and the students may be unfamiliar with using recorders in band. The language structure of the standard item is more complex than most of the 3rd grade items but the length of the item is similar to several. The gap variable in the access-based item, while not significant (p=.07), is close and is negative in its coefficient. This suggests that, overall, students with accommodations did better on the access-based item relative to their non-accommodated peers. One other 3rd grade pair illustrated a significant gap impact in the standard that did not lesson in the access-based. In no other item pair is the gap influence found at this or a significant level in one of the items and not the other. When correct response is differentiated by mathematics ability, the access item seems to slightly favor the students will least ability (by 4%) and have a slight negative effect of the higher two levels (3% and 2%, respectively). While a majority of the lowest students are receiving accommodations, the first-person format, perhaps, of the access-based item may be disconcerting to a few students at the higher ability levels.

VAELL3Bears Standard

Patty just received a letter in the mail telling about a new promotion with stuffed animals. When Patty has collected and shown proof of owning 125 stuffed animals she will receive the new Million Dollar Bear free. Patty has 79 animals right now. Which of the following equations show how many more animals Patty will need to collect to get her free Million Dollar Bear?

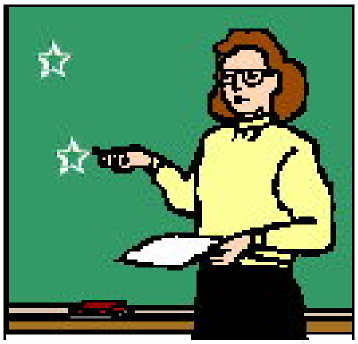
A. □ - 125 = 79

B. 79 + □ = 125

C. 79 - □ = 125

D. 125 + 79 = □

VAELL3Bears Access-Based



A class has 79 stars.

They need 125 stars.

How many more stars do they need? Choose the correct equation.

A. □ - 125 = 79

B. 79 + □ = 125

C. 79 - □ = 125

D. 125 + 79 = □

		VAELL3 Bears		Total
		0	1	
Math3.1a	Rarely	76	17	93
	Sometimes	325	116	441
	Almost Always	450	299	749
Total		851	432	1,283

		VAELL3 Bears		Total
		0	1	
Math3.1a	Rarely	64	29	93
	Sometimes	273	168	441
	Almost Always	365	384	749
Total		702	581	1,283

Standard						Access-Based				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
57.026*	1	-1.793	0.176	**0.000**	Constant	45.191*	1	-1.106	0.156	**0.000**

		0.921	0.127	**0.000**	Target			0.763	0.116	**0.000**
35.27*	6	-2.545	0.312	**0.000**	Constant	20.788*	6	-1.478	0.261	**0.000**
		0.375	0.170	**0.027**	Target			0.297	0.156	0.057
		0.323	0.070	**0.000**	Reading			0.203	0.068	**0.003**
		0.087	0.222	0.694	Testwiseness			0.290	0.201	0.148
		0.053	0.050	0.284	Psychosocial			0.021	0.045	0.650
		-0.049	0.192	0.796	Context			-0.153	0.174	0.381
		-0.271	0.182	0.136	Gap			0.130	0.099	0.190
		0.126	0.065	**0.053**	ELL			0.006	0.058	0.918

VAELL3Bears

This is the access-based item where, while the target information in the independent variable is not significantly different from 0, it is very close (p=.057). Reading continues to play an influential role in both items. The other significant variable is ELL status, which impacts the standard but not the access-based item. This suggests that the change in format or lessoning of the amount of language may be beneficial in the access-based item, where the lengthy item in the standard seems to have inhibited some of the newer ELLs from responding correctly. Substantially less language may have been problematic for some students and may be why this target was less clearly defined. However, in differentiating the impact of each item by mathematics ability, it appears that many students from all of the ability levels benefited from the design of the access-based item: 13% of the lowest ability, 12% of those in the middle ability group, and 11% of students at the highest level responded correctly on this item versus it's standard counterpart. This finding may benefit those who are not ELL as well as those who are; it will be interesting to see how poorer readers performed on each of these items.

VAELLN33 Standard VAELLN33 Access-Based

Ms. Cho lives in Hagerstown and has to drive to Salisbury. It is 75 miles from Hagerstown to Baltimore. It is 129 miles from Baltimore to Salisbury. How many miles is it from Hagerstown to Salisbury if Ms. Cho travels through Baltimore?

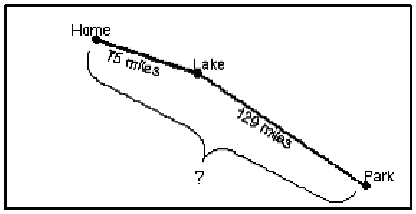
A. 54

B. 64

C. 194

D. 204

Ms. Cho's Trip



How many miles does Ms. Cho drive from home to the park?

A. 54

B. 64

C. 194

D. 204

		VAELLN33		Total
		0	1	
	Rarely	53	33	86
Math3.4a	Sometimes	207	235	442
	Almost Always	313	442	755
Total		573	710	1,283

		VAELLN33		Total
		0	1	
	Rarely	56	30	86
Math3.4a	Sometimes	236	206	442
	Almost Always	321	434	755
Total		613	670	1,283

Standard						**Access-Based**				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
8.938*	1	-0.106	0.147	0.471	Constant	30.739*	1	-0.636	0.149	**0.000**
		0.332	0.111	**0.003**	Target			0.618	0.113	**0.000**
18.113*	6	-0.140	0.256	0.585	Constant	11.365	6	-0.773	0.248	**0.002**
		0.369	0.154	**0.017**	Target			0.373	0.153	**0.015**
		0.053	0.067	0.433	Reading			0.095	0.067	0.159
		0.241	0.198	0.223	Testwiseness			-0.352	0.194	0.070
		0.065	0.044	0.145	Psychosocial			0.026	0.044	0.550
		-0.648	0.176	**0.000**	Context			0.141	0.171	0.410
		0.000	0.168	0.999	Gap			0.079	0.096	0.410
		-0.034	0.057	0.554	ELL			0.043	0.056	0.443

VAELLN33

The target continues to be defined when other variables are added. Interestingly, neither item in this pair indicates that reading made a significant difference on how students performed. However, those who have limited testwiseness skills appear to do somewhat better on the access-based item (p=.07) while in the standard condition, they score lower than those who have these skills. Over the 11 item pairs, students with limited testwiseness skills did better on 5 of the access-based items; in 2 of the items the impact was significantly different from 0 while it was close in another 3. For this item pair, though, context is significantly different from 0 in the standard and negative in its coefficient. Surprisingly, this suggests that, for those students whose teachers said they may have contextual problems, they did better on the standard than on the access-based item. On the other hand, in general, context did not pose a problem for any students on the access-based item which is what is desired. In reviewing the correct response ratios by mathematics ability level, the access-based item was found to be more difficult for all levels and particularly for the middle group (7% less students responded correctly). To complicate matters, review of the access item reveals that the graphic was fuzzy and “home” was spelled incorrectly. While it is possible that these errors could have had a significant impact on how students performed on the access-based item, this item pair is included to make the point that simply reducing language and replacing with a graphic may not always be the best approach in general, although something about this item is appealing to those with limited testwiseness skills. Finally, the lack of impact of reading on either item is interesting and deserves further consideration.

VAELL331 Standard

The third-grade class is going to raise money for the homeless in a shelter. They want to buy each person a hat for five dollars. What is the largest number of hats the third grade class can buy if they have $90? Choose the correct equation.

A. 90 + 5 = □

B. 90 – 5 = □

C. 90 × 5 = □

D. 90 ÷ 5 = □

VAELL331 Access-Based

Ann has $90 to buy books.
Each book costs $5.

How many books can Ann buy?
Choose the correct equation.

A. 90 + 5 = □

B. 90 – 5 = □

C. 90 × 5 = □

D. 90 ÷ 5 = □

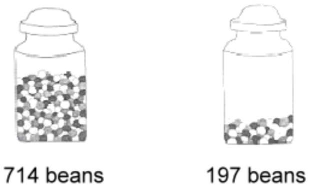
		VAELL331		Total
		0	1	
Math3.1d	Rarely	755	170	925
	Sometimes	244	52	296
	Almost Always	24	38	62
Total		1,023	260	1,283

		VAELL331		Total
		0	1	
Math3.1d	Rarely	764	161	925
	Sometimes	218	78	296
	Almost Always	28	34	62
Total		1,010	273	1,283

Standard						**Access-Based**				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
25.789*	1	-2.250	0.209	**0.000**	Constant	73.899*	1	-2.974	0.237	**0.000**
		0.723	0.147	**0.000**	Target			1.282	0.161	**0.000**
26.646*	6	-2.552	0.361	**0.000**	Constant	20.099*	6	-3.105	0.367	**0.000**
		0.181	0.198	0.360	Target			0.812	0.211	**0.000**
		0.342	0.081	**0.000**	Reading			0.305	0.080	**0.000**
		0.549	0.278	**0.048**	Testwiseness			0.010	0.267	0.971
		0.043	0.059	0.466	Psychosocial			0.087	0.063	0.167
		-0.029	0.225	0.896	Context			-0.243	0.233	0.298
		-0.176	0.218	0.420	Gap			-0.008	0.135	0.952
		-0.106	0.076	0.164	ELL			-0.061	0.079	0.437

VAELL331

This is the other pair where the target is not significantly distinguished from 0 in the standard. Reading continues to be significant in both items. “Five” is not represented numerically which probably has some impact. Additionally, it appears to be a consideration of complexity of language, as the item is no longer than several other standard problems. Testwiseness has a significant impact in the standard item instead, suggesting that students that are fluent in testwiseness skills score better than those with limited skills. No difference in testwiseness is found in this item's access-based counterpart. For students whose teachers say they can demonstrate mastery sometimes in this mathematics skill, they responded correctly to the access-based item 9% more of the time than they did to the standard. There was little change for the other levels of mathematics ability.

VAELL310 Standard	VAELL310 Access-Based
Mrs. Hill and Mr. Smith have jars of candy in their classrooms. There are 714 pieces in Mrs. Hill's jar and 197 pieces in Mr. Smith's jar. How many more pieces of candy are in Mrs. Hill's jar than in Mr. Smith's jar? A. 517 B. 527 C. 627 D. 911	 What is the difference? A. 517 B. 527 C. 627 D. 911

		VAELL310		Total
		0	1	
Math3.4b	Rarely	175	43	218
	Sometimes	393	169	562
	Almost Always	205	298	503
Total		773	510	1,283

		VAELL310		Total
		0	1	
Math3.4b	Rarely	172	46	218
	Sometimes	374	188	562
	Almost Always	223	280	503
Total		769	514	1,283

Standard						**Access-Based**				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
165.936*	1	-2.388	0.190	**0.000**	Constant	112.126*	1	-1.986	0.178	**0.000**
		1.604	0.137	**0.000**	Target			1.280	0.129	**0.000**
61.533*	6	-2.919	0.318	**0.000**	Constant	46.094*	6	-2.411	0.291	**0.000**
		0.786	0.174	**0.000**	Target			0.657	0.169	**0.000**
		0.411	0.073	**0.000**	Reading			0.279	0.070	**0.000**
		-0.102	0.225	0.649	Testwiseness			-0.477	0.216	**0.027**
		0.123	0.051	**0.015**	Psychosocial			0.122	0.050	**0.014**
		0.132	0.194	0.498	Context			0.081	0.191	0.670
		0.068	0.183	0.711	Gap			0.129	0.107	0.227
		-0.009	0.067	0.893	ELL			0.037	0.064	0.557

VAELL310

In addition to reading and the target in each item, the psychosocial variable significantly impacts both items. There does not appear to be an easy explanation for why this latter variable is important for predicting this item. Testwiseness is significantly active in the standard but not in the access-based, indicating that students with limited skills do better in the latter item. However, the fact that reading is still significant, when almost all language has been stripped away in the access item, suggests that the reading variable may be a proxy for some other consideration, beyond psychosocial variable as it has been measured here, and beyond testwiseness. There appears to be little change in percent correct at each level of mathematics knowledge.

VAELL331b Standard	VAELL331b Access-Based
There are 31 railroad cars on a train. Twelve railroad cars are carrying oranges. How many railroad cars are not carrying oranges? A. 43 B. 29 C. 23 D. 19	A class has 31 students. 12 students are boys. How many are not boys? A. 43 B. 29 C. 23 D. 19

		VAELL331b		Total
		0	1	
Math3.4b	Rarely	161	57	218
	Sometimes	335	227	562
	Almost Always	175	328	503
Total		671	612	1,283

		VAELL331b		Total
		0	1	
Math3.4b	Rarely	122	96	218
	Sometimes	200	362	562
	Almost Always	97	406	503
Total		419	864	1,283

Standard						Access-Based				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
143.288*	1	-1.766	0.170	**0.000**	Constant	110.253*	1	-0.694	0.156	**0.000**
		1.417	0.127	**0.000**	Target			1.274	0.127	**0.000**
74.293*	6	-2.639	0.297	**0.000**	Constant	72.079*	6	-1.695	0.267	**0.000**
		0.604	0.165	**0.000**	Target			0.536	0.168	**0.001**
		0.468	0.074	**0.000**	Reading			0.324	0.083	**0.000**
		-0.110	0.209	0.597	Testwiseness			-0.443	0.208	**0.034**
		0.125	0.048	**0.009**	Psychosocial			0.171	0.047	**0.000**
		0.011	0.182	0.952	Context			-0.102	0.182	0.576
		-0.137	0.174	0.434	Gap			0.118	0.102	0.249
		0.095	0.062	0.127	ELL			0.185	0.060	**0.002**

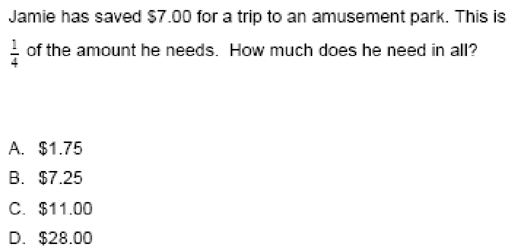
VAELL331b

In this pair, reading, target, and the psychosocial variable impact both items. ELLs had more trouble than their peers on the access-based item. It is probably because of the “not” in the item question, which is difficult for limited English speakers. These students may have noticed this in classroom condition but not under the pressure of large-scale administration. Upon first glance it is unclear why the psychosocial variable is impacting the items in this way, although closer inspection indicates that, over items, a significant psychosocial impact is paired with testwiseness considerations in 4 cases. This occurs in each case, as well as here, that the access-based item where students with limited testwiseness skills do better; the psychosocial variable alone is highlighted in the standard.

Grade 5

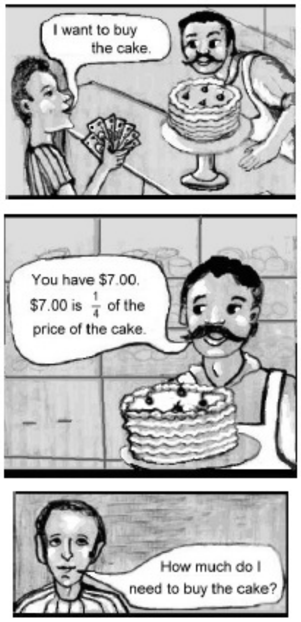
Similar to third grade, 11 of the items on the access-based form and 10 on the standard predicted target mathematics ability at a slope significantly different from 0. However, when other ancillary variables were added to the regressions, target ability continued to be significantly different from 0 for 5 of the standard and 6 of the access-based items. As in grade 3, no variables were found to be pervasive across one administration condition versus the other. Reading was found to be significantly different from 0 in 10 of the 11 pairs. Results from 5 item pairs will be reported here. The remaining items not illustrated here present similar issues as those discussed below.

VAELL5Y2 Standard VAELL5Y2 Access-based



Jamie has saved \$7.00 for a trip to an amusement park. This is $\frac{1}{4}$ of the amount he needs. How much does he need in all?

A. \$1.75
B. \$7.25
C. \$11.00
D. \$28.00


I want to buy the cake.
You have \$7.00. \$7.00 is $\frac{1}{4}$ of the price of the cake.
How much do I need to buy the cake?

A. \$1.75
B. \$7.25
C. \$11.00
D. \$28.00

		VAELL5Y2		Total			VAELL5Y2		Total
		0	1				0	1	
Math5.4	Rarely	223	73	296	Math5.4	Rarely	190	106	296
	Sometimes	378	207	585		Sometimes	311	274	585
	Almost Always	164	180	344		Almost Always	116	228	344
Total		765	460	1,225	Total		617	608	1,225

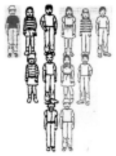
Standard						**Access-based**				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
81.420*	1	-1.544	0.142	**0.000**	Constant	112.034*	1	-1.163	0.133	**0.000**
		0.925	0.107	**0.000**	Target			1.057	0.105	**0.000**
46.347*	6	-2.208	0.304	**0.000**	Constant	66.234*	6	-2.169	0.283	**0.000**
		0.343	0.146	**0.019**	Target			0.290	0.145	**0.046**
		0.307	0.075	**0.000**	Reading			0.486	0.078	**0.000**
		-0.663	0.254	**0.009**	Testwiseness			0.449	0.239	0.061
		0.150	0.051	**0.003**	Psychosocial			0.061	0.047	0.196
		0.347	0.212	0.101	Context			-0.303	0.191	0.113
		-0.063	0.169	0.709	Gap			0.064	0.096	0.504
		0.072	0.066	0.273	ELL			0.043	0.062	0.492

VAELL5Y2

In this pair, the target and reading significantly impact each item. Testwiseness and psychosocial variable are both significant in the standard, and testwiseness is close to significance in the access-based item (p=.06). The standard item favors students with limited testwiseness skills while the alternative appears to favor those with fluent skills. The psychosocial variable, on the other hand, indicates that students with psychosocial concerns do more poorly on the standard item, but that the access-based item does not impact either those with or without psychosocial issues. In reviewing the percent correct by mathematics skill level, it is clear that all groups benefited from the changes in the standard item: both the low and intermediate improved in their correct response by 11% while the highest group improved by 14%. It will be important to understand if there is a particular profile of students who benefited from the particular presentation of the item requirements in the access-based item. It is possible that it eased psychosocial considerations, but the testwiseness change is unsettling.

VAELL505 Standard

The gymnastics class stood in rows to have their team picture taken. The photographer told 2 people to stand in the first row, 4 people to stand in the second row, and 6 people to stand in third row.



The photographer continued the pattern. How many people did the photographer tell to stand in the sixth row?

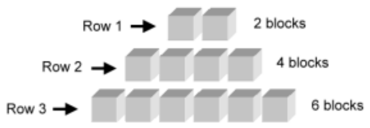
A. 8

B. 10

C. 12

D. 14

VAELL505 Alternative



Continue the pattern.

How many blocks are in row 6?

A. 8

B. 10

C. 12

D. 14

		VAELL505		Total			VAELL505		Total
		0	1				0	1	
Math5.6a	Rarely	54	110	164	Math5.6a	Rarely	62	102	164
	Sometimes	177	366	543		Sometimes	163	380	543
	Almost Always	103	415	518		Almost Always	106	412	518
Total		334	891	1,225	Total		331	894	1,225

Standard						**Alternative**				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
20.960*	1	0.534	0.130	**0.000**	Constant	34.359*	1	0.260	0.147	0.077
		0.497	0.109	**0.000**	Target			0.634	0.110	**0.000**
11.655	6	0.711	0.285	**0.013**	Constant	12.881*	6	0.691	0.272	**0.011**
		0.299	0.155	**0.054**	Target			0.470	0.156	**0.003**
		0.218	0.085	**0.010**	Reading			0.079	0.083	0.340
		-0.202	0.250	0.417	Testwiseness			-0.157	0.242	0.516
		0.060	0.049	0.227	Psychosocial			0.014	0.049	0.782
		-0.079	0.203	0.696	Context			0.209	0.199	0.292
		-0.085	0.176	0.629	Gap			0.196	0.100	**0.051**
		-0.102	0.065	0.116	ELL			-0.162	0.064	**0.011**

VAELL505

For this pair, reading goes from being significant in the standard to not significant in the access-based item while the target ability continues to be significant in both. Reading skills that would appear to be needed for each item seem to change dramatically, particularly as the standard item utilizes more complex language structures in addition to using more language overall. The gap variable suggests, however, that students who don't need accommodations tended to score higher on this item as compared to those who need accommodations. The latter finding is difficult to reconcile with the former. Overall there was little change in correct response by level of mathematics skill between the two items in the pair. The pair both appear to be rather easy items—for this reason, perhaps reading is less of a factor for some reason.

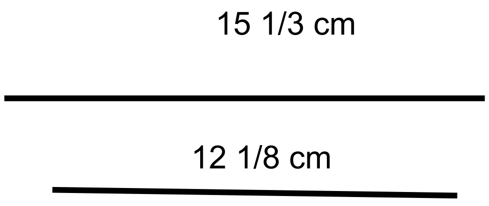
VAELL506 Standard	VAELL506 Access-Based
Mark returned a video 3 days late and paid $6 in late charges. Linda returned a video 5 days late and paid $10. Their friend Eric returned a video 9 days late. How much did Eric pay in late charges? A. $4 B. $8 C. $14 D. $18	3 notebooks cost $6. 5 notebooks cost $10. How much do 9 notebooks cost? A. $4 B. $8 C. $14 D. $18

		VAELL506		Total			VAELL506		Total
		0	1				0	1	
Math5.6a	Rarely	78	86	164	Math5.6a	Rarely	91	73	164
	Sometimes	265	278	543		Sometimes	186	357	543
	Almost Always	220	298	518		Almost Always	116	402	518
Total		563	662	1,225	Total		393	832	1,225

Standard						**Access-Based**				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
2.193	1	0.046	0.120	0.704	Constant	72.778*	1	-0.162	0.125	0.197
		0.141	0.095	0.139	Target			0.890	0.108	**0.000**
5.573	6	-0.085	0.253	0.737	Constant	31.202*	6	-0.565	0.255	**0.026**
		0.086	0.135	0.527	Target			0.374	0.150	**0.013**
		0.102	0.071	0.154	Reading			0.214	0.082	**0.009**
		-0.082	0.221	0.713	Testwiseness			-0.121	0.233	0.604
		0.050	0.045	0.266	Psychosocial			0.048	0.047	0.308
		-0.157	0.182	0.390	Context			0.019	0.193	0.921
		-0.280	0.157	0.075	Gap			0.298	0.096	**0.002**
		0.026	0.057	0.654	ELL			-0.008	0.060	0.899

VAELL506

The target goes from non-significant in the standard to significant in the access-based item; however, reading goes from non-significant in the standard as well to significant in its counterpart. The gap variable in the standard tends to favor the students who receive accommodations (although it is not-significant at p=.08), while it significantly favors those who don't need accommodations in the access-based item. It is unclear as to why the standard does not measure the target mathematics skill. Upon reviewing the correct response tables, they indicate that the access-based item seems to be especially beneficial for students with the highest ability at mastering this mathematics concept—the scores improve by 21% from standard to alternative. Students with some skills also benefit with a 15% increase while students with low levels of skill decrease over items by 8%. For some reason, the graphic here is particularly beneficial to students with more mathematics skill, and these students appear to be especially those who do not need accommodations. Whatever is being measured by the reading variable is probably connected to the profile of these students.

VAELL543 Standard	VAELL543 Access-Based
Andrea had two paper-clip chains. The first chain had a total length of $15\frac{1}{3}$ centimeters (cm), and the second chain had a total length of $12\frac{1}{8}$ centimeters (cm). What was the difference in the lengths of the two chains? A. $3\frac{1}{5}$ cm B. $3\frac{1}{24}$ cm C. $3\frac{2}{11}$ cm D. $3\frac{5}{24}$ cm	Lily drew two lines. The first line was **15 1/3** cm long. The second line was **12 1/8** cm long.  What was the **difference** in the length of these two lines? A. 3 1/5 cm B. 3 1/24 cm C. 3 1/11 cm D. 3 5/24 cm

		VAELL543		Total			VAELL543		Total
		0	1				0	1	
Math5.7a	Rarely	232	73	305	Math5.7a	Rarely	263	42	305
	Sometimes	403	179	582		Sometimes	431	151	582
	Almost Always	159	179	338		Almost Always	189	149	338
Total		794	431	1,225	Total		883	342	1,225

Standard						**Access-Based**				
Incremental x^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental x^2	df	Coefficient	S.E.	Sig.
73.42*	1	-1.612	0.144	**0.000**	Constant	75.236*	1	-2.095	0.163	**0.000**
		0.889	0.108	**0.000**	Target			0.973	0.118	**0.000**
54.545*	6	-1.999	0.302	**0.000**	Constant	40.455*	6	-2.498	0.329	**0.000**
		0.276	0.149	0.064	Target			0.359	0.162	**0.027**
		0.477	0.077	**0.000**	Reading			0.476	0.080	**0.000**
		-0.113	0.262	0.667	Testwiseness			0.086	0.277	0.756
		0.026	0.050	0.604	Psychosocial			0.014	0.054	0.803
		0.508	0.216	**0.019**	Context			-0.142	0.223	0.524
		-0.328	0.174	0.059	Gap			-0.084	0.112	0.454
		-0.080	0.066	0.225	ELL			-0.007	0.072	0.920

VAELL543

Here, the target also goes from non-significant in the standard (although it is close with p=.06) to significant in the access-based item. Reading is significant in both cases. In this pair, as well as the pair reported next, context is significant in the standard but not in its counterpart, and the coefficient in the standard significantly favors those who don't have contextual challenges. It is probable that the idea of paper-clip chains is unfamiliar. The gap variable was not significant in the standard item but close (p=.06), and it favors those who needed accommodations. This gap disappeared in the access-based item, but the reason for the difference in influence is not clear. The correct response tables indicate that all mathematics skill levels did more poorly on the access-based item. Inspection of the item in the booklet indicates that it was not clearly distinguished from another item that was above it, and here, correct response for all levels also declined. No other cause seems forthcoming, and this finding, may suggest the importance of proper spacing in test booklets as well as in items.

VAELL513 Standard	VAELL513 Access-Based
Denise and Nona play on their school's soccer team. Denise scored 4 goals last season. Nona scored 5 goals. Together Denise and Nona scored half of the team's total number of goals last season. How many goals did the team score altogether last season? A. 9 goals B. 13 goals C. 15 goals D. 18 goals	Denise collected **4** cans. Nona collected **5** cans. They have **half** the cans they need. How many **total** cans do they need? A. 9 cans B. 13 cans C. 15 cans D. 18 cans

		VAELL513		
		0	1	Total
Math5.4	Rarely	251	45	296
	Sometimes	400	185	585
	Almost Always	184	160	344
Total		835	390	1,225

		VAELL513		
		0	1	Total
Math5.4	Rarely	222	74	296
	Sometimes	324	261	585
	Almost Always	138	206	344
Total		684	541	1,225

Standard						**Access-Based**				
Incremental χ^2	df	Coefficient	S.E.	Sig.	**Variable**	Incremental χ^2	df	Coefficient	S.E.	Sig.
118.363*	1	-2.135	0.161	**0.000**	Constant	136.458*	1	-1.576	0.142	**0.000**
		1.191	0.117	**0.000**	Target			1.196	0.109	**0.000**
84.632*	6	-3.612	0.385	**0.000**	Constant	91.526*	6	-2.501	0.300	**0.000**
		0.284	0.159	0.073	Target			0.258	0.151	0.087
		0.488	0.079	**0.000**	Reading			0.600	0.080	**0.000**
		-0.345	0.296	0.243	Testwiseness			-0.104	0.251	0.680
		0.099	0.056	0.075	Psychosocial			0.060	0.049	0.221
		0.513	0.242	**0.034**	Context			0.013	0.203	0.947
		-0.045	0.182	0.806	Gap			0.092	0.101	0.360
		0.182	0.077	**0.018**	ELL			0.047	0.066	0.472

VAELL513

In this pair, neither target is significant, although they are close (p=.07 in standard and p=.09 in access-based), and reading is a significant factor in both. Here, context is significant in the standard and favors the students without contextual problems whereas context does not significantly impact the access-based item. Likewise, the ELL status variable is significant in the standard but not in its counterpart, and indicates that non-ELLs tend to perform significantly better on the standard item than do ELLs. Taking the context and ELL variables together, it appears that this population may benefit from this change in items; it is not clear exactly why, however, as soccer is certainly not necessarily contextually unfamiliar. Perhaps girl's school soccer teams are more unusual for ELLs—this will need to be investigated. In reviewing the mathematics skills by ability level, the tables indicate that particularly students at the highest levels benefited from the change with a 13% increase in correct response. While little change occurred for the middle group, a 10% change in the lowest group is also evident. It will be interesting to identify who, besides ELLs are benefiting here.

Discussion

Impact of Language

Overall, reading continued to impact most of the items over and above the influence of the target criterion and regardless of whether they were standard or access-based. This is disheartening for the access-based items, especially when the readers consider that accommodations for poor readers included oral and other similar supports. Two clues, however, suggest that this variable, at least in part, may be actually measuring something else (perhaps particularly in 3rd grade). First, the box plots in grade 3 suggests that the variable is almost acting as a constant across items, even though items differed dramatically in how much reading they required. The second clue is in access-based items where virtually no reading was required (for instance, VAELL310). Complexity of language structures in some standard items seemed to have an influence, although the influence often showed up in variables other than reading (for instance testwiseness or psychosocial). Also, as the 3rd grade box plots indicate, the impact of the reading variable does appear to be more varied in the standard items. In fifth grade, the reading impact continues to be prominent and its influence is somewhat more varied in the large-scale test. However, as compared to grade 3, reading impacts almost all items while the target is only a factor in approximately half. This is disturbing because one could surmise that the test is becoming a measure of whatever the "reading" variable is measuring, reading or otherwise. Finally, on a few items reading doesn't appear to be as influential, but there doesn't appear to be an easily understood rationale for why this is the case. Inspection of the items suggests that the reading load on the access-based items appears to decrease so that may have led to the rather monotonic impact of reading. It is also possible that this may be a proximate variable that reflects the ongoing debates about social-economic-status and related opportunity to learn issues.

VAELL3Bears, in 3rd grade, is an example of a trend that is seen in a number of the items when changes were made from standard to access-based. When standard items have either a complex language structure or when they include, relatively, a great deal of language, the language changes and compensatory support in the access-based items appear to be especially useful. In this item, the only ancillary variable that significantly impacted it besides reading was ELL status, with ELLs scoring better relative to others and to how they behaved on the standard version. Changes in the item were perhaps a little disconcerting for students (as evidenced by the marginal target impact), though we surmise this is probably because the item format is not in a typical large-scale test form. However, it seemed to help a great number of students across mathematics ability levels. It will be interesting to see if and how this conclusion of use holds up for poorer readers in general.

The results for VAELL3Bears and others, across grades, also suggest that when items are not particularly lengthy, and when they have a straightforward sentence or phrasal structure, additional compensatory support or language reduction may not make as much of a difference. Another interesting finding was that ELLs appeared to struggle more with items that required estimation (e.g., VAELL543, VAELLN33, and VAELL310) but had pictures that showed more concrete countable concepts. This may have confused students who may have attempted use the pictures as actual representations as opposed to illustrative examples. Because of the two different administrations however, this is very tentative because other characteristics of the classroom environment may be addressing the compensatory function, at least to a point. The issues of reading and the impact of compensatory supports when language is complex, may begin to explain why "access-basing" constructed response items produced even more clearly useful results. Sophisticated language is often seen in constructed response items and adapting it without changing the intent of the item seems to have benefits similar to those just discussed. Logic would suggest that the issues surrounding reading would dictate that students who have trouble reading and who then have to read less on multiple choice items should perform more validly on these types of items than they would if they must not only read open ended items but write responses as well. However, preliminary work (Kopriva and Lowrey, 1994) suggests that ELLs respond quite differently to distractors than do native speakers. Further, a survey of these students suggests they prefer to explain themselves rather than choose a response. To some extent, the "reading" variable or several of the other ancillary factors may be masking a systematic response of some students to having to choose among particular distractors. This choice is not in play in constructed response. Analyses of the type completed here should help provide more information about these items to help address this possibility.

Impact of Target Measure

As mentioned above, in 5th grade only approximately ½ of the items appear to be measuring skills as defined by the mathematics criterion measure, and that both the teachers and the district said the students had learned. Approximately 1/3 of these regressions impact both items in the pair in this fashion. The analyses here tried to focus especially on items where the target is being measured on at least one of the items in a pair, so that the change in target could be examined and where, in all cases, the impact of other ancillary variables can be ascertained. Therefore many of the "non-target" relevant items are not presented in this paper. However, it is sobering and thought provoking to consider that such a percentage of items are behaving in a way that is so counter to the judgment of the teachers and the district. Of interest is that the constructed response items in the project behaved better than the multiple-choice items. One explanation for the multiple choice findings is that perhaps students are not attending to the content in the items and are simply filling in bubbles.

Impact of Testwiseness

Testwiseness, as it is defined here, seems to be particularly a factor in grade 3 where it behaves differently between the two items for 6 of the 11 pairs. Considerations seem to go both ways to some extent, with 5 of the 11 favoring less skilled students the access-based changes, but the sixth favoring students with fluent abilities. There is a pairing in 4 of these pairs of testwiseness and psychosocial, where psychosocial is evident in both items and where testwiseness is lessoned in access-based. Further, the correlational tables in both grades suggest that testwiseness is paired differentially across administrations with several variables, including ELL status and reading. More work needs to be completed to understand all of these relationships.

Impact of Context Supports

Two of the objectives of the access-based work were to provide more universal contexts and use graphics to provide compensatory support. More work on identifying why some graphics work and other don't is needed, as well as understanding why some groups benefit by certain graphics and other don't. For instance, in VAELL506 in grade 5, why was the format and graphic most useful for students with higher mathematics ability and for those who didn't need accommodations? For third grade, context changes appeared to have limited usefulness, whereas, across items, the 5th grade box plots demonstrate that the context variable favored students with limited contextual skills in both standard and access-based. Probably, in some cases, the item contexts were fine in both sets of items; the plots suggest, however, that the range in scores was certainly greater in standard and sometimes significantly higher for those with a broad range of context knowledge and skill vs. those with limited skills. It will be important to tease out these distinctions as well as to distinguish variations that are occurring because of administration conditions rather than because of the items per se.

Impact of Accommodation Need

. The box plots provide an interesting summary about how the "gap" variable is operating in the items. In both grades the standard condition seems to favor students who need accommodations. This underscores that the classroom environment may be providing useful support for some students and it will be important to understand this more fully. However, the intent, of course, is to not have a gap between those who need accommodations and those who do not. In grade 3 the large-scale test seemed rather close to this objective; in grade 5, on the other hand, the variable clearly favors those who do not need accommodations. Most troubling is that this difference in 5th grade looks significantly different across the two administrations.

Perhaps the accommodations were not useful in the access-based test for this grade. Or perhaps they helped mitigate an even more polarizing impact. It will be important to understand this relationship and continue to address this concern in item building as well as within test administrations.

Conclusion

The first half of this article discusses considerations and guidelines, developed from both theory and practice, which may be useful in building items which are more accessible. It briefly outlines the challenges faced by ELLs and some students with language-based disabilities, and then identifies the multi-dimensional compensatory issues and factors that should be taken into account when developing items for these populations. It is necessary to recognize that not only does the mechanics of this type of item development need to be considered but that the development of skills associated with these mechanics is an absolutely essential part of successful access-based item writing. This is because the access requirements for many of the items are different and no fixed checklist or template will suffice for each situation. As with any complex task, experts are those who know how to successfully navigate among a large amount of possible alternatives, and those who know how to skillfully select and apply the correct choices. Development of these skills entails on-going training and iterative application of the skills and factors with a wide variety of item situations.

The results of the study which applied these item writing principles provide a great deal of information about how test administration conditions and how target and ancillary variables in both items and students interact in complex set ways. Most research focuses on test level results to address some of these issues; it is hoped that this paper will focus readers on how many of these characteristics have their genesis at the item level. Although the empirical results are complex, they provide further support for continued examination of accessibility factors related to item development. It is the intent of the authors that this paper provide a blueprint of the factors and some insight into the process skills writers need to develop when constructing access-based items for this particular population. It is also the hope that the findings provide seeds for future research which will continue to examine and unravel these complex relationships. Recognizing the limitations of the current methodology, namely the differing test conditions, we believe that the explanations and findings noted here provide direction to fulfilling both intents. We recommend that ongoing work continue to build upon many of the questions that have been raised, and that this paper can help shed light on some of the more key concerns and how future work might attend to addressing the issues considered here.

References

Abedi, J., Courtney, M., & Leon, S. (2003). *Research-supported accommodation for English language learners in NAEP* (CSE Tech. Rep. No.586). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education,* 14(3). 219-234.

Achieve, Inc. (June 2004). *Do graduation tests measure up? A closer look at state high school exit exams.* Washington, D.C.: Achieve, Inc.

Bejar, I. I. (2002) Generative testing: from conception to implementation. In S.H. Irvine & P. Kyllonen (Eds), *Item generation for test development* (pp. 199-218). Mahwah, NJ: Lawrence Erlbaum Associates.

Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning and Assessment,* 2(3), 1-29.

Bielinski, J., Thurlow, M.L., Callender, S., & Bolt, S. (2001). *On the road to accountability: Reporting outcomes for students with disabilities* (Technical Report 32). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42*(1), 75-83.

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (Eds.) (1999). *How People Learn: Bridging Research and Practice.* National Research Council. Washington:

National Academy Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 300–396.

Emick, J., Monroe, R., Kopriva, R., & Sprehn, M. (2005). *The culture of the U.S. testing system: A novel response.* Manuscript submitted for publication.

Enright, M. K., Morely, M., & Sheehan, K. M. (2002). *Items by design: The impact of systematic feature variation on item statistical characteristics. GRE Research Report No. 95-15.* Princeton, NJ: Educational Testing Service.

Farr, B. P., & Trumbull, E. (1997). *Assessment Alternatives for Diverse Classrooms* Christopher-Gordon Publishers, Inc. Norwood, MA.

Filippatou, D., & Pumfrey, P. D. (1996). Pictures, titles, reading accuracy, and reading comprehension: A research review (1973-1995). *Educational Research, 38*, 259–291.

Forte, E., & Popham, J. (2006). Wyoming's new accountability tests provide "traffic signals" to help teachers improve instruction. *Harvard Education Letter, March/April.*

Haladyna, T. M., & Shindoll, R. R. (1989). Shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions*, 12, 97-104.

Haertel, E. H., & Wiley, D.E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.) *Test theory for a new generation of tests* (pp.30-59*).* Hillsdale, NJ: Earlbaum, .

Heath, S.B. (1983). *Ways with words*. NewYork: Cambridge University Press.

Heath, S. B. (1989). Oral and literate traditions among Black Americans living in poverty. *American Psychologist, 44*, 367-373.

Hipolito-Delgado, C. (2006, April). *Assessing the Selection Taxonomy for English Language Learners (STELLA).* Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Houang, R.T. (2004). *The holy grail of curriculum measurement: Issues of matching and alignment.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Johnstone, C. J. (2003). *Improving the validity of large-scale tests: Universal design and student performance* (Tech. Rep. No. 37). Minneapolis, MN: National Center on Educational Outcomes.

Kopriva, R.J. (October, 1996). *Variant Methodology for Different Testing Populations.* Presentation for Meta-SCASS meeting, Washington, D.C.

Kopriva, R.J. (2000). *Ensuring Accuracy in Testing for English Language Learners*. Washington, D.C.: Council of Chief State School Officers.

Kopriva, R.J. (2006). *Improving Large-Scale Achievement Tests for English Language Learners.* Manuscript in preparation.

Kopriva, R.J., Cho, M., & Carr, T. (2006). *Application of STELLA System and Relevant Findings*. Presentation at the National Conference on Large Scale Assessment, San Francisco, CA.

Kopriva, R.J., & Lara, J. (1997). Scoring English language learners' papers more accurately. In Y.S. George & V.V. Van Horne (Eds.) *Science education reform for all.* (pp. 77-82). Washington, D.C.: American Association for the Advancement of Science.

Kopriva, R.J., & Lowrey, K. (1994). *Investigations of language sensitive modifications in pilot study of CLAS, the California Assessment System*. Sacramento, CA: Department of Education, California Learning Assessment System Unit.

Kopriva, R.J., & Martin (1998). Validity and Non-Equivalent Issues in Large-Scale Testing. Panel sponsored by the Council of Chief State School Officers.

Kopriva, R., & Mislevy, R. (2005). *Narrative final performance report Valid Assessment of English Language Learners* (PR #R305T010846). U.S. Department of Education.

Kopriva, R., & Winter, P. (2003).*Construct Validity: What Are We Really Measuring* Paper presented at the National Conference on Large-Scale Assessment, San Antonio, TX.

Linn, R. L. (1993). The use of differential item functioning statistics: a discussion of current practice and future implications. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 349-364*).* Hillsdale, NJ: Lawrence Earlbaum.

Malcolm, S. (1991). Equity and excellence through authentic science assessment. In E. Kulm and S. Malcolm (Eds.), *Science Assessment in the Service of Reform* (pp. 313-330.) Washington, D.C.: American Association for the Advancement of Science.

Mann, H., Emick, J., Cho, M., & Kopriva, R., (April, 2006). *Addressing the Validity of Test Score Inferences for English Language Learners with Limited Proficiency Using Language Liaisons and Other Accommodations*. Presentation at the meeting of the American Educational Research Association, San Francisco, CA.

Mislevy, R., & Center for the Study of Assessment Validity and Evaluation (2005). *Access-Based Item Development.* Training Presented to the South Carolina

Department of Education.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Haertel, G., Hamel, L., et al. (2003). *Design Patterns for Assessing Science Inquiry* (Technical Report): SRI International.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (1999). *Evidence-centered assessment design.* Princeton, NJ: Educational Testing Service.

Monroe, R. (2004). *Classroom practices and large-scale assessment.* Unpublished manuscript. University of Maryland College Park.

Pellegrino, J., Baxter, G., & Glaser, R. (1999). Addressing the "Two Disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P.D. Pearson (Eds.), *Review of research in education* (pp. 307-353). Washington, DC: American Educational Research Association.

Popham, J., Farr, R., & Lindquist, M. (2003). *Crafting curricular aims for instructionally supportive assessment.* Wyoming Department of Education. Retrieved January 13th, 2006 from http://www.k12.wy.us/SA/Paws/docs/CraftingCurricula.pdf

Popham, J. Pellegrino, J. Berliner, D., Flick, M., & Kopriva, R. (January, 2006). Technical Advisory Committee Meeting. Wyoming Department of Education, Jackson Hole, WY.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37-

75). Boston: Kluwer.

Samuelsen, K., & Kopriva, R.J. (October, 2004). *Making Sure Students Receive Appropriate Accommodations on Academic Tests.* Presentation at the National Summit Sponsored by the Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students, Washington D.C.

Schmidt, W. H., McKnight, C.C., Houang, R. T., Wang, H.C., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning.* San Francisco: Jossey-Bass, a John Wiley and Sons, Inc. Company.

Shaw, J. (1997). Reflections on performance assessment of English language learners. In B. Farr and E. Trumbull (Eds.), *Assessment Alternatives for Diverse Classrooms* (pp. 334-342). Norwood MA: Christopher-Gordon Publishers.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Sireci, S.G., Li, S., & Scarpati, S. (2003). *The effects of tests accommodations on test performance: A review of the literature.* Commissioned paper by the National Academy of Sciences/National Research Council's Board on Testing and Assessment. Washington, DC: National Research Council.

Siskand, T. (2004). *Application for Achieving Accurate Results for Diverse Learners: Accommodations and Access Enhanced Item Formats for English Language Learners and Students with Disabilities (AARDL).* U.S. Department of Education, Title IV, Subpart 1, Section 6112: Enhanced Assessment Instruments. Washington, D.C.

Solano-Flores, G., Jovanovic, J., Shavelson, R. J., & Bachman, M. (1999).

On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*(3), 293–315.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38(5),* 553-573.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32(2),* 3-13.

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The effects of content, format, and inquiry level on performance on science performance assessment scores. *Applied Measurement in Education, 13(2),* 139-160.

Tindal, G., Health, B., Hollenbeck, P.A., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: an experimental study. *Exceptional Children*, 64, 439-451.

Tindal, G., & Ketterlin-Geller, L.R. (2004). *Research on Mathematics Test Accommodations Relevant to NAEP Testing*. Washington, D.C.: National Assessment Governing Board.

Valverde, Gilbert A. (2005). Curriculum policy seen through high-stakes examinations: Mathematics and biology in a selection of school-leaving examinations from the Middle East and North Africa. *Peabody Journal of Education* 80 (1): 29-55.

Winter, P., Kopriva, R., Chen, S., Emick, J. (in press). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences.*

Wong Fillmore, L., & Snow, C. E. (2000). *What teachers need to know about language. ERIC Clearinghouse on Languages and Linguistics Special Report.* Washington, DC: U.S. Department of Education Office of Educational Research and Improvement. (ERIC Document Reproduction Service Number ED 444379)

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In R. Mitchell (Ed.), *Implementing performance assessments: Promises, problems, challenges* (pp. pp.61-89). Hillsdale, NJ: Erlbaum.