Towards accessible assessments: Promises and limitations of test item adaptations for students with disabilities and English language learners

Abstract

When do item adaptations veer from their intent and, instead of increasing access, modify the construct being measured? This study analyzed early elementary student achievement data from a statewide field test containing both standard and adapted science items. Four student groups were included in this analysis: English language learners, students with learning disabilities, students who are deaf or hard-or-hearing, and a control group of non-IEP native English speakers. This study included a qualitative evaluation of the items by experts with backgrounds in working with students with disabilities. Small sample sizes and restricted numbers of items available made generalizations of the results difficult, but findings suggest unexpected focal group differences in performance on the adapted items. The authors suggest that attention needs to be paid to whom items are adapted for, and steps need to be taken during item development to ensure that changes have the intended effect.

Towards accessible assessments: The promises and limitations of test item adaptations for students with disabilities and English language learners

Large-scale, standardized assessments play a significant role in how elementary and secondary schools measure academic progress in the United States. The past twenty years have seen a movement towards a common assessment experience for students with diverse characteristics, culminating in the development of assessments for the Common Core Standards (Christensen, Lazarus, Crone, & Thurlow, 2008; Rigney, Wiley & Kopriva, 2008; Council on Chief State School Officers, 2012; Thurlow, Lazarus, Thompson, & Robey, 2002). Whereas in previous generations many students were excluded from the large-scale assessment process, when it occurred at all, current policy and practice have expanded to provide meaningful assessments for students with a wide range of abilities and backgrounds. Students with disabilities, including those students who are deaf or hard of hearing (DHH), and students who are English language learners (ELL) have been the focus of strategies to increase the inclusivity of assessment process (e.g., Cawthon, Ho, Patel, Potvin, Trundt, 2009; Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007; Shaftel, Belton-Kocher, Glassnap, & Poggio, 2006; Sireci, Scarpati, & Li, 2005). Unfortunately, "standard" test item formats often have features that confound the scores of students from diverse educational backgrounds. Given the nature of traditional assessment formats (e.g., in written English) and the characteristics of these student groups (e.g., have difficulty reading English at grade level), large-scale, standardized assessments can provide misinformation about the academic proficiencies of students from diverse backgrounds (Elliott & Roach, 2007; Fairbain & Fox, 2009; Kopriva, 2008a; Liu & Anderson, 2008; Shaftel, Yang, Glassnop, Poggio, 2006).

Policy Context of Access and Assessments

In response to the need for a more accessible assessment system, states now implement different forms of assessments. These efforts have been guided by a series of legislation and policy guidance. The latest ESEA (Elementary and Secondary Education Act) reauthorization (NCLB; 2002) originally required all students to participate in a standardized statewide assessment, including English language learners and most students with disabilities, with exceptions for students who were “most significantly disabled.” The Individuals with Disabilities Education Act (IDEA, 1997) and its revision, The Individuals with Disabilities Education Improvement Act (IDEIA; 2004) governs how students with disabilities receive services in K-12 education settings. IDEIA articulates how decisions about accommodations and, where needed, alternate assessment formats for the most significantly disabled, are documented and implemented for each eligible student.

The development of alternate and other assessment formats saw significant changes in the period between 2005 and 2010. Students with disabilities who fell in between the gaps of “most significantly disabled” and “on grade level and able to use assessment accommodations with the general test forms” did not have many accessible assessment options in the early years of NCLB. Although accommodations provided greater access for many ELLs and students with disabilities who were on grade level, these accommodations did not ensure meaningful participation for all such students within the assessment framework. In response to that need, the U.S. Department of Education allowed states to develop two versions of alternate assessments for some students with disabilities: Alternate Assessments Based on Alternate Academic Standards (AA-AAS) and the Alternate Assessments Based on Modified Academic Standards (AA-MAS). An example of one state’s modified assessment for selected students with disabilities, and methods that were used to implement and investigate its measurement traits, can be found in DePascale, 2010a. Even with pending changes in the assessment programs under the Common Core Standards and most likely the next reauthorization of ESEA, test item modifications for some students with disabilities during the test development process are a key part of how items are reviewed and revised to fit the needs of a diverse set of test takers.

Additionally, a number of states have implemented 'simplified language' versions of general test forms at some points during NCLB (for instance, see Pennock-Roman & Rivera, 2011; Kopriva, 2008a). These alternative test forms were meant to measure the same measurement target as those on one or more general test forms, and were seen as an accommodation for some English language learners or students with disabilities who were at grade level in their schooling but did not have the language capacity to meaningfully engage with the language on the test forms administered to the majority of students (Kieffer, Lesaux, Rivera, & Francis, 2009; Sireci, Li, & Scarpati, 2003; DePascale, 2010b). The versions usually included some type of linguistically simplified English text, often with compensatory reliance on relevant graphics in the stem or prompt, the answer options, or both. Linguistic features typically include vocabulary, syntax, discursive structure, and sometimes code or dialect specifications. Writers trained in linguistically simplified item writing were often taught when and how tradeoffs between these features, item format, and various visual elements may be effectively accomplished (for instance, see Abedi & Lord, 2001, Kopriva, 2008b & c). In many cases the linguistically simplified items were interspersed with selected items edited according to universal design principles that were on the general test forms (discussed below). Often these universally designed items were common across the various form versions.

Conceptualization of Access

The policy shift towards inclusive assessment practices has led to a significant rise in the number of research articles measuring the influence of test format on measures of student knowledge and skill. Alternate and modified assessment policies for some students with disabilities often rely on assessments that, at the very least, change the format of standard test items for a number of reasons. Depending on the goals of the assessment, it may also be desirable to change the content rigor or the cognitive complexity of the test items as well. However, the research literature on the effects of test item modifications *does not* typically assume that the revised test items are meant to be easier. The research literature associated with these types of items has tended to be more focused on improving accessibility, even if the target depth of knowledge or cognitive load is modified, with efforts made to maintain item difficulty while changing the features of the test item (for instance, see DePascale, 2010a; Kopriva & Albers, 2013). Other literature is concerned about maintaining the intent of the target knowledge, skills and abilities across versions as well as keeping the cognitive complexity constant. In this case more constrained changes in conditions or formats are considered (see Winter, 2010, for a summary of comparability issues and methods).

In both cases, the assumption in the research literature is that increased item accessibility will lead to fewer barriers to measuring the intent of the items and perhaps increases in student performance assuming students have had access to opportunity to learn the material being tested. However, much of the literature focuses on an item's accessibility as a general construct situated within the item, and does not consider the extent to which an item is *differentially accessible to members of various groups*. Yet assessment policy and the practical considerations of test item development suggests that a single group of items will need to be used with students with a diverse set of characteristics, even within student subgroups such as ELL or students with disabilities. In other words, at this time, the relative dearth of studies that investigate the impact of item format changes on a heterogeneous population of test takers makes it is difficult to defensibly match specific student characteristics with specific aspects of accessible test items. Authors such as Solano-Flores (2013) and Erikan, Roth, Simon, Sandilands, and Lyons-Thomas (2013) have argued for a more nuanced approach to item selection for students with particular characteristics.

Scholars in the field have wrestled with this challenge of diverse student assessment needs and approaches to item development via two main conceptual frameworks for ensuring item access: Universal Design and test item adaptations. The principles of Universal Design emphasize building accessible items ‘from the ground up’ and broadening the accessibility of test items to meet the assessment needs of as many students as possible on the general test forms. From the start of the item development process, items are written with a variety of student groups (e.g., native English speaking students without disabilities, students with disabilities, English Language Learners) in mind. This Universal Design approach guides the work of Kortez and Barton (2003), Liu and Anderson (2008), Johnstone, Thompson, Bottsford-Miller, and Thurlow (2008), among others. The Universal Design principles for assessment center mainly on providing students multiple means of action and expression, including options for how they demonstrate their knowledge. For example, one of the questions listed in the UD guidance is ‘can the item be put into Braille or other tactile form?’. This question stem focuses on whether the item adequately addresses access needs of students who are blind or visually impaired. Johnstone et al. (2008) laid out several approaches for incorporating these Universal Design principles into existing mandated testing structure, particularly for low incidence disability populations, such as students who are blind, deaf or have cognitive disabilities.

In contrast with the Universal Design approach, item adaptations, including versions that maintain the same target construct and cognitive complexity and modifications which entail format changes that may change the nature of the construct, typically make changes to *existing or concurrently developed* test items to make them more accessible for a particular purpose. Item adaptations are common in research when item changes are being systematically investigated and often compared to other versions, and this type of development sometimes occurs in practice, for instance the Texas Education Agency's long standing methods of developing English and Spanish test forms concurrently (see TEA's website for more explanation). Item adaptations that are developed to measure the same target and cognitive complexity are derived from Messick's theory of measurement validity (1989) and on Mislevy's Evidence Centered Design (1994). Abedi et al.'s (for example, 2001; 2011) and Kopriva's (2008b and c) explanations of principled linguistic simplification of text and use of compensatory visuals articulate some of the tenets associated with this type of adaptation that is developed to *not* alter the content and depth of knowledge. Using cognitive load theory, with an emphasis on removing extraneous cognitive load from the test taking process, scholars at Vanderbilt University articulated a sequence of steps in item modification process (e.g., Kettler, Elliott & Beddow, 2009). In order to quantify what it means have an accessible item for students with diverse characteristics, Kettler et al. (2009) developed a Test Accessibility and Modification Inventory-Accessibility Rating Matrix (TAMI-ARM) to assess the accessibility of existing test items. Their inventory breaks items down into several components and analyzes them on a four-point scale ranging from not accessible to maximally accessible. The components include item stimulus, item stem, visuals, item layout, answer choices, and item fairness with an opportunity for the reviewer to explain what features can be modified to increase accessibility scores.

Abedi et al. (2011) used a componential approach in examining the role of cognitive, grammatical, lexical and textual/visual features in grade eight statewide reading assessments. They grouped accessibility features into two major categories: (a) features that are incidental to the measurement and may be changed without threatening the intended target, and (b) features that are crucial to the construct measured and cannot be changed without altering the complexity of the item. Item adaptations that seek to be used for students taking a regular standardized assessment, and not one guided by alternate assessment content, should fall in the former category and not the latter. To address the first category, the researchers indicated select textual and visual features of test items evaluated this study that best discriminated between students with and without disabilities without altering the intent of the content and cognitive load, including font point size, font type, and the number of unnecessary visuals in the test item. Features that fall into the second category where the construct could be threatened are of great relevance to researchers examining alternate assessments (Elliott et al., 2010; Kettler, Rodriguez, Bolt, Elliott, Beddow, & Kurz, 2011). These types of changes may be used to address the needs of students with significant cognitive disabilities or students that require modifications to depth of knowledge as well as an item's textual and visual features.

When do adaptations that don't affect the construct for some students cross the line and impact the target intent of the item for others? The effects of research on item adaptations have been mixed. Sato, Rabinowitz, Gallagher, and Huang (2010) found that students across levels of English proficiency (ranging from beginning to advanced levels) performed better on mathematics test items that had lower levels of linguistic complexity than those with higher levels. Jamal Abedi and his colleagues have focused on linguistic complexity in math and science word problems, a large component of item format structure and an area often targeted to increase accessibility. Abedi, Bailey, Butler, Castellon-Wellington, Leon, & Mirocha (2005) revised items so that they use simplified language. Related adaptations investigated include chunking text/shortening amount of reading (Abedi, Kao, Leon, Sullivan, Herman, Pope, Nambiar, & Mastergeorge, 2008) or adding visual supports such as graphics or tables (Siegel, 2007).

However, others have found minimal to no effect in adapting text in ways such as this, even when the intent is to measure the same construct (e.g. see Keiffer et al., 2009). In fact, Cawthon, Ho, Patel, Potvin, and Trundt (2009) found unintended interactions with disability type. It is also quite possible that the test item adaptation process varies in the extent to which it is successful, including among individual items on a specific assessment. Because general design principles need to be applied to test items that vary in content demand, it is possible that they are not equally effective across an entire battery of items on an assessment, thus diluting possible main effects (Solano-Flores, in press; Carr, 2008).

Measuring Access

Finally, in addition to the different theoretical perspectives on access, there are a variety of ways to measure access. One basic approach to measuring access is to compare student performance on the original form of the item with scores on the adapted form. However, although this may reflect a more quantitative approach, it may not fully explain why student performance changes or does not change with adaptations to different features of an item. In order to gain more explanatory results on effects of item changes, researchers may also utilize a panel of experts to evaluate test items according to several characteristics such as item grammar and vocabulary (Abedi et al., 2011) or inclusion of visual representations (Kettler et al., 2009). These experts may evaluate the items in a ‘round table’ format (Johnstone et al, 2008) or by individually rating items using protocols (Kettler et al, 2009; Liu & Anderson, 2008). Additionally, students may be invited to ‘think aloud’ as they solve the test items to verbalize difficulties that they encounter (Wright, Staehr-Fenner, Moxley, & Carr, 2012; Wright & Kopriva, 2009; Johnstone et al., 2008). In comparison with measuring changes in test scores, these other methods of evaluating access focus more on the *process* of changing the features of the test item and their perceived impact on item accessibility.

Throughout the conversation regarding item accessibility is a concern for students who may not be able to fully demonstrate their knowledge and skill without a change to the general standardized test format. Yet even after the items have been revised to reduce target-extraneous load, the question still remains: How does one know if the items have had their desired effect? Interpreting the results of test item changes on student test scores depends, in great part, on the students who participate in the assessment. While it is critical to look at changes on test performance within the target population, those findings may not be sufficient in understanding whether the item adaptation has changed the construct being measured for others. In fact, the same could be said for all test items on the general forms as well.

Many researchers in the field draw upon the concept of “differential boost” as salient evidence of a “good” item adaptation, all else being equal (e.g., Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Elliott, Kettler, Beddow, & Kurz, 2011). Differential boost refers to the idea that there should be an interaction between the student group and the format of the test (or test item). Differential boost studies administer both sets of items (or both kinds of tests) to two groups of students, such as, students without a particular disability and students with the disability. A “good” item change increases the test scores of those students with the disability more than it increases the test scores of students without the disability. If, in contrast, all scores increase across both groups at about the same rate, the test item adaptation may be interpreted as making the test item easier, or somehow pervasively changing the original intent of the construct being measured. Differential decline would perhaps be the most disappointing effect, with the item creating further barriers to access for some students over others.

Historically, many item adaptation studies have focused on the effects of item changes for an entire student group, such as English Language Learners or students with disabilities. One reason for focusing on a group is that it is easier to make connections between the kinds of item changes, such as simplified language, and the perceived needs of the students. Yet the ELL group, as almost all others, is heterogeneous for a number of reasons. Kopriva, Emick, Hipolito-Delgado, & Cameron (2007) and Solano-Flores's work (for instance, Solano-Flores & Li, 2009) are two of few studies that look at differential access within groups, in this case ELLs. As Kopriva et al. (2007) emphasizes, no item adaptation or accommodation is necessarily relevant or good for all students in a given group. What is important is to know which adaptation or accommodation is useful for which students. This requires knowing more about the students' characteristics and more about the characteristics of the accommodations. It is useful, therefore, to look at the performance of different groups, as well as subgroups that were most likely to benefit from particular item changes. This allows for some greater understanding of how item accessibility efforts differentially affect students with some common struggles (i.e., reading items with complex English text), even if those barriers are due to different etiologies (such as having a learning disability, being deaf or hard of hearing, or being a (hearing) English Language Learner. For English learners, the level of English proficiency is an important intervening variable as is literacy in the home language for students with less English proficiency. For others, such as students who are deaf or hard of hearing, cochlear implants or time spent signing may be important intervening variables.

In sum, mixed results in the research literature on accessible assessments may well reflect the variety of theoretical perspectives on how to increase access for students with diverse characteristics, the types and combination of features emphasized in the item adaptation or item development to begin with, accompanying accommodations and how they are delivered, the tools used to measure accessibility, and the characteristics of the student samples. Yet despite the different approaches to item accessibility, whether from Universal Design or a variety of item linguistic or item modification approaches, both sets of researchers seem to agree on several themes regarding item access. Both approaches highlight the importance of the language properties of the item such as the sentence structure, vocabulary and whether the item includes only relevant item content. This review of access frameworks, measurement, and student characteristics serves to contextualize the research study discussed here.

Overview of the Study

As part of a federally funded grant, a southern state partnered with researchers on an item adaptation project. The original study investigated strategies designed to increase access to test content for English language learners at different levels of English language proficiency (ELP) who may encounter barriers to how items are typically written for their statewide assessment (Carr, 2008). Data for this article came from a 2006 administration of the adapted items that were part of a larger statewide field test collecting data on items developed for a few different purposes. Seasoned state item writers, researchers, and ELL experts concurred that, for ELLs, the adapted versions continued to measure the same content at the same level of cognitive complexity. Findings showed that, in general, ELLs with lower levels of ELP most benefitted from the adaptations as opposed to their performance on the standard items, and that this differential effect lessoned as ELP increased. The difference in item versions for High English proficient ELLs was similar to the differences for the control group (Carr, 2008). While two groups of students with disabilities (students with learning disabilities in reading (LD) and students who are deaf or hard of hearing (DHH)) were not the focus of the original study, data were disaggregated for these two groups as well. Overall, these groups of students with disabilities performed better on the adaptations as opposed to the non-IEP, native English speaking control group. However, subsequent evaluation of the individual items indicated different patterns in their responses as compared to their ELL peers, with differences between LD and DHH students as well (Carr, 2008).

The objective of the current study was to conduct a secondary analysis of the effects of some of the item versions used in the 2006 administration for the two groups of students with disabilities, ELLs as a group, and the control group. To replicate the original evaluation of adapted features, and address effects by group, three research questions guided this investigation: (1) To shed light on the differential patterns of response the original study found across ELLs, LDs, and DHHs, did the secondary qualitative evaluation of item adaptations find that the resulting item features appeared to minimize barriers while retaining the same intent as the standard items, and, if so, how might this be explained? (2) After controlling for student proficiency in science and English language arts, were there interactions within format for different groups? (3) After controlling for student proficiency in science and English language arts, did the groups perform differentially better (or worse) on the standard vs. adapted assessment items?

The analysis focused on the standard and adapted science test item versions for third, fourth, and fifth grade. As a starting point, item accessibility was analyzed using the TAMI-ARM evaluation matrix (Beddow, Kettler, & Elliott, 2009) as an external check on the effects of the item adaptation process. The study then measured potential differential effects of the standard vs. adapted assessment item formats for each of the study's four student groups, using logistic regression and contrast chi-square significance tests, with independent measures of science and English language arts (ELA) achievement scores included as covariates. In order to further account for individual student differences on characteristics (beyond group membership) that might further influence test performance, holding science and ELA constant, the test format by student group interaction were analyzed as a measure of potential differential boost as a result of the item changes.

Methodology

Student Demographics

This analysis includes test performance data from 16,369 students in 3rd, 4th, and 5th grades. A summary of the demographics of the overall student population is provided in Table 1; the demographic distributions were relatively equal between each of the three grades. There were a total of 14,239 students in the control group (students without a disability and not ELLs), 679 students were designated as English language learners, 1,357 were students with a learning disability, and 193 students were deaf or hard of hearing. While most field test forms were randomly assigned across the state, ELLs were oversampled on the forms containing the adapted items in order to ensure sample sizes suitable for analyses in the original study. The LD and DHH students were evenly sampled on all field test forms.

In the data used in this study students were nearly evenly split by gender in the Control and ELL groups, but there were twice as many students with LD that were male (*n* = 935) as female (*n* = 422) and three times as many DHH that were male (*n* = 148) than female (*n* = 45). The student population was ethnically diverse, with 6,322 (39%) students identified as Black, 798 Hispanic (5%), 9,070 White (55%), and a small number of students from other backgrounds (*n* = 278, 1%). About half of the students were on free or reduced lunch, including almost all (86%) of the ELL and the majority (70%) of the students with LD. A small number of students in each category were also designated as gifted (too few to be reported according to NAEP restricted data guidelines); for students with disabilities this is considered a "twice exceptional" condition.

Table 1.

Demographics of Student Participants

	Grades 3-5 Overall Demographics (*N* = 16,468)				Total
	Control N = 14,239	ELL N = 679	LD N = 1,357	DHH N = 193	
Ethnicity					
Black	5642	7	602	71	6322
Hispanic	224	557	13	4	798
White	8179	41	733	117	9070
Other	194	74	9	1	278
Gender					
Male	6873	349	935	148	8305
Female	7341	330	422	45	8138
Gifted					
Academic	2349	4	10	4	2367
Artistic	166	4	8	0	178
Both	120	0	0	0	120
School Lunch					
Free	6208	520	842	97	7667
Reduced	1234	64	118	33	1449
Migrant Status	23	31	2	0	56
Science Performance Level					
Below Basic	4711	430	937	132	6210
Basic	5178	176	308	47	5709
Proficient	2312	19	44	8	2383
Advanced	2032	6	24	5	2067
English Language Arts Performance Level**					
Below Basic	1834	289	725	96	2944
Basic	5663	269	501	79	6512
Proficient	5791	70	127	14	6002
Advanced	931	2	0	1	934

**Columns may not add to total *N* due to missing data.

Field Test Forms and Performance Data

Relevant field test forms with adapted and standard items for science, grades 3-5, were provided to the PI. These items were used in the evaluation activity. Performance data of students participating in the 2006 statewide field test administration in elementary science were compiled and delivered. Besides demographic data, and ELL and SwD status, the state also sent the state's science and ELA scale scores from the 2006 statewide assessment for those students whose field test performance data were part of this study, as well as their placement in one of the four state's performance standards levels, below basic, basic, proficient and advanced.

For science, approximately a third (30%) of Control group students were either proficient or advanced, whereas 4% of ELL, 5% of students with LD, and 7% of DHH students met those criteria. For English Language Arts, 47% of Control group students were either proficient or advanced, whereas 11% of ELL, 9% of LD, and 8% of DHH met those criteria. There are two interpretations that are relevant here: the first is that one might expect higher performance from students in the Control group on the statewide tests than students from the three comparison groups. This is not surprising, but gives a sense of scope to the achievement gap between students without disabilities and their peers. The second meaningful reading of these scores is that there are not wide differences between the ELL, LD, and DHH groups. Although there are very different etiologies for difficulties in reading English, the proficiency scores indicate that they may have similar degree of difficulties with tests in written English.

Item Development, Scaling, and DIF Analyses

Although the current study was conducted as a secondary data analysis, it is helpful to understand the original item and test development, scaling, and DIF evaluation processes. Assessment content specialists, EL specialists in the state assessment unit, and researchers from the University of Maryland, participated in this development using items the state had recently received from a contractor that appeared to have features they believed would pose barriers for ELLs. The objective for this project was to develop the adapted items in such a way that neither the target content nor the complexity level of the standard items would be changed. Using the standard items as a starting point, eventually 201 adapted item versions in four subject areas (language arts, math, science, and social studies), for six grade levels (3rd-8th grades) were completed, or about 9 items per subject/grade. .

Item Development. The adapted item development was a multi-step process. Initially, a two-way training session in the State Department of Education was convened. Researchers presented an overview of the item adaptation methods and training materials to guide the process, and state staff explained the constraints and approach they took to develop items as per their regulations. Then staff and researchers worked together on a set of items in each of the four content areas to consider what changes might be made that would not alter the intended constructs of the standard items and would fit within the state's guidelines.

For the next few weeks, state staff worked separately and then reconvened with researchers to discuss their work, and together they continued to work on new adaptations. From that point on state staff selected standard items to adapt and completed the first round of development. Researchers reviewed and commented on the products and state staff completed the final items. This included contracting with an artist to complete the visuals that had been informally storyboarded to date, and selecting and ordering items on forms. There were regular phone meetings to discuss any outstanding issues. In all, sometimes items went through more than one round of revisions and sometimes item drafts were discarded. While staff internal to the SEA and the research project judged that the completed adaptations were measuring the same intended content and cognitive complexity as the original standard items, there were no external reviews undertaken.

The types of changes made to these items included linguistic semantic and structural techniques to lessen the language load, visual and graphic additions or changes, formatting, and other techniques to provide increased access to the item without altering the construct being measured. For example, the adapted version might incorporate pictures or diagrams into the stem and/or as answer choices (with captions). No new items were created in this process, focusing instead on the item targets already addressed in the standard items. Field test forms associated with this project included both standard and adapted items, where the standard item for one item pair was placed on one form and the adapted item was placed on another.

Scaling. After the data collection in winter 2006, both sets of items were scaled using the Rasch item response theory (IRT) methodology (Taylor, 2007). Scaling parameters from the 2005-2006 administration of the statewide assessment were used to guide the scaling. The experimental items were scaled along with the operational items to compare the fit of the experimental items and the operational items to the model. The difficulty parameters that resulted from running these analyses were also compared across items to determine comparability in difficulty.

Differential Item Functioning. To analyze for differential item functioning or DIF the Mantel-Haenszel (MH) method was used for the following focal groups: students with an individualized education program (IEP) and ELLs at different levels of proficiency. (Note that these are slightly different groups than the analysis groups in the current study in that LD and DHH are not disaggregated). The MH approach to identifying DIF requires that students be matched on ability in order to attribute any remaining variation in item performance to DIF. The matching criterion used here was the student's total test score on the operational items. The operational test score continuum was divided into 10 categories for matching purposes. The approach used here was external matching as the items under study, the experimental items, were not included in the operational test score used for matching. As was expected, with the demographic makeup of the state, that the samples for groups of interest would be small. Small sample sizes proved to be problematic throughout this project, including early in the analysis of differential item functioning. When sample sizes are small, the MH statistic lacks the power necessary to identify differentially functioning items (Parshall & Miller, 1995; Roussos & Stout, 1996). Therefore, while few items were identified using this method it is unclear whether this is because of the sample sizes or lack of DIF.

As such, in addition to using the MH statistic to flag item problems, item characteristic curves from the various versions were compared graphically. Finally, t-tests and p-values were used to test for differences in item difficulty across the versions. Comparisons between the various versions of the experimental items were conducted using t-tests for each two-by-two pairing. Specifically, the t-test compared the difference in item difficulties and whether that difference was statistically significantly different than zero. In the case where the t-test was significant, the interpretation is that there is a significant difference in the difficulty of the two items being compared. These comparisons were made for the entire sample as well as at the disaggregated level. In all, the comparisons for the overall sample and those of the disaggregated groups were within acceptable levels.

Both infit and outfit statistics were also examined to determine the fit of the items to the Rasch model. Infit is a weighted mean square that is sensitive to unexpected responses by students when items are closely matched to their ability level. Problems with infit suggest greater threats to validity than problems with outfit as it is difficult to explain why students are performing as expected on items that are matched to their ability. Problems with infit are often more difficult to diagnose than problems with outfit. Alternatively, outfit is an unweighted mean square that is sensitive to unexpected student responses when items are not matched to their ability (i.e., when items are very easy or very hard for them). Items whose mean squares were greater than 1.30 or less than 0.70 were flagged as problematic. When all students were included in the analysis very few items were flagged for infit or outfit. However, when the data were disaggregated to the subpopulations of interest more items were flagged. Generally, though, more items were flagged for outfit than for infit. Levels of infit appeared fairly acceptable.

Evaluation of Item Accessibility

As an external check on the item adaptation process, a number of different methods were employed. We first included a measure of test item accessibility for the Standard and Adapted items. Although the evaluation tool was originally developed to develop and assess item modifications, the purpose of the TAMI-ARM (Beddow, Kettler, & Elliott, 2009) is to analyze tests and test items on the basis of how accessible they are to *all* students (however, in this case, it is not possible for the items to be accessible to all students, particularly those with visual impairment). TAMI-ARM breaks item accessibility into six categories; passage/item stimulus, item stem, visuals, answer choices, page/item layout and fairness. The instrument does not evaluate items to determine if the cognitive load has been retained or not. Within each category, TAMI-ARM gives the criteria that correspond with each level of accessibility. For example, the criteria for a maximally accessible for all test-takers in the Visual category are: included visual(s) are necessary for responding to the item, visual(s) clearly depict the intended image(s) and are as simple as possible, and contain only text that is necessary for responding (Beddow, Kettler, &

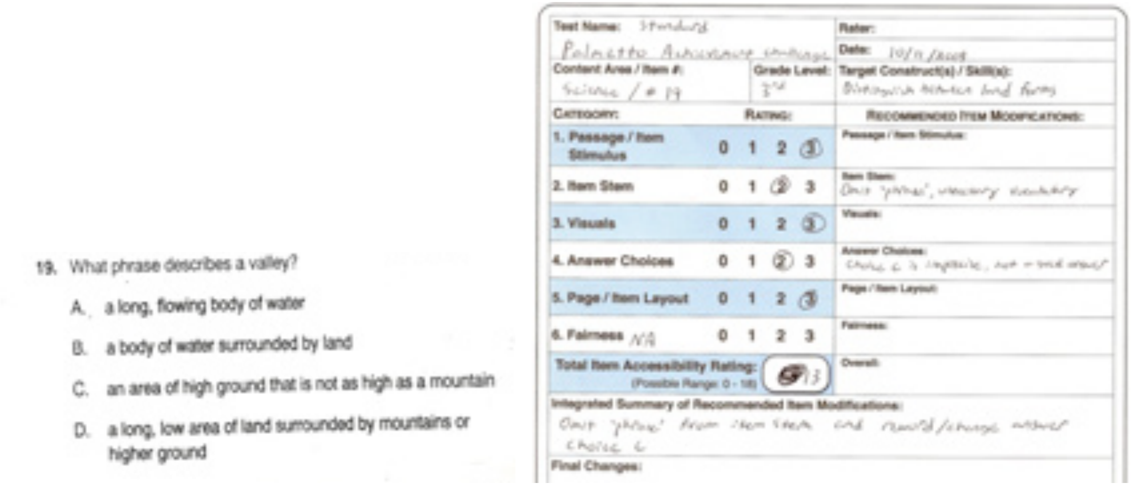
Elliott, 2009). This allows for some consistency across items and an operationalization of what is meant by minimally vs. moderately vs. maximally accessible within each category. Higher scores (e.g., 3-4) indicate items that were perceived as more accessible to learners than lower scores (e.g., 1-2). The Accessibility Rating Matrix extends this idea of range of accessibility to include the rater's perspective as to how many test-takers the item is maximally accessible for (i.e., Some, Most, Nearly All). This extension does not capture specific student groups, but rather, the proportion of all students who might participate in this assessment.

The raters were graduate students in School Psychology and the PI, whose backgrounds have primarily been working with or identifying students who may have particular disabilities. The PI and students working in this lab had experience working with tools from across the field, including those focused on test item linguistic complexity and textual features and identifying features that may be salient for SwDs and others (Cawthon, Kaye, Lockhart, & Beretvas, 2012). After a round of training and reliability checks where coders were in agreement at least 90% of the time, two raters coded the items independently. The project staff coded a total of 27 science item pairs (Standard vs. Adapted), nine pairs for each of the three grades in this study. This analysis used the TAMI-ARM categories and yielded scores from 1 to 4 as well as an overall 'average' score for each item.

In addition to the TAMI-ARM quantitative measures, qualitative comments from project staff were noted for each test item pair. These comments were related to the item's accessibility and reported the particular changes that the original item development had applied during their process. The comments referred to any specific category of item accessibility (i.e., illustrations or page layout) or the items' overall accessibility. Each item had a comment column for both the item's original (Standard) form and the adapted (Adapted) form. For illustrative purposes, we have included two released items that were evaluated by the project staff. Below we include the items in their original form with the TAMI-ARM rating form for the item, followed by the adapted item and its corresponding TAMI-ARM rating form. The first example, the Valley test item, was an item that benefited from the adaptation process and received a higher access score (Figures 1 and 2). The second item, Human Body Test, was one of the 'trouble' items, in which the changes did not improve the access scores despite the efforts to simplify the item format (Figures 3 and 4).

Figure 1

Valley Test Item and TAMI form, Standard Format

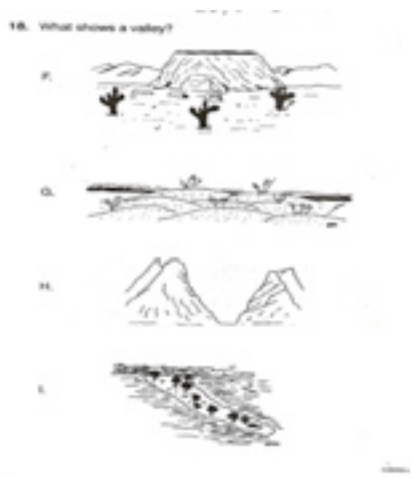


19. What phrase describes a valley?
    - A. a long, flowing body of water
    - B. a body of water surrounded by land
    - C. an area of high ground that is not as high as a mountain
    - D. a long, low area of land surrounded by mountains or higher ground

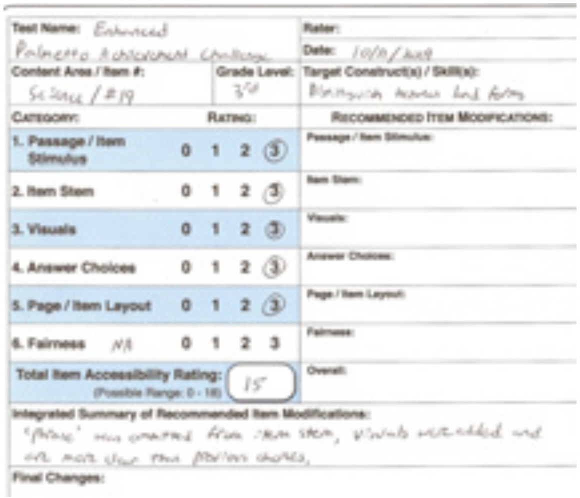
Test Name: Standard Palmetto Achievement Challenge		Rater:
		Date: 10/11/2008
Content Area / Item #: Science / # 19	Grade Level: 3rd	Target Construct(s) / Skill(s): Distinguish between land forms
CATEGORY:	RATING:	RECOMMENDED ITEM MODIFICATIONS:
1. Passage / Item Stimulus	0 1 2 [✓ 3]	Passage / Item Stimulus:
2. Item Stem	0 1 [✓ 2] 3	Item Stem: Omit 'phrase', vocabulary [illegible]
3. Visuals	0 1 2 [✓ 3]	Visuals:
4. Answer Choices	0 1 [✓ 2] 3	Answer Choices: Choice C is [illegible], not a [illegible] answer
5. Page / Item Layout	0 1 2 [✓ 3]	Page / Item Layout:
6. Fairness N/A	0 1 2 3	Fairness:
Total Item Accessibility Rating: (Possible Range: 0 - 18)	13	Overall:
Integrated Summary of Recommended Item Modifications: Omit 'phrase' from item stem and reword/change answer choice C		
Final Changes:		

Figure 2

Valley Test Item and TAMI form, Adapted Format



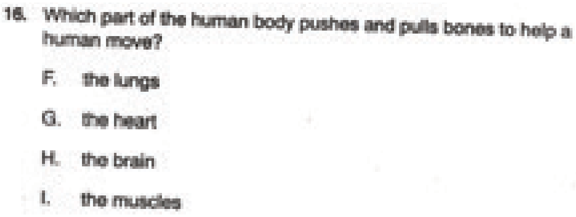
18. What shows a valley?



Test Name: Enhanced Palmetto Achievement Challenge		**Rater:** **Date:** 10/11/2012
Content Area / Item #: Science / #19	**Grade Level:** 3rd	**Target Construct(s) / Skill(s):** Distinguish among land forms
CATEGORY:	**RATING:**	**RECOMMENDED ITEM MODIFICATIONS:**
1. Passage / Item Stimulus	0 1 2 [✓ 3]	Passage / Item Stimulus:
2. Item Stem	0 1 2 [✓ 3]	Item Stem:
3. Visuals	0 1 2 [✓ 3]	Visuals:
4. Answer Choices	0 1 2 [✓ 3]	Answer Choices:
5. Page / Item Layout	0 1 2 [✓ 3]	Page / Item Layout:
6. Fairness N/A	0 1 2 3	Fairness:
Total Item Accessibility Rating: (Possible Range: 0 - 18)	15	Overall:
Integrated Summary of Recommended Item Modifications: 'Picture' was omitted from item stem, visuals were added and are more clear than previous choices.		
Final Changes:		

Figure 3

Human Body Test Item and TAMI form, Standard Format



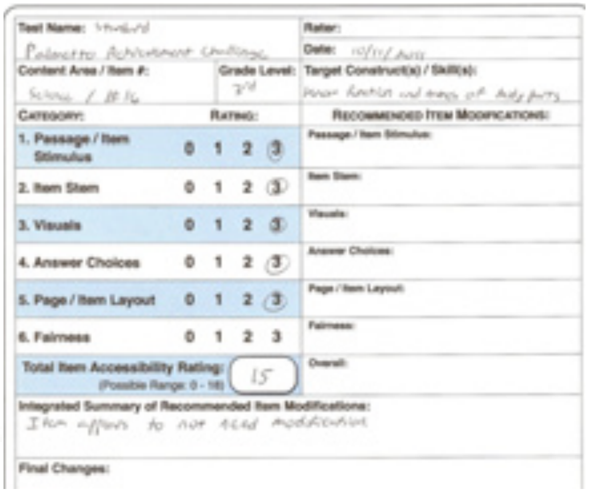
16. Which part of the human body pushes and pulls bones to help a human move?

F. the lungs

G. the heart

H. the brain

I. the muscles

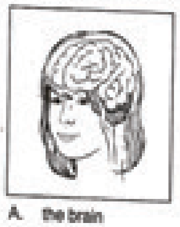


Test Name: Standard Palmetto Achievement Challenge		**Rater:** **Date:** 10/11/2012
Content Area / Item #: Science / #16	**Grade Level:** 3rd	**Target Construct(s) / Skill(s):** Know function and types of body parts
CATEGORY:	**RATING:**	**RECOMMENDED ITEM MODIFICATIONS:**
1. Passage / Item Stimulus	0 1 2 [✓ 3]	Passage / Item Stimulus:
2. Item Stem	0 1 2 [✓ 3]	Item Stem:
3. Visuals	0 1 2 [✓ 3]	Visuals:
4. Answer Choices	0 1 2 [✓ 3]	Answer Choices:
5. Page / Item Layout	0 1 2 [✓ 3]	Page / Item Layout:
6. Fairness	0 1 2 3	Fairness:
Total Item Accessibility Rating: (Possible Range: 0 - 18)	15	Overall:
Integrated Summary of Recommended Item Modifications: Item appears to not need modification		
Final Changes:		

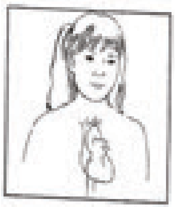
Figure 4

Human Body Test Item and TAMI form, Adapted Format

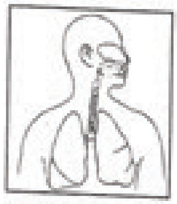
What pushes and pulls bones so a person can move?



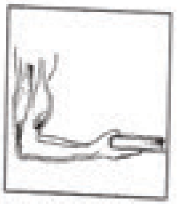
A. the brain



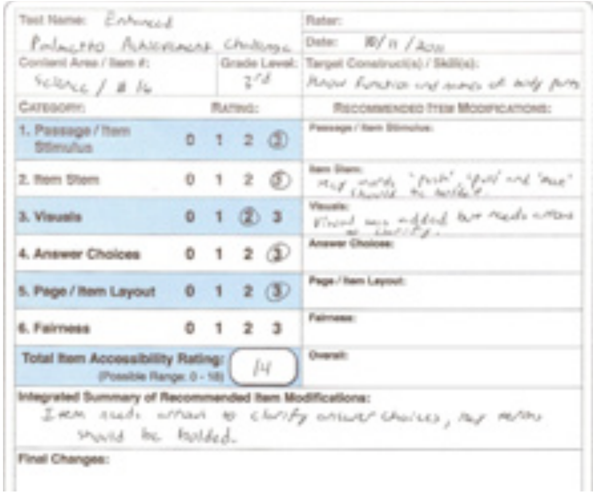
B. the heart



C. the lungs



D. the muscles



Test Name: Enhanced Palmetto Achievement Challenge		Rater: Date: 10/11/2011
Content Area / Item #: Science / # 16	Grade Level: 3rd	Target Construct(s) / Skill(s): Know function and names of body parts
CATEGORY:	RATING:	RECOMMENDED ITEM MODIFICATIONS:
1. Passage / Item Stimulus	0 1 2 [✓ 3]	Passage / Item Stimulus:
2. Item Stem	0 1 2 [✓ 3]	Item Stem: Key words "push", "pull" and "move" should be bolded.
3. Visuals	0 1 [✓ 2] 3	Visuals: Visual was added but needs arrows to clarify.
4. Answer Choices	0 1 2 [✓ 3]	Answer Choices:
5. Page / Item Layout	0 1 2 [✓ 3]	Page / Item Layout:
6. Fairness	0 1 2 3	Fairness:
Total Item Accessibility Rating: (Possible Range: 0 - 18)	14	Overall:
Integrated Summary of Recommended Item Modifications: Item needs arrows to clarify answer choices, key terms should be bolded.		
Final Changes:		

Evaluation Findings. Overall, items scored in the upper range of the TAMI-ARM scale for each of the analyzed components (see summary in Table 2). As a note, the total scores in this study are average scores of the five components analyzed for the AARDL items: Item Stem, Item Stimulus, Visuals, Answer Choices, and Page/Item Layout. Our study did not include the last "Fairness" criterion because it did not fit with the intent of the original project, which was to address the format components of the item and not the overall "fairness" of the item content or approach. The study analysis indicated a significant difference in group means between the original (*M* = 3.62) and adapted items (*M* =3.87), such that the adapted items obtained higher scores on the TAMI-ARM (*t* = 3.9, *p* < .01). This indicated that, overall, changes made to the items may have raised the item accessibility of the science items in grades three through five.

Table 2.

TAMI-ARM Scores on Standard vs. Adapted Items

Grade	Format	Item	Item	Item	Answer	Item	Average
		Stem	Stimulus	Visual	Choices	Layout	Component
		M	*M*	*M*	*M*	*M*	Score
		(SD)	*(SD)*	*(SD)*	*(SD)*	*(SD)*	*M*
							(SD)
3rd (*n* = 9)	Standard	3.22	3.67	2.75	3.33	4	3.47
		(0.83)	(0.5)	(1.23)	(1)	(0)	(0.40)
	Adapted	3.89	4	3.67	4	3.7	3.87
		(0.33)	(0)	(0.5)	(0)	(0.67)	(0.20)
4th (*n* = 9)	Standard	3.78	3.78	3.57	3.22	4	3.68
		(0.44)	(0.67)	(0.53)	(0.83)	(0)	(0.23)
	Adapted	3.89	4	3.89	4	4	3.96
		(0.33)	(0)	(0.33)	(0)	(0)	(0.09)
5th (*n* = 9)	Standard	3.78	3.67	3	3.78	3.78	3.73
		(0.44)	(0.71)	(0)	(0.67)	(0.44)	(0.27)
	Adapted	4	3.78	4	3.78	3.78	3.87
		(0)	(0.67)	(0)	(0.67)	(0.67)	(0.2)
Overall (*N* = 27)	Standard	3.59	3.70	3.10	3.44	3.93	3.63
	Adapted	3.93	3.93	3.85	3.93	3.85	3.90

The results of a t-test indicated that there was a significant difference in group means between the standard (*M* = 3.62) and adapted items (*M* = 3.87) such that the adapted items obtained higher scores on the TAMI ARM (*t* = 3.9, *p* < .01). This indicated that, overall, the changes made to the items may have raised the item accessibility. As indicated previously, the above scores reflect scores in five of the six categories on the TAMI-ARM scale. Items were not rated in the sixth category, fairness, as this was judged to be outside of the purview of the process in the original study.

Even with the overall increase in TAMI-ARM scores between the Standard and Adapted items, some adapted items did not have increased accessibility scores. For example, one item received a 4.0 access score in its original form, but a 3.6 in its changed form. In addition to the

TAMI-ARM, we also added a comment section to help capture some of the changes we saw happening in the adapted item development process (Table 3). The reasons for problematic increases or decreases in item accessibility varied across the items. In Table 3, the labels for the items (i.e., rocks) refer to the topic covered in the item. For example, an item with the label ‘rock’ corresponded to a test question that required the student to demonstrate their knowledge about rocks. For the example above, the corresponding comment in the adapted section was “the oppositional phrasing makes the item ‘tricky’.” Many of the comments related to specific criteria within the TAMI-ARM process, such as issues with the visual layout or the item stem.

Table 3.

Example TAMI-ARM Comments

Item Construct	Standard		Adapted	
	Comment	Category	Comment	Category
Valley	The word ‘phrase’ distracts from the item stem and is not necessary vocabulary to the question.	Item Stem	The omission of ‘phrase’ and the pictographic representation of the answer choices help to clarify the target concept.	Item Stem & Visual
Rocks	The box is visually confusing.	Visual & Page/Item Layout & Item Stimulus	The picture is a better representation of the question than the table.	Visual & Page/Item Layout & Item Stimulus
Electromagnet	‘Toaster’ is not a viable answer choice.	Answer Choice	All answer choices are now equally viable.	Answer Choice
Weathering	The pictures in the answer choices are abysmal.	Visual	The pictures in the answer choices were much improved. But, these pictures may give the test-taker too much information compared to the original.	Visual

The 'trouble item' criterion in this study is a decrease in the TAMI-ARM framework from the Standard version to the Adapted version. Four items met this criterion, two in third grade and two in fifth grade. Because a decrease in accessibility ratings indicate a potential negative effect on item accessibility, these items were removed from subsequent secondary data analyses on changes in student test scores.

Results

Descriptive Data Overall, the test items were relatively difficult for a fair proportion of the students in this sample, as indicated by their mean scores in Table 4. Results of student performance on the two assessment formats, by group, are provided in Table 4. The data represent the average scores for each group and assessment condition (percent of items correct out of the seven items in the third and fifth grades and the nine items in the fourth grade). The unadjusted average scores ranged from 48.8% to 69.9% across the Control group, 38.1% to 51.5% for ELL, 27.1% to 46.8% for LD, and 39.1% to 52.2% for DHH.

Table 4.

Mean Percent Correct by Grade and Test Format Condition: Unadjusted Means

Grade	Test Format	Student Group			
		Control M (SD)	ELL M (SD)	LD M (SD)	DHH M (SD)
3rd Grade	Standard	62.3 (25.4)	48.9 (28.1)	44.4 (25.7)	41.3 (24.2)
	Adapted	62.5 (22.7)	44.6 (20.9)	46.8 (22.2)	47.9 (24.3)
4th Grade	Standard	60.7 (19.4)	42.9 (20.3)	43.1 (21.6)	51.3 (21.5)

	Adapted	69.6 (19.0)	51.5 (22.7)	27.1 (24.4)	52.2 (25.3)
5th Grade	Standard	53.0 (20.6)	38.9 (18.2)	36.6 (19.1)	46.9 (20.8)
	Adapted	48.8 (20.2)	38.1 (18.5)	38.7 (19.8)	39.1 (18.4)

The science scale score (SCI) and the English language arts scale score (ELA) were used as a covariate in this analysis. The mean SCI was 302 (*SD* = 14.73), 403 (*SD* = 18.78), and 502 (*SD* = 20.69) for 3rd, 4th, and 5th graders, respectively. The mean ELA was 310 (*SD* = 20.2), 403 (*SD* = 18.8), and 504 (*SD* = 24.4). Correlations between the two covariates were moderate across grade levels, at .549, .598, and .670 for 3rd, 4th, and 5th grade, respectively. These values are below the level recommended by Mertler and Vanatta (2002) for removal of a covariate ($r < .80$), but are still high enough to warrant some concern as to the additive value of a second covariate in this analysis. Adjusted means are shown in Table 5. The adjusted average scores ranged from 47.8% to 67.5% across the Control group, 44.0% to 64.5% for ELL, 44.5% to 61.1% for LD, and 45.7% to 65.9% for DHH.

Table 5.

Percent Correct by Grade and Test Format Condition: Adjusted Estimated Marginal Means

Grade	Test Format	Student Group							
		Control		ELL		LD		DHH	
		N	*M (SE)*	N	*M (SE)*	N	*M (SE)*	N	*M (SE)*
3rd Grade	Standard	1996	60.0 (.4)	33	64.5 (3.2)	175	57.3 (1.4)	27	56.9 (3.7)
	Adapted	2091	60.1 (.4)	217	58.1 (1.3)	233	61.1 (1.2)	34	65.9 (3.2)
4th Grade	Standard	2463	59.7 (.3)	37	51.1 (2.7)	241	53.7 (1.1)	37	60.2 (2.7)
	Adapted	2522	67.5 (.3)	192	61.9 (1.2)	220	58.6 (1.1)	36	65.9 (2.7)

5th Grade	Standard	2630	52.0 (.4)	25	46.4 (.4)	263	46.6 (1.2)	21	54.1 (4.1)
	Adapted	2535	47.8 (.4)	175	44.0 (1.5)	225	44.5 (3.1)	38	45.7 (3.1)

Interactions and Main Effects

This study utilized the Generalized Linear Model (GZLM) approach to analysis of test performance. We chose the GZLM approach because of the wide range of sub-sample sizes and strong degree of non-normality within the data set. We chose not to address nesting in this design due to absence of information about students' classrooms. Instead, we ran a separate GZLM for each of the three grades in order to maintain the differences in test content between the grades. The models used the percent of targeted items correct as the dependent variable, student group and test item format condition (Standard vs. Adapted) as independent variables, and the science scale score and English language arts scale scores as covariates. The percent correct variable was treated as a categorical variable for each percentage (e.g., 11%, 22%), using the multinomial, ordinal logistic GZLM model. The assumptions behind this model allowed for analysis of the non-normal distribution of the performance on the set of Standard vs. Adapted test items. Furthermore, because of the unequal sample sizes between student groups, many other analysis techniques were not applicable.

For each GZLM model, we first tested for model goodness of fit using the Likelihood ratio chi-square omnibus test and the deviance ratios. The Likelihood ratios chi-square for all three grades were significant at $p < .001$ and the deviance ratios were close to 1.0, indicating good model fit. The overall tests of significance for the GZLMs, with the "trouble" items removed, are provided in Table 6. The main effect of Student Group in grades 4 and 5, which evaluates the general effect of group over both formats, reflects that the percent correct of

Control and the DHH groups are significantly higher than the ELLs and LDs, when the state achievement scores in science and ELA are used as covariates. The significant main effect of Test Format (over groups) in grade 4 and approaching significance in grade 5 indicates that the adapted version is higher in fourth while the reverse may be true in fifth. The interaction term between student group and test format approaches, but does not quite reach significance across all three grades. The results for the Wald Chi Square tests of significance of this interaction were 7.158 (3), $p < .067$ for third grade, 7.556 (3), $p <. 056$ for fourth grade, and 7.505 (3), $p = .057$ for fifth grade. However, subsequent contrasts tests were not significant anywhere, almost certainly because of small sample sizes and perhaps the non-parametric logistic methods that are less sensitive than other tests. Nevertheless, the three accompanying data plots suggest some interesting trends (Figures 5, 6, and 7, for grades 3-5, respectively).

First, there seems to be differences between focal groups for grades 3 and 4. With Deaf and Hard of Hearing students scoring highest in both grades, the largest differences with DHH students in grade 3 are with ELLs and with LDs in grade 4. If the trend of focal group differences is found elsewhere, the fact that adaptations may impact groups in different ways has large implications. Second, in third grade there appears to be a format interaction across groups, which may or may not be significant with larger sample sizes. Here, the Control group appears to have very similar performance on both format types, the ELLs did better on the standard versions of the items, whereas the LD and DHH group did better on the Adapted versions. In contrast, the fourth and fifth grade plots clearly show no format interaction between groups. Third, all three grades show different relationships between the standard and adapted items, with the main effect of test format being significant in grade 4 (adapted higher) and approaching significance for fifth grade (standard higher) as noted above. These figures suggest that, without more data (probably quantitative and qualitative) to support or question expert judgments about the effectiveness of adaptations ‘there are no guarantees’ that adapted items in and of themselves will result in performances that can be defensibly interpreted. Whether or not the various associations in the figures are significant here or not, they suggest that, without more information, assuming that the higher score or the score from the adapted vs. the standard versions is always the more valid representation of student knowledge is premature. This is true of general test forms as well, of course, which is why adapted items were considered in the first place. It seems that the solution for building valid tests for diverse groups is more nuanced.

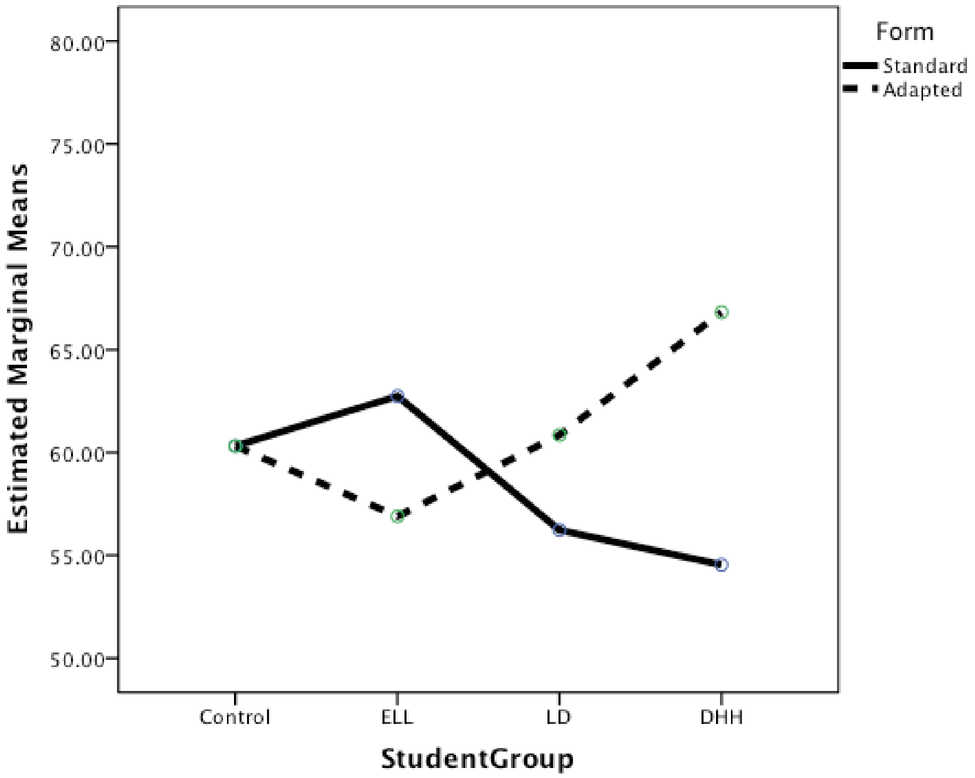
Table 6.

GZLM Statistics Summary

Grade	Main Effects Wald Chi Square (df), Sig.				Interaction
	Student Group (Control, ELL, LD and DHH)	Test Form (Standard vs. Enhanced)	Science Proficiency	English Language Arts Proficiency	Student Group * Test Form
3rd	.944 (3), *p* = .815	.550 (1), *p* = .458	1369.7 (1), *p* <.001	89.23 (1), *p* <.001	7.158 (3), *p* =.067
4th	58.8 (3), *p* <.001	26.6 (1), *p* <.001	1277.5 (1) *p* <.001	78.16 (1) *p* <.001	7.556 (3), *p* =.056
5th	17.99 (3), *p* <.001	3.624 (1), *p* = .057	626.1 (1), *p* <.001	25.63 (1) *p* <.001	7.505 (3), *p* = .057

Figure 5

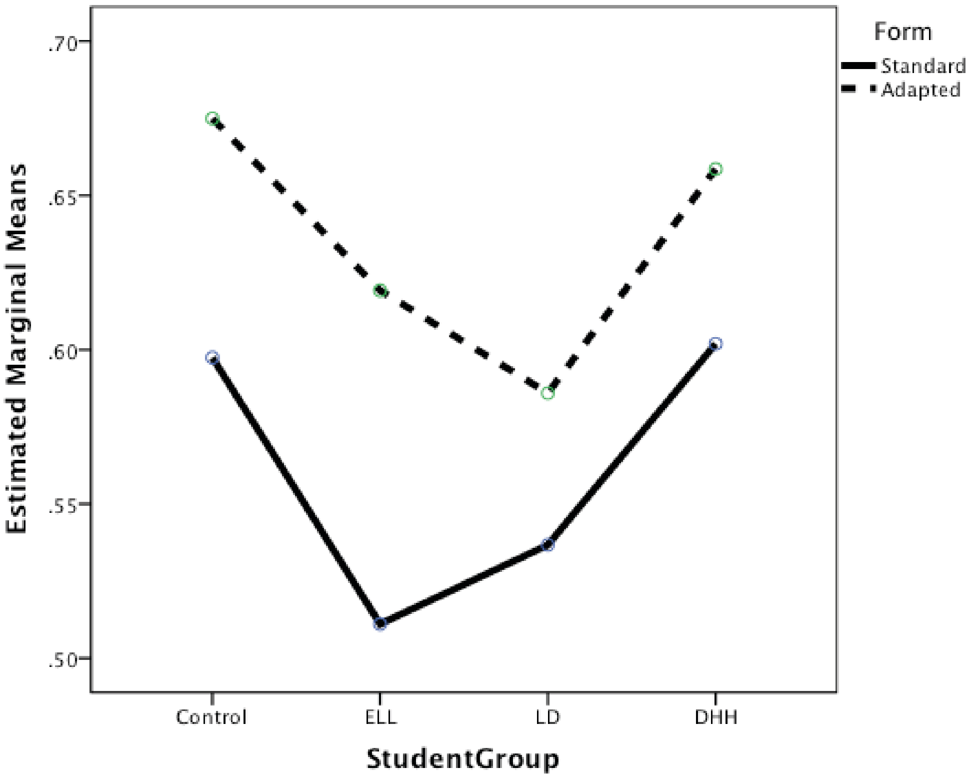
Grade Three Mean Percent Correct



Covariates appearing in the model are evaluated at the following values: Science Scale Score =
301.89, ELA Scale Score = 309.63

Figure 6

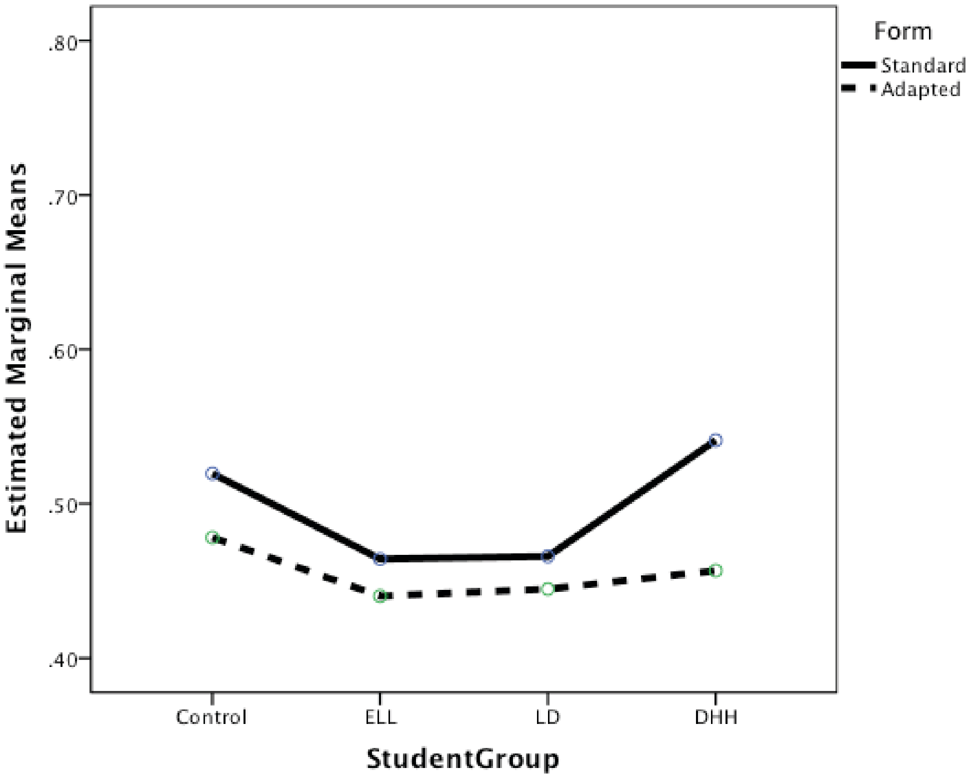
Grade Four Mean Percent Correct



Covariates appearing in the model are evaluated at the following values: ELA Scale Score = 404.42, Science Scale Score = 403.16

Figure 7

Grade Five Mean Percent Correct



Covariates appearing in the model are evaluated at the following values: Science Scale Score = 502.35, ELA Scale Score = 503.65

Discussion

The main purpose of this study was to investigate the effects of science item changes for groups of elementary grade students who might benefit from more accessible test items. Of the 27 item pairs analyzed, 23 of them resulted in an increase in accessibility ratings using the TAMI-ARM, though rater comments indicated that some of the adapted items may have confused the construct being measured or confused some test takers. Of interest is that the raters here commented on features not seen or considered noteworthy by the original team of developers. This is important because whereas state subject, assessment, and ELL specialists, and assessment researchers trained to identify and adapt items for ELLs agreed that the adapted tasks they created as part of the original study were accessible for ELLs and measured the same content at the same levels of cognitive complexity, this study used raters with special education backgrounds and saw issues not considered before.

Significant Group main effects in fourth and fifth grades found that, for the focal groups, DHH students generally score higher across than their ELL and LD peers, implying that the diversity in groups that experts think might benefit from format changes. Significant main effects in Format for the same grades highlighted that, across groups, students performed better on adapted items grade 4 but standard items in grade 5, even though the methodology used to develop them in the original study was the same, and grade differences were not identified as part of the evaluation in this study. While there were no significant interaction contrasts (most likely because of sample size and methodology choice), the graphs in grades 3 and 4 showed substantial differences in the adjusted mean differences of the three focal groups on the adapted items. This trend suggests that adaptations may not be equally effective or useful for all students needing them. Carr (2008) found this in her analysis of the item responses from the three focal groups, and the original study found that within the ELLs, low English proficient students performed better on the adapted items in grade 3, but the reverse was true with high English ELLs (Kopriva & Cameron, 2006).

Implications

There appear to be a few important implications from this study. First, Carr (2008) asked if ELLs, students with learning disabilities and students who are deaf and hard of hearing may focus on different characteristics in the item visuals. While ELLs seem to use the overall picture to get the gist of the question, she suggested that DHH students may focus on more specific qualities within the visuals, particularly when they are pictures of people, and students with learning disabilities seemed bothered more easily by location of different elements in illustrations. Although this study didn't go into that level of detail, the differences in adapted responses by the three groups to sets of items over the different grades seems to suggest that something like this may be occurring. For instance, since the visuals in this study were of mediocre quality (at best) this may have impacted some students more in some groups than others, just as lack of text may have done. Both of these characteristics were noted by the raters on the TAMI-ARM.

Further, using the TAMI-ARM, it appeared that in many cases the standard items were not that inaccessible to begin with, so changes observed in the item adaptations were not considerable when using the accessibility rating scales. Furthermore, features of items that did or did not increase accessibility ratings were not consistent, a concern for this evaluation measure but not for the linguistically simplified methodology under which the adapted items were developed. This finding raises the question of how to measure "access" in the context of standardized test items, an issue currently under discussion in the field. It also raises the question of "adequate access" and how to make decisions about what test items need to be changed, for whom, to what degree, and to make what impact.

Given the two types of development and evaluation matrices from this study and the original, and given the differential findings between focal groups from the GZLM, one conclusion seems clear. That is, more needs to be understood about how various students who all need other modes of information besides English text to comprehend questions or provide their responses, differentially ‘hear’ the variety of semiotic representations used in adaptations. This conclusion seems to be an essential factor in explaining the differences in performance between these three groups. As noted in the literature, it is also a challenge researchers found in smaller scale studies within the ELL and DHH populations, and it appears to be a salient issue for the students with learning disabilities group as well. However, to date, this issue does not seem to have been generally considered in large-scale development.

Second, the issue of modifying the cognitive load or not is an essential consideration. The evaluation here suggested that some of the adaptations may have ‘crossed the line’ in regards to changing the cognitive demand in the items. This investigation has brought up that what might be ‘crossing the line’ for some students may not be for others. Those with little English language may benefit from one word notations instead of a sentence and still retain the target content in the item at the same level of cognitive rigor; however, for others the difference in detail in the adapted versus the standard text may actually alter the construct. It seems that to complete effective adaptations for students who are variously responding to different elements, first item writers must be VERY clear as to what is the intended target and cognitive demand. As adaptations are considered for students with different characteristics, the target and cognitive load must remain constant and immovable for adaptations that are intended to retain the integrity of the original item and it’s target. This is the goal of many accommodations in use today—flexible conditions that do not threaten the validity of the constructs. For adaptations where modifications to cognitive complexity are acceptable, this is accomplished deliberately and not ‘maybe yes or maybe no’. It seems that this calls for increasing the precision of adapting by being cognizant of target and cognitive complexity, and also considering specific needs and characteristics of groups or profiles that adaptations are targeted for.

Third, this study points out the importance of using both relevant empirical and judgment data to determine, a priori if possible, how changes (in this case item changes) are impacting students with particular characteristics. It seems necessary to utilize experts from each pertinent group in the judgment phases of item development, in-depth empirical data such as think-alouds to understand if and how changes are impacting constructs or creating new barriers for each group or subgroup, and large-scale empirical data to be able to generalize effects. In all cases, differential interpretations need to be ruled out and/or minimized so readers and consumers can be confident of the findings.

Limitations

This study carried with it many of the difficulties of a secondary data analysis. There were vastly uneven sample sizes across student groups, with very small prevalence of students in the DHH group. This dataset was reflective of the reality that DHH are a low-incidence population, a fact that poses as a challenge even when using large-scale assessment databases such as those at the state or national level. However, with uneven samples across the groups (Control, ELL, LD, and DHH), analysis procedures were limited to those that could accommodate such different group sizes easily. The GZLM logistic regression method was selected because it is robust even under the conditions of this dataset. Second, this study reviewed only 27 pairs of items across 3 grades, which of course limits the generalizability of the findings overall. A third limitation of this study is that more in depth background data for the SwD students were not collected which could have given the study more information to contextualize the evaluation. Fourth, the covariates chosen in this analysis, the statewide science and ELA achievement test scores, were an attempt to help account for variations within each group. However the state has acknowledged that they completed the original study precisely because they believed that their general test forms may not be accessible ‘enough’, so the statewide scores may already have a built in bias. Additionally, while the scope of the ELA test, particularly for 3rd grade, is partly a test of literacy skills, tests of this subject also include an assessment of literature and other skills only tangentially related to either science or reading or writing. Finally, from a study design perspective, this analysis evaluated item accessibility after the fact, using a tool that did not clearly match up with all of the criteria used for revision in the original process. Occasionally, TAMI-ARM scoring criteria did not fit apparent substantive changes in the item format. In other words, their item scores did not seem to reflect the item’s accessibility because the feature that was changed was not clearly a part of the TAMI-ARM framework. This reflects a limitation of the tool, on one hand, although it allowed for a different point of view that enhanced the analysis in this study on the other.

Future Directions for Research

The scope and quality of future research on the effects of item adaptations will depend on results from item development and revision projects that are available on a large scale, and those that are addressing finer-grain, more particular elements in smaller scale investigations. Rarely are multiple groups included in an analysis at this level, and for low-incidence groups such as DHH, it is very helpful to see how they respond to item adaptations compared to other student groups with different etiologies, but potentially similar challenges with standardized assessments. This study included a set of nine items per grade that were available in both the Standard and Adapted conditions. Full assessments, with enough items to provide the kind of reliability and coverage needed to provide information about how well the test functions in a condition intended to retain the cognitive complexity of the standard items, and perhaps in a modified condition for diverse subgroups of students, is a next step in the research process. It may not be practical to develop and administer such assessments to entire grades at every level, but for meaningful analysis of their impact on low-incidence populations, it would be necessary to have minimum cohorts of students participate in a research initiative.

In depth studies that use think-alouds or other procedures designed to understand how students with diverse needs and strengths are understanding standard items, adapted items with the complexity consistent to the standard, or adapted items with modified complexity will give us clues as to how to encourage access while still allowing for as full an array of depth of knowledge as possible. Additionally, qualitative studies that focus in a fine-grain way on discourse or text/other semiotic representations differences and outcomes would also be helpful in refining why and how item features work or don't and for whom.

Future research might also look at other subject areas, particularly those that combine both content area knowledge and have potentially varying demands on reading skills such as social studies or mathematics. In these more "applied" contexts, the types of adaptations made in the original project, such as adding graphics or reformatting tables of information, might add additional information to the test item without altering the construct being measured. This is in contrast with a reading comprehension test, where the test developer's intention may explicitly be to assess how well students read text and interpret their findings. Adding additional sources of information may reduce the reliance on text and thus make the item easier to solve in ways that are not threatening the target or complexity level. The original project revised test items for multiple subject areas across the elementary and middle grades; and future analyses will focus on how the effects of adaptations may differ between subject areas.

The mixed results of this analysis, as well as the limitations noted above, do point towards a need to consider the fundamental malleability of paper-and-pencil test formats. If best practices in test item development and administration calls for a tailored approach between the test item and the student, it will be very difficult, if not impossible, to create that match without computer-based assessments. Universal Design principles are difficult to implement without some access to multiple modes of presentation and response to test items. While the multiple-choice format of questions is also a potential capacity limitation, especially for lower English ELLs who face language as well as literacy challenges, it is highly unlikely that statewide, standardized assessments will move away from a predominantly multiple-choice approach. What computerized assessment can bring is a mechanism by which interactive, multi-modal, and animated test items can be developed that can, in fact, result in greater accessibility. In a sense, accessible items provide students with tools to represent knowledge, skills, and abilities. The ability to control these features so that they match the needs of individual students with disabilities or English language learners, utilizes these tools in a deliberate manner. Preliminary results from the ONPAR project are promising (Kopriva & Winter, 2013; Kopriva, Gabel, & Cameron, 2009; Wright, 2009). The ONPAR projects focus on providing students with individualized levels of item features, such as audio in English or in one's native language for emerging English language users, roll over graphics, symbols or glossed text, and novel response formats for more challenging constructs.. The initial target groups include ELLs; later ONPAR studies included LD and other students with disabilities that teachers thought might benefit. As piloting continues, template item formats may be revised and adjusted to meet the needs of students with different language and literacy challenges, including perhaps DHH.

References

Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2005). *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives.* CSE Report 663 for Center for Research on Evaluation Standards and Student Testing (CRESST) at the University of California Los Angeles.

Abedi, J. & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25 (4)*, 36-46.

Abedi, J., Kao J.C., Leon, S., Sullivan L., Herman, J.L., Pope, R., Nambiar, V.,& Mastergeorge, A.M. (2008). Exploring factors that affect the accessibility of reading comprehension assessments for students with disabilities: a study of segmented text. CRESST Report 746 for Center for Research on Evaluation Standards and Student Testing (CRESST) at the University of California, Los Angles.

Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2011). Accessible reading assessments for students with disabilites: The role of cognitive, grammatical, lexical, and textual/visual features. CRESST Report 785 for Center of Research on Evaluation Standards and Student Testing (CRESST) at the University of California, Los Angles.

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

Carr, T.G. (2008). Qualitative review of items that worked and didn't work, National Council of Measurement in Education Annual Meeting, New York, NY, March 25-27

Cawthon, S., Kaye, A., Lockhart, L. & Beretvas, S. N. (2012). Effects of linguistic complexity and accommodations on estimates of ability for students with learning disabilities. *Journal of School Psychology, 50,* 293-316.

Cawthon, S., Ho, E., Patel, P., Potvin, D., & Trundt, K. (2009). Multiple constructs and effects of accommodations on accommodated test scores for students with disabilities. *Practical Assessment, Evaluation, and Research 14 (21).* Available online: http://pareonline.net/genpare.asp?wh=0&abt=14.

Christensen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

DePascale, C. (2010a). Modified tests for modified achievement standards: Examining the comparability of scores to the general test. In P.C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations*. Washington, DC: Council of Chief State School Officers.

DePascale, C. (2010b). Evaluating linguistic modifications: An examination of the comparability of a plan English mathematics assessment. In P.C. Winter (Ed.), *Evaluating the comparability of scores from educational achievement test variations*. Washington, DC: Council of Chief State School Officers.

Elliott, S.N., Kettler, R.J., Beddow, P.A., & Kurz, A. (Eds). (2011). *Handbook of Accessible Achievement Tests for All Students: Bridging the Gaps Between Research, Practice, and Policy.* New York: Springer.

Elliott, S.N., Kettler, R.J., Beddow, P.A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M.R., Bolt, D., & Roach, A.T. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children,* 76 (4), 475-495.

Elliott, S.N., & Roach, A.T. (2007). Alternate assessments in students with significant disabilities: Alternative approaches, common technical challenges. *Applied Measurement in Education, 20*, 301-333.

Erikan, K., Roth, W-M, Simon, M., Sandilands, D, & Lyons-Thomas, J (in press). Tests fair for all linguistic minority students? Validity and fairness of measurements for diverse linguistic minority students. In *Applied Measurement in Education*.

Fairbairn, S. & Fox, J. (2009). Inclusive achievement testing for linguistically and culturally diverse test takers: Essential considerations for test developers and test makers. *Educational Measurement: Issues and Practice, 28 (1)*, 10-24.

Fuchs, L., Fuchs, D., Eaton, S., Hamlett, C. L., & Karns, K. (2000). Supplementing teachers' judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29*, 65-85.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum.

Individuals with Disabilities Education Act of 1997 (20 U.S.C. § 1400 et. seq.)

Individuals with Disabilities Education Improvement Act (IDEIA) of 2004, Public Law 108-446.

Johnstone, C.J., Thompson, S.J., Bottsford-Miller, N.A., & Thurlow, M.L. (2008). Universal design and multi-method approaches to item review. *Educational Measurement: Issues and Practice,* 25-36.

Keiffer, M.J., Lesaux, N.R., Rivera, M. & Francis, D.J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79(3), 1168-1201

Kettler, R.J., Elliott, S.N., & Beddow, P.A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education, 84*, 529-551.

Kettler, R.J., Rodriguez, M.R., Bolt, D.M., Elliott, S.N., Beddow, P.A., & Kurz, A. (2011). Modified multiple-choice items for alternate assessments: Reliability, difficulty and differential boost. *Applied Measurement in Education*, 24, 1-25.

Kopriva, R.J. (2008a). *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments.* New York, NY: Routledge Publishers.

Kopriva, R.J. (2008b). Providing a foundation of principled test construction: Maintaining the integrity of the item targets. In R.J. Kopriva, *Improving testing for English language learners.* New York, NY: Routledge Publishers.

Kopriva, R.J. (2008c). Access-based item development. In R. J. Kopriva, *Improving testing for English language learners*. New York, NY: Routledge Publishers.

Kopriva R.J. & Albers, C. (in press). Considerations of achievement testing for students with individual needs. *APA handbook of testing and assessment in psychology*.

Kopriva, R., & Cameron, C. (June, 2007). *Comparing standard and enhanced access items for diverse students: Item analyses in 6 grades & 4 subjects*. Paper presentation at the annual Council of Chief State School Officers Large Scale Assessment Conference: Nashville, TN.

Kopriva, R., Emick, J., Hipolito-Delgado, C., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26 (3)*, 11-20.

Kopriva, R., Gabel, D. & Cameron, C. (2011) *Designing dynamic & interactive assessments for English learners & others which directly measure targeted science constructs.* Paper presented at American Educational Research Association (AERA) Annual Meeting, New Orleans, LA.

Koretz, D., & Barton, K. (2003). *Assessing students with disabilities: Issues and evidence.* Technical report for the Center for the Study of Evaluation, National Center for Research on Evaluation, Standards and Student Testing, Los Angeles.

Liu, K.K. & Anderson, M. (2008). Universal Design considerations for improving student achievement on English language proficiency tests. *Assessment for Effective Intervention, 33* (3) 167-176.

Messick, S. (1989). Validity. In R.L.Linn (Ed.), *Educational measurement (3rd edition)*, 13-103.

Mislevy, R.J. (1994). Evidence and inference in educational measurement. *Psychometrika*, 51, 177-195.

No Child Left Behind Act of 2001, 20 U.S.C. 6301 et seq. (2002).

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small sample conditions. *Journal of Educational Measurement, 32*, 302-316.

Pennock-Roman, M. & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practices*, 30 (3), 10-28.

Rigney, S., Wiley D.E., and Kopriva, R.J. (2008). The past as preparation: Measurement, public policy and implications for access. In R.J. Kopriva, *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments.* New York, NY: Routledge Publishers.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Sato, E., Rabinowitz, S., Gallagher, C. and Huang, C.-W. (2010). *Accommodations for English language learner students: The effect of Linguistic modification of math test item sets.* Washington, DC: NCEE [online]. Available: http://ies.ed.gov/ncee/edlabs/regions/west/pdf/REL_20094079.pdf

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

Shaftel, J., Yang, X., Glasnapp, D., & Poggio, J. (2005). Improving assessment validity for students with disabilities in large-scale assessment programs. *Educational Assessment, 10,* 357-375.

Siegel, M. (2007) Striving for equitable classroom assessments for linguistic minorities: strategies for and effects of revising life science items. *Journal of Research in Science Teaching. 44*(6), 864-881

Sireci, S.G., Li, S. and Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature.* Center for Educational Assessment Research Report no. 485. Amherst, Ma: University of Massachusetts, School of Education.

Sireci, S. G., Scarpati, S. E. & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490.

Solano-Flores, G. (in press). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. In *Applied Measurement in Education*.

Solano-Flores, G. & Li, M. (2009). Language variation and score variation in the testing of English language learners, native Spanish speakers. *Educational Assessment*, 14, 1-15.

Taylor, M. (2007). *Achieving Accurate Results for Diverse Learners: A Study of Differential Item Functioning.* Final report to South Carolina Department of Education under the Enhanced Assessment Instruments Grant Report.

Thurlow, M.L., Lazarus, S., Thompson, S., & Robey, J. (2002). *2001 state policies on assessment participation and accommodations* (Synthesis Report 46). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/NCEO/OnlinePubs/Synthesis46.html

Winter, P.C. (2010). *Evaluating the comparability of results from educational achievement test variations.* Washington, DC: Council of Chief State School Officers.

Wright, L. Staehr-Fenner, D., Moxley, & Carr, T.G. (2012). ONPAR cognitive interview final report summary. At www.onpar.us.

Wright & Kopriva, R.J. (2009). ONPAR elementary and middle school science project, 2006-2009: Cognitive lab report summary. At www.onpar.us.