Targeted Linguistic Modifications of Science Test Items for English Learners

Tracy Noble
TERC
Stephen G. Sireci
Craig S. Wells
University of Massachusetts, Amherst
Rachel Kachchaf
Smarter Balanced Assessment Consortium
Ann Rosebery
TERC
Yang Wang
Education Analytics

Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April, 2016

Abstract

In this study, targeted linguistic and visual modifications of state science test items were used to explore how specific features of test items affect the performance of English Learners (ELs) in Grade 5. A prior statistical analysis of the relationship between the presence of specific linguistic and visual features of test items and the performance of ELs compared to non-ELs was used to select linguistic and visual features for modification. Analyses were carried out at the test, modification type, and individual item level, and were complemented by findings from interviews with ELs. Overall, the results indicate that modifications in which visuals were added to answer choices were the most successful at improving performance of ELs, without altering performance of non-ELs.

Targeted Linguistic Modifications of Science Test Items for English Learners

Introduction

Since the passage of the No Child Left Behind Act (NCLB) in 2002, there has been a great deal of interest from the research and policy communities in mathematics and English language arts testing. Science testing has received less attention, due in part to the relatively late inclusion, in the 2007-2008 school year, of science testing in NCLB-mandated accountability measures. Science tests scores are currently included in the accountability measures that can result in substantial consequences for students, teachers, and schools, including the denial of high school diplomas, the firing of school faculty and staff, and takeover of schools by other entities (Center on Education Policy, 2011; NCLB: Academic Assessment and Local Educational Agency and School Improvement, 2002), and for this reason, science testing needs increased attention from the assessment research community.

Standards-driven education reform legislation such as NCLB is intended to address persistent gaps in test scores between White middle-class students who speak English as a first language and students from historically underserved communities. One such underserved group is English Learners (ELs), who are classified by their schools as Limited English Proficient (LEP), based on a set of criteria that typically include a home language survey and a test of English proficiency, but that can vary from state to state and even across school districts within a given state (Linquanti & Cook, 2015; Sireci & Faulkner-Bond, 2015). Science tests at the state and national levels show persistent test score gaps between ELs and non-ELs, (Caldas, 2013; Chudowsky & Chudowsky, 2010, 2011; Kobe, Chudowsky, & Chudowsky, 2010), similar to those seen on mathematics and English language arts tests (Hemphill & Vannerman, 2011).

As a result of test score gaps between ELs and non-ELs on science tests, the high-stakes consequences of testing differentially affect ELs and their teachers, schools, and communities (Caldas, 2013; Center on Education Policy, 2011). The differences between the test scores of groups of students are typically referred to as achievement gaps, but we believe that they are more accurately described as test score gaps, because achievement is a complex construct best measured using multiple criteria, rather than a single test score (Massachusetts Department of Education, 2006; Noble, Suarez, Rosebery, O'Connor, Warren, & Hudicourt-Barnes, 2012; Pellegrino, Chudowsky, & Glaser, 2001). Despite the stated goals of NCLB to close such test score gaps, they persist on large-scale assessments at the state and national level (Caldas, 2013; Chudowsky & Chudowsky, 2010, 2011; Kober, Chudowsky, & Chudowsky, 2010; Lee & Reeves, 2012), resulting in increasingly substantial consequences for EL students, their teachers, and schools (Huddleston, 2014; Pennsylvania Clearinghouse for Education, March, 2013; Sims, 2013).

Validity and Assessment of ELs

Many researchers have questioned the validity of interpreting ELs' test scores on state tests as measures of ELs' knowledge of a content domain such as science or mathematics, because these tests are written, in English, by and for speakers of English as a first language (Hakuta & Beatty, 2000). Despite a significant body of research demonstrating that students may need six or more years to become fluent in spoken and written English (Hopkins, Thompson, Linquanti, Hakuta, & August, 2013), many states test ELs' science and mathematics knowledge and skills with tests written in English as early as students' first year in U.S. schools (Solórzano, 2008; Wolf, et al., 2008). It is important to note that most states do offer some form of accommodation for ELs— that is, some change to the circumstances of testing—although the

majority of offered accommodations were originally designed for students with disabilities and include, for example, the provision of extra time for testing and small group test administration. It is much rarer for ELs to be offered accommodations such as translated tests or customized glossaries, that directly address the language issues of testing (Rivera, et al., 2006; Wolf, et al., 2008).

The language of NCLB itself highlighted the challenges of creating assessments for ELs in content domains such as science and math "in the language and form most likely to yield accurate data on what students know and can do in academic content areas" and suggests that states should develop yearly student academic assessments in all "the languages other than English that are present in the participating student population" (NCLB: State Plans, 2002). Despite the intentions of the law as written, the underfunding of its mandates to state and local school systems is well-documented (Duncombe, Lukemeyer, & Yinger, 2008; Fritzberg, 2004), and may explain why most states have not had the funds to offer all ELs the option of taking high-stakes science and math tests in their first languages (Wolf et al., 2008).

Following the *Standards for Educational and Psychological Testing*, hereafter referred to as the *Standards*, we use the term validity to refer to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014, p. 11). A series of reports commissioned by the National Research Council question the validity of inferences and actions based on content-based assessments for students who are learning English (August & Hakuta, 1997; Hakuta & Beatty, 2000). These reports caution that a test assessing content knowledge cannot provide valid

information about students' knowledge if "a language barrier prevents the students from demonstrating what they know and can do" (Hakuta & Beatty, 2000, p. 20).

The assessment community has expressed the concern that ELs' scores on high-stakes science tests are affected by construct-irrelevant variance (CIV) (Abedi, 2006; Solano-Flores, 2008). Construct-irrelevant variance (CIV) refers to systematic errors that arise when an assessment actually measures something *other than the construct the test is intended to measure* (AERA, APA, & NCME, 2014; Haladyna & Downing, 2004; Sireci & Faulkner-Bond, 2015). If the scores of EL students on science tests are in part measuring ELs' levels of English proficiency, then English proficiency acts as a source of CIV that threatens the validity of interpreting ELs' test scores as measures of their knowledge and skills in science. The Standards state that in cases when EL students are given a test written in English, "[i]f the test is not intended to also be a measure of the ability to read in English, then test scores do not represent the same construct(s) for examinees who may have poor reading skills, such as limited English proficient test takers, as they do for those who are fully proficient in reading English" (AERA, APA, & NCME, 2014, p. 60).

Sociocultural Perspective on Assessment

A sociocultural perspective on assessment can help to explain how English language proficiency and the language of test items interact to affect the test scores of ELs. Viewed from a sociocultural perspective, test-taking is a language-mediated interaction between the structure and content of test items and students' sense-making resources, including language, culture, and life experiences (Basterra, Trumbull, & Solano-Flores, 2011; Gee, Michaels, & O'Connor, 1992; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003). The interpretation of a test item, like the interpretation of any text, is not fixed and unitary, but emerges from the

interaction of a student with the text of an item in a given context at a particular time. The student brings his or her life experiences and linguistic repertoires of practice (Gutiérrez & Rogoff, 2003) to this interaction with test items, which are written in the specialized Discourse (capitalized, following Gee, (1996)) of science testing. This Discourse includes cultural and linguistic expectations about how students will interpret the items and select their answers. The expectations in the Discourse of testing may or may not be consistent with the Discourses with which ELs, and in some cases non-ELs, are familiar. In this way, mismatches may arise between the expectations of test writers and students' interpretations of test items, leading students to choose incorrect answers even when they have the knowledge and skills targeted by the test items (Noble, Rosebery, Suarez, Warren, & O'Connor, 2014; Noble et al., 2012; Solano-Flores, 2016; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003).

Construct-irrelevant Variance and the Language of Science Test Items

Complementing the studies cited above, a number of large-scale studies of ELs' interactions with test items have provided evidence for the role of the language of test items in contributing to CIV in the test scores of ELs. For example, research on language-focused accommodations such as linguistic simplification of test items, provision of a glossary of non-mathematical or non-scientific words used in test items, and translation of tests into students' first languages, can positively affect the test scores of ELs (Abedi, 2008; Kieffer, Lesaux, Rivera, & Francis, 2009; Li & Suen, 2012; Pennock-Roman & Rivera, 2011; Sireci, Li, & Scarpati, 2003), demonstrating that the language of test items can contribute to CIV for ELs by depressing the scores of ELs who are not provided with these types of accommodations.

A number of other empirical studies of ELs and testing provide evidence for the contribution of specific linguistic features of science and mathematics test items to CIV in test

scores of ELs. Noble, Kachchaf, and Rosebery (2015) reviewed 176 reports on the topic of ELs and assessments in science and mathematics, from 2000-2013. They identified a subset of 9 reports of large-scale studies demonstrating the effects of specific linguistic and visual features of mathematics and science test items on the performance of ELs on those items, and thus the role of the language and other semiotic features of test items in contributing to variation in test scores of ELs. However, only two of these focused specifically on the linguistic features of science test items (Abedi, Courtney, & Leon, 2003; Abedi, Courtney, Leon, & Goldberg, 2005). Given that language is used in different ways in science and mathematics practice and pedagogy (Halliday & Martin, 1993; Schleppegrell, 2007), it is important to investigate how the language of *science* test items affects the performance of ELs.

The language demands of science testing are ever-increasing, and new science tests influenced by the Next Generation Science Standards (NGSS) (Next Generation Science Standards Lead States, 2013) are expected to have an increased emphasis on students' use of language in science tasks (National Research Council, 2014a, 2014b). Researchers and educators alike have expressed concern that such tests may introduce language demands that are not part of the construct the tests are measuring and may differentially affect EL test-takers (Abedi & Linquanti, 2012). In addition, the Every Student Succeeds Act (2016) allows states greater flexibility in the use of assessments for federal reporting purposes, and state departments of education will need to use the findings of research to select from among the range of available assessments those that will yield valid scores for ELs as well as non-ELs.

To support state departments of education in the development and selection of tests as these tests continue to increase in linguistic complexity, we need to identify the specific linguistic and other semiotic features such as visual representations that affect the test

performance of ELs. The goal of the study reported herein is to understand which linguistic features of science test items contribute to CIV for ELs and which features may mitigate these effects and reduce CIV for ELs, with the goal of allowing test item writers and state departments of education to identify and modify these features of their science assessments items, or provide appropriate language-based accommodations, in order to create more valid and fair assessments for all students.

Targeted Linguistic and Visual Modifications

The focus of this article is an experimental study of the effects of targeted linguistic and visual modifications of Grade 5 science test items from one state's high-stakes science test. This study was designed to contribute to the research literature on the assessment of ELs in science, and to identify specific linguistic and visual features of science test items affecting ELs' test performance. This study diverges from previous studies of the process of linguistic modification of test items in science and mathematics (e.g., Abedi, Courtney, & Leon, 2003; Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005; Abedi & Lord, 2001; Sato, Rabinowitz, Gallagher, & Huang, 2010), by creating targeted modifications to test the effects of specific features on the performance of ELs. The work of Solano-Flores, Wang, Kachchaf, Soltero-Gonzalez, and Nguyen-Le (2014) on the effect of adding illustrations to high-stakes science test items on the scores of ELs has used this emerging targeted approach to test item modification.

In a previous study, we identified five linguistic and visual features of Grade 5 multiple-choice science test items from one state's high-stakes science test that were correlated at a statistically significant level with Differential Item Functioning (DIF) for ELs compared to non-ELs (Kachchaf et al., 2015). Three of these features were correlated with *higher* levels of DIF favoring non-ELs over ELs, that is, these three features appeared more often in test items on

which ELs scored lower than non-ELs who scored the same on the rest of the items on the test. These three features are: Forced Comparison (FC), Reference Back (RB), and Low Frequency Non-Technical vocabulary (LFNT). The FC feature occurs in test items requiring students to compare all answer choices in order to select the one that is, in the wording of the question, the *best, most likely*, etc., that is, that is at an extreme on a scale defined by the test item. The RB feature occurs in test items in which the question sentence requires students to refer back to information given earlier in the item. LFNT vocabulary occurs infrequently in 5th grade texts, but does not have a primarily scientific meaning (e.g., the words *hose, repeatedly,* and *unusually* qualified as LFNT vocabulary for 5th grade students).

The other two features were correlated at a statistically significant level with *lower* levels of DIF favoring non-ELs over ELs, that is, they appeared more often on test items that ELs scored well on compared to non-ELs who scored the same on the remaining items on the test. These two features are: the presence of a Visual (e.g., picture, diagram, or table) in the item, and Technical vocabulary, or words with a primarily scientific meaning. These five features are the main focus of the test item modifications undertaken in the study reported on herein. Each feature will be more fully defined in the Method section that follows and is described in detail in Appendix A.

Method

The primary goal of the study reported is to identify the linguistic and visual features of multiple-choice science test items that have the largest impacts on the performance of ELs. To evaluate these impacts, we modified released Grade 5 science test items from one state's science test, administered original and modified versions of these items to ELs and non-ELs and

evaluated the effects using a variety of statistical procedures. Our research questions are as follows:

1) What are the effects of targeted modifications of multiple-choice science test items on the performance of ELs and non-ELs?
2) Which specific linguistic and visual feature modification types and individual item modifications had the greatest impact on the performance of ELs?
3) How do ELs with varying levels of English proficiency respond to the test item modifications?
4) How do non-ELs with varying levels of performance on the state English Language Arts (ELA) test respond to the test item modifications?

In this section, we first describe the assessment and the modifications made to the test items. Next, we describe the student samples. Finally, we describe the statistical analyses conducted to evaluate the effects of the modifications.

State 5th-Grade Assessments

State science assessment. The focus of this work was the multiple-choice science test items from the 5th grade MA state Science and Technology/Engineering (STE) test, the Massachusetts Comprehensive Assessment System (MCAS). The STE MCAS is administered every year to MA students in grade 5 and grade 8, and discipline-specific science MCAS tests are administered to students in high school (e.g., biology, chemistry, physics). These MCAS tests fulfill federal and state accountability requirements for science testing. The 5th grade STE MCAS currently includes 38 multiple-choice and four open-response test items that count towards student scores each year. These 42 test items are common across all 5th grade STE MCAS tests

in any given year[1]. Each correctly answered multiple-choice test item is worth one raw score point, and each correctly answered open-response test item is worth four raw score points.

The STE MCAS is aligned with the MA STE Curriculum Frameworks (Massachusetts Department of Education, 2006) and test items are aligned with the learning standards associated with four different STE strands: Earth and Space Science, Life Science, Physical Sciences, and Technology/Engineering. Approximately 85% of test items each year are associated with the first three strands, and approximately 15% are associated with the Technology/Engineering strand (Massachusetts Department of Elementary and Secondary Education (MA DESE), 2015). The focus of the research reported herein is the multiple-choice test items associated with the Earth and Space Science, Life Science, and Physical Sciences strands[2]. The state currently releases half of the test items annually, and prior to 2009, full STE tests were released to the public each year. MCAS scaled scores range from 200 – 280, and correspond to four achievement levels: Advanced, Proficient, Needs Improvement, and Warning/Failing associated with specified ranges of scaled scores (MA DESE, 2015). All test items used in this study were released 5th grade STE MCAS test items from the years 2004 to 2012.

State English proficiency assessment for ELs. In addition to the 5th grade STE MCAS, we used ELs' scores from the state's English language proficiency assessment, the Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS) test (WIDA, 2014), to investigate the differences between original and modified test scores when ELs were matched for their English proficiency level as measured by the ACCESS

[1] Each test form also includes four test items that vary from test booklet to test booklet and consist of field test and equating items.

[2] We chose to focus our analysis of 5th grade STE MCAS test items on the science test items, and excluded the Technology/Engineering test items, due to significant variation across classrooms and districts in the coverage of Technology/Engineering topics, according to an informal survey of MA elementary school teachers. For this reason, we felt that differences in opportunity to learn technology and engineering would be so pronounced as to interfere with our ability to understand the role of linguistic features in ELs' performance on Technology/Engineering test items.

test. The ACCESS test is administered every year to all ELs in MA who are classified as LEP, and consists of a total of 65 test items, 48 of which are multiple-choice and intended to assess a student's listening and reading comprehension, 4 of which are open response and intended to assess a student's writing proficiency, and 3 of which are scripted face-to-face interview questions administered to assess students' speaking proficiency in English. Students are given final ACCESS scores in listening, speaking, reading, and writing, as well as composite scores that include an overall English language proficiency score[3]. The ACCESS score used in our analysis is the ACCESS reading score, which is provided as a scale score ranging from 100 to 600 for Grade 5, and converted to a level score ranging from 1 to 6.

State English language arts assessment. The state English Language Arts (ELA) assessment, the ELA MCAS, was used to group non-EL students by their level of performance on this assessment. Each year, all 5th grade students in MA, except for first year ELs, are given the ELA MCAS, a test of English language proficiency and reading comprehension, consisting of 36 multiple-choice and four open response test items each year. Scores on this assessment were used to group the non-EL students in our sample into non-ELs who scored at the Proficient level or above on the ELA MCAS and non-ELs who scored below the Proficient level on the ELA MCAS, with the goal of identifying those non-ELs who would be likely to benefit from the linguistic simplification of test items. As for the STE MCAS, the ELA MCAS has a scaled score range from 200 to 280, and to qualify for the Proficient achievement level, students' scaled score must be 240 or above (MA DESE, 2015).

[3] With reading and writing scores each comprising 35% of this overall score and speaking and listening scores each comprising 15% of this overall score.

Test Item Modification

The modification process included several steps: 1) identifying items to modify, 2) developing item modification methods and initial modified test items, 3) coding original and modified items for the required science knowledge and science task, 4) interviewing ELs about the original and modified forms of each item, and 5) updating selected modifications based on the results of the interviews.

Identifying items. We identified an initial set of MCAS STE test items to modify based on a set of criteria including: 1) non-negligible (index >.05 or < -.05 (Dorans & Holland, 1993)) DIF favoring non-ELs over ELs, 2) the presence of one or more of the three features found to be correlated with higher levels of DIF favoring non-ELs (i.e., FC, RB, and LFNT vocabulary), and 3) the lack of one or more of the two item features found to be correlated with lower levels of DIF favoring non-ELs over ELs (Technical vocabulary, Visuals).

Item modification methods. We developed methods to modify each item feature (Visual, FC, RB, and LFNT vocabulary) based upon the definitions of each item feature and the constraints of the test items. Our goal was to remove or reduce each feature found to be associated with lower relative performance for ELs and to add Visuals, a feature correlated with higher relative performance for ELs. We found that very few items could accommodate the addition of Technical vocabulary, and as a result, we did not modify items by adding this feature. A brief description of each of these modification types is given in the sections that follow, and a more detailed description is provided in Appendix A.

Visual modifications. The visual modifications of test items were conducted in one of the following ways: 1) adding a visual (picture, diagram, table, or graph) to the stem of the item, that

is the text of the item exclusive of the four answer choices[4], 2) adding visuals to each of the four answer choices, or 3) maximizing a visual already present in the item, for example, by enhancing an existing visual, such as a table of names of animals, by adding visual images to illustrate each of the animals listed in the table. In all cases, the added visuals were created to illustrate objects rather than processes. The visuals were designed to avoid presenting science content knowledge targeted by the test item.

Forced Comparison modification. For each FC item, we assessed whether the extreme value descriptor (e.g., *best, most, most likely, greatest*) could be removed without changing the scientific content and the correct answer for the item. In cases in which this extreme value descriptor could be removed, we also modified several other aspects common in the FC feature, including the question format *Which of the following,* and the noun or verb associated with the extreme value descriptor in order to clarify the question.

Reference Back modification. For each RB item, in which the question sentence referred back to information in prior sentences (e.g., *these* conditions), we attempted to include in the question sentence the information from prior sentences (e.g., which conditions are referred to). We created a RB modification for the item when this addition could be made without making the question sentence too long to be comprehensible.

Low Frequency Non-Technical vocabulary modification. Following procedures used by Butler, Bailey, Stevens, Huang, and Lord (2004), we identified words in test items that appear infrequently in 5th grade texts (Zeno et al.,1995). We then excluded from this set words that were identified as Technical due to their primary meaning being associated with a scientific discipline or one of the MA state science standards (See Kachchaf et al., 2015 for further details). To

[4] In some cases, an original item with a visual illustration was contrasted with a modified item without the illustration, to demonstrate the effect of the illustration on the performance of ELs.

modify LFNT vocabulary, we replaced each word classified as LFNT with a new word that had the same meaning and was not classified as LFNT. We excluded from modification any LFNT word that, despite being non-Technical was judged by the research team and/or expert coders and advisers to be constitutive of the science content of the test item.

Refinement of modifications through cognitive interviews. As part of our modification development process, we interviewed 52 Grade 5 ELs at 11 different schools in 3 school districts about 32 different test items, each of which was modified in one to four different ways as we refined our modifications. Each student was interviewed about 4-12 items, depending upon the time available for the interview, and each student was interviewed about a combination of test items in their original forms and test items in their modified form. Student participants spoke a total of 10 different languages, including Spanish, Portuguese, Haitian Creole, and Cape Verdean Creole, and were interviewed, whenever possible, by an interviewer fluent in both the student's first language and English or by a combination of an interviewer and a translator fluent in the student's first language and English. We conducted two waves of interviews, allowing time for data analysis and refinement of modifications between the first and second wave. The interviews provided valuable feedback to the modification process, because interview questions were designed to assess students' comprehension of the test item language, allowing us to systematically code students' interpretations of test item language to judge their comprehension of the question asked by the test item.

Through the process of revising test item modifications in response to feedback from student interviews, we found that the modifications leading to the greatest improvements in student performance on test items and comprehension of the questions asked by the test items were: the Visual modification, the combination of the FC and RB modifications, and the

combination of the FC and LFNT modifications. In addition, we developed a fourth category of modifications inspired by the findings of interviews, which we call the *Interview-based* modification. Modifications of the FC, RB, and LFNT features independently were found to be insufficient to improve student performance and students' interpretations of item text.

Interview-based modification. In this modification, multiple features (Visual, FC, RB, and LFNT vocabulary) were modified, and additional aspects of the items, such as specific words, sentence structures, or aspects of the visual illustrations, were also modified, in response to student interview data. For example, in one test item, the word *damp* was critical to comprehending the test item, but was unknown to a number of the students we interviewed. This word had not been coded as LFNT vocabulary, and thus had not been previously identified for modification, but was modified in the Interview-based modification of this test item.

Final modification set. Once the selection of modifications was complete, our final set of item modifications included four types: 1) FC and RB, 2) LFNT and FC, 3)Visual, and 4) Interview-based modifications (See Appendix A for more details).

Coding science knowledge and task. We coded each original and modified test item for (a) the science knowledge needed to answer the item (e.g., opposite poles of magnets attract), and (b) the science task required by the item (e.g., identify the effect of a given cause), based upon the standard associated with each test item and professional judgment. Two experienced science coordinators coded all original versions of the test items prior to coding all modified versions. In cases in which either the science knowledge or task had been altered in the modification, we revised the modification until the two were equivalent. We also asked three scientists (experts in Life Science, Earth and Space Science, and Physical Sciences) to each review all original and modified versions of items in their respective area of expertise to verify

that the science knowledge and skills had not been altered in the modified version. Finally, a panel of experts in linguistics, STE assessment design, and EL education reviewed our original and modified test items to judge the effectiveness of the modifications and their consistency with the modification methods.

Participants

Four public, urban MA school districts participated in the study. Districts were recruited based upon their size and the proportion of ELs in the district. All participating districts had larger percentages of ELs and larger percentages of students receiving free or reduced lunch than in the state as a whole. District directors of STEM and directors of EL/Bilingual education programs were contacted about the study. Participation in the study was not possible in some districts which had initially expressed interest in participating due to the pressures of PARCC mathematics and ELA field-testing in their districts during the year of the study. In two districts, all schools serving 5th grade students participated in the study. In two districts, principal volunteers were sought by the district, and two principals in one district and three principals in another district chose to have all of their 5th grade classrooms participate in the study. In three of the districts, students and parents could choose to opt out of the study. In the fourth district, the assessment was given as one of the district's benchmark assessments and for this reason, students did not have a similar opt-out option. A total of 2359 students at 39 schools in four districts participated in the study.

We retrieved the state demographic data for each participating student. We excluded from the final data set the following students: 27 students for whom we could not locate demographic data in the state data set, 429 students who were receiving Special Education services at the time of testing, four students who used word-to-word bilingual dictionaries during

testing (not part of our testing protocol), and six students who either did not complete at least half of the test or did not take the test at the scheduled time. The exclusion of the 466 students described above resulted in the final data set including test data for 1,893 students, including 310 ELs and 1583 non-ELs. The final data set included 991 female students, 902 male students, and 1631 (86% of the sample) students who received free or reduced lunch. The ELs in the sample spoke 17 different first languages, including Spanish (234 students), Vietnamese (18 students), Portuguese (15 students), and Somali (13 students). The majority of the ELs in the sample were in Sheltered English Immersion programs (284 students), with the remaining 26 students either in other bilingual programs or opting out of all EL programs.

Research Design and Procedure

To evaluate differences in students' performance across the original (unmodified) and modified items, we used an experimental design in which both ELs and non-ELs took tests consisting of both original and modified items. Two different test versions were created and randomly assigned to student participants in each school: Test Version A and Test Version B. Each test had twenty experimental test items, consisting of 10 test items in their original form and10 test items in their linguistically modified form. Six anchor items were common to both test forms, yielding a total of 26 items in Test Version A and Test Version B. The anchor items were chosen for their lack of linguistic complexity and negligible levels of DIF favoring non-ELs over ELs. All experimental test items that were presented in their original form in Test Version A, were presented in their modified form in Test Version B, and vice versa. In addition, the modified items were presented in two different counterbalanced orders in each of the two Test Versions, and these two different orders were randomly distributed to students, to control for fatigue effects. The anchor items remained in fixed locations across all tests.

The tests were administered between February 3, 2014 and March 7, 2014 to 2,356 students in 39 schools in four Massachusetts school districts. Tests were administered in one school district by teachers as the district benchmark assessment, and in the other three school districts by teams of trained test administrators who were current and former teachers and school administrators. A maximum of one hour was allowed for test administration, and the overwhelming majority of students finished the test in less than an hour. Students who finished ahead of time were asked to read a previously selected book.

Three students did not complete at least 50% of the test items, and their scores were not included in the full study. A small number of students skipped some items on the test. Four percent of the students (71) did not fill in answers for between one and three test items. Half of one percent of students (8) did not fill in answers for four or more test items, but did complete a majority of the test items. Skipped test items were distributed throughout students' test booklets, rather than being grouped at the end of students' tests. Skipped answers were given a 0 score, as were incorrect answers.

All tests were electronically scored and data were entered into spreadsheets for data analysis. We then used Student Information Management System (SIMS) data provided by the Massachusetts Department of Elementary and Secondary Education (DESE) as part of a data-sharing agreement with the DESE, to match students in our data set with the demographic data collected from schools by the state. We used SPSS to match the two data sets on the students' first and last names and made 84 additional matches by hand, when the spelling of or the order of students' names as written on their test booklets differed from the state data file.

Data Analysis

To evaluate the effects of the linguistic modifications, we conducted analyses at the total test score level, the modification type level, and the individual item level, to address our goal of exploring the effects of specific linguistic and visual features of test items on ELs' performance. At the total test score level, we used analysis of variance (ANOVA) to compare the raw scores of ELs and non-ELs on the original and modified versions of the items. These analyses were also conducted using students' scores on the 6 anchor items as a covariate. In addition, we conducted an analysis on ELs only, using their ACCESS scores as a covariate. For the modification-type analyses, IRT was used to evaluate the differential difficulty of original and modified versions of the items grouped by modification type. For the item-level analyses, we used a differential item functioning procedure based on item response theory (IRT) to evaluate the effects of individual item modifications. Each of these analyses is described in the sections that follow.

Raw score analyses across test forms and EL status

As described earlier, each student answered 10 original items, 10 modified items, and six anchor items. Two 10-item scores were created for each student: one "original item score" equal to the sum of the student's scores on the 10 original items, and one "modified item score" equal to the sum of the student's scores on the 10 modified items. Thus, students who received one set of original items received a different set of items in modified form. We labeled one set of 10 items "Item Set 1" and the other set of 10 items "Item Set 2." Item Set 1 appeared in original form on Test Version A and in modified form on Test Version B. Item Set 2 appeared in original form on Test Version B and in modified form on Test Version A. Table 1 illustrates the structure of the data for the ANOVA analyses.

Table 1. Structure of Data for ANOVA Analyses

Item Form	Dependent Variable

	Item Set 1 Score	Item Set 2 Score
Original	Test Version A	Test Version B
Modified	Test Version B	Test Version A

In our analysis, the Item Set 1 scores of students who saw these items in original form were compared to the Item Set 1 scores of students who saw these items in modified form. That is, when Item Set 1 Score is the dependent variable, Test Version A designates the test with the original versions of these items, and Test Version B designates the test with the modified versions of these items. Similarly, the Item Set 2 scores of students who saw these items in original form (Test Version B) were compared to the Item Set 2 scores of students who saw these items in modified form (Test Version A). In this way, scores on original items were compared to the scores on the same items when they appeared in modified form.

We conducted separate ANOVAs for each Item Set (i.e., Item Set 1 Score, Item Set 2 score), using raw score across the ten items as the dependent variable and Test Version (which designates whether the raw scores were based on original or modified items) and EL status as the independent variables. The raw score dependent variable was calculated by summing up the scores for the 10 items of interest, on each of which a student may score 0 for an incorrect answer choice, or 1 for a correct answer choice. Four sets of ANOVAs were conducted:

1) a two-way ANOVA with EL status and Test Version (designating the original or modified item sets) as the two independent variables,
2) a two-way ANOVA with student group (ELs, non-ELs who scored Below Proficient in ELA, and non-ELs who scored Proficient or Above in ELA) and Test Version as the two independent variables,

3) a two-way ANCOVA with student group and Test Version as the independent variables and with the raw score across the common items as a covariate, and
4) an ANOVA including ELs only, with Test Version as the independent variable and ACCESS Reading score as a covariate.

Item Response Theory and DIF Analyses

In addition to looking at overall differences in mean performance across the student groups on original versus modified items, we were also interested in exploring the effects of specific modification types and individual test item modifications. Such effects could occur at the modification type level or the individual test item level, but be overshadowed at the total test score level. Therefore we used an IRT approach to evaluate groups of items defined by modification type and two different methods for evaluating the effects of individual item modifications using differential item functioning.

Comparison of modification-type TCCs. To evaluate whether modification effects were linked to one of the four specific modification types (i.e., FC and RB, FC and LFNT, Visual, and Interview-based), we used IRT to calibrate all the test items so that "mini-test characteristic curves" (TCCs) based on items of the same modification type could be created. Item difficulty and proficiency scores for examinees were estimated using the 1-parameter logistic (1PL) item response theory (IRT) model,

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$

where P_i is probability of correct answer on item i; θ (theta) is examinee proficiency; b_i is item difficulty parameter; D is a scaling factor equal to 1.7; and e is base of the natural logarithm.

The *b*-parameter estimates from the separate calibration of the 1PL for each group were used to create mini-test characteristic curves (TCCs) based on items of the same modification type: FC&RB, FC&LFNT, visual, and interview-based, for the ELs in our sample, in order to better understand the effects of different modification types on the test scores of EL students. The item parameter estimates for Test Version B were placed onto the same scale as Test Version A using the mean-sigma method. The item characteristic curves (ICCs) were summed for proficiency values ranging from -3 to 3 in increments of 0.1 for each item associated with each modification type: FC&RB, FC&LFNT, Visual, and Interview-based.

Differential item functioning analyses. Differential item functioning (DIF) refers to a situation where an item is more difficult for one group of students compared to another, after taking into account any overall differences between groups. Clauser and Mazor (1998) described DIF as being present "when examinees from different groups have differing … likelihoods of success on an item, *after they have been matched on the [proficiency] of interest*" (p. 31, emphasis added). The italicized phrase is key to understanding DIF, because it represents an *interaction* between group membership and the likelihood of a particular response on an item, conditional on the attribute measured.

We used DIF analyses in the present study to determine whether the modified versions of the items were more or less difficult than their original counterparts. Unlike more typical uses of DIF analysis, we compared scores on two different *forms* of an item for *one* group of students (e.g., ELs), rather than using DIF to compare the scores on the same item for two different groups of students (e.g., ELs and non-ELs). The hypothesis underlying our modifications is that

the items would be easier for ELs after modification. We used both logistic regression and an IRT-based DIF detection method (Lord's chi-square), but the results were similar and so we only describe the logistic regression method here.

Logistic regression estimates the probability of a correct response given an examinee's proficiency level. When testing for DIF between a reference (e.g., no modification) and focal (e.g., modification) group, the logistic regression model may be specified as:

$$P\left(u_{ij}=1\middle|\theta_j\right)=\frac{e^{\left[\tau_0+\tau_1\theta+\tau_2 g_j+\tau_3\left(\theta_j g_j\right)\right]}}{1+e^{\left[\tau_0+\tau_1\theta+\tau_2 g_j+\tau_3\left(\theta_j g_j\right)\right]}}, \quad (4)$$

where $P\left(u_{ij}=1\middle|\theta_j\right)$ represents the probability of correctly answering item *i* given examinee *j*'s proficiency, denoted θ_j; g_j is a dummy code used to represent whether examinee *j* is in the reference (*g*=0) or focal (*g*=1) group; τ_2 represents the difference between the reference and focal group, controlling for proficiency level; and τ_3 corresponds to the interaction between group and proficiency, denoted $\theta_j g_j$. Uniform DIF is indicated by $\tau_2 \neq 0$ and $\tau_3 = 0$ while non-uniform DIF is represented by $\tau_3 \neq 0$, whether or not $\tau_2 = 0$. In addition to evaluating items for statistically significant DIF, we also used effect size criteria to flag non-negligible DIF items. Following Jodoin and Gierl's (2001) effect size guidelines, items with effect sizes greater than .035 (signifying the modification status of the item accounted for 3.5% of the variation in item performance) were used to identify items as "medium DIF" and effect sizes greater than .07 were used to identify items displaying "large DIF."

For logistic regression, the covariate (θ_j) was based on raw scores. In addition, a two-stage purification approach was applied where the DIF items were initially identified using all items to represent the raw score. A second analysis was conducted in which the raw score was based on only the non-DIF items. An alpha level of 0.05 was used to flag DIF items and the change in *R*-squared values was used to classify statistically significant items as exhibiting negligible (< 0.035), moderate (0.035-0.07), and large (> 0.07) DIF. The R package *difR* was used to implement the logistic regression procedure.

Results

The results are organized first by reporting the ANOVAs based on overall scores on the original and modified item sets, then we report the results for sets of items defined by modification type, and finally the item-level (DIF) results.

Analyses of Effects of Modifications on Item Set Scores

We conducted ANOVAs on the (raw) total scores for original and modified item sets as described in the method section. We ran the analyses with and without the 6 common items as a covariate. However, the results were essentially the same and so we only report the results without the common items as a covariate here.[5]

ANOVA results for EL status and item set format. Descriptive statistics for the EL/non-EL groups and original and modified raw scores are presented in Tables 2 and 3 for Item Set 1 Score and Item Set 2 Score, respectively. For Item Set 1 Score, both ELs and non-ELs did slightly better on the modified items. The ANOVA results indicated a statistically significant main effect for EL status ($F_{(1, 1889)}$=167.6, p<.001), a significant main effect for item set

[5] We also conducted ANOVAs based on separate "original" and "modified" subscores for students calculated using item response theory. Those results were also similar to those reported here and so are omitted due to space considerations.

format ($F_{(1, 1889)}$=13.72, p<.001), and a non-significant student-by-format interaction ($F_{(1, 1889)}$=1.97, p=.161). The eta-squared effect size associated with the item format factor was .01, which indicates a negligible increase for both groups on the modified items.

Table 2

Descriptive Statistics for ANOVA on Item Set 1 Score

EL Status	Item Format	Dependent Variable: Item Set 1 Score
		Mean (SD)
EL	Original	6.15 (2.26)
	Modified	6.71 (2.04)
Non-EL	Original	7.74 (1.80)
	Modified	8.00 (1.61)

For Item Set 2 Score, ELs did slightly better on the original versions of the items, but non-ELs did slightly better on the modified versions. The ANOVA results indicated the main effect for EL status was statistically significant ($F_{(1, 1889)}$=141.0, p<.001), but the main effect for item format ($F_{(1, 1889)}$=0.71, p=.40), and the student-by-format interaction ($F_{(1, 1889)}$=3.04, p=.08) were not.

Table 3

Descriptive Statistics for ANOVA on Item Set 2

EL Status	Item Format	Dependent Variable: Item Set 2 Score
		Mean (SD)
EL	Original	5.97 (2.23)
	Modified	5.65 (2.25)
Non-EL	Original	7.20 (1.95)

	Modified	7.31 (1.86)

ANOVA results for student group and item format. The two-way ANOVAs with student group (defined as EL, non-EL below proficient on ELA test (BP non-EL), or non-EL proficient and above on ELA test (PA non-EL)) and item set format (original or modified) revealed statistically significant differences across the three EL groups for both test forms (Item Set 1: $F_{(2, 1790)}$=206.0, p<.001; Item Set 2: $F_{(2, 1790)}$=257.6, p<.001), with the PA Non-ELs scoring statistically significantly higher than the other two groups, and the BP Non-ELs scoring higher than the ELs. The interaction effects were not statistically significant for either comparison (Item Set 1: $F_{(2, 1790)}$=2.3, p=.10; Item Set 2: $F_{(2, 1790)}$=2.2, p=.11), although the Test Version effect was statistically significant for Item Set 1 ($F_{(1, 1790)}$=16.3, p<.001) but not for Item Set 2 ($F_{(1, 1790)}$=0.1, p=.74). These results are summarized in Tables 4 and 5. As can be seen from Table 4, the modified versions of the items were easier for Item Set 1, but the interaction did not suggest they were differentially easier for any of the three student groups. The modifications of Item Set 1 had a small effect size (eta-squared=.009) accounting for just under 1% of the variance in students' performance.

Table 4

Descriptive Statistics for 3-Group ANOVA on Score 1

EL Status	n	Item Format	Dependent Variable: Score 1
			Mean (SD)
EL	163	Original	6.15 (2.26)
	147	Modified	6.71 (2.04)
Non-EL, Not Proficient	334	Original	7.01 (1.92)
	321	Modified	7.35 (1.72)
Non-EL, Proficient	401	Original	8.42 (1.40)
	431	Modified	8.54 (1.27)

Table 5

Descriptive Statistics for 3-Group ANOVA on Score 2

EL Status	n	Item Format	Dependent Variable: Score 2
			Mean (SD)
EL	147	Original	5.97 (2.23)
	163	Modified	5.65 (2.25)
Non-EL, Not Proficient	321	Original	6.22 (1.97)
	334	Modified	6.40 (1.87)
Non-EL, Proficient	431	Original	8.03 (1.52)
	401	Modified	8.09 (1.48)

ANCOVA results for ELs only and English reading proficiency scores as covariate.

Reading proficiency scores from the ACCESS English proficiency test were available for the ELs who participated in the study. We used these scores as a covariate to look for differences in performance across modified and original versions of the item sets when English reading

proficiency scores were controlled for. As with the previous analyses, separate analyses were conducted for each item set. For both item sets, English reading proficiency scores were highly correlated with performance on the test items. For Item Set 1 Score, the correlation was .67; for Item Set 2 Score it was .65.

The ANOVA results for Item Set 1 Score are summarized in Table 6. The difference across item formats was statistically significant ($F_{(1,306)}$=8.22, p<.01), but the effect associated with this difference was relatively small (eta-squared=.026, indicating the modification accounted for 2.6% of the variance in the raw scores, after accounting for ACCESS reading score).

Table 6

Summary of ANOVA Results for ELs for Item Set 1 Score

Item Format	Mean (SD)	F	η^2
Original	6.15 (2.26)	8.22*	.026
Modified	6.73 (2.04)		

*p<.01.

The results for Item Set 2 Score are presented in Table 7. The difference due to item format was not statistically significant ($F_{(1,306)}$=1.95, p=.16).

Table 7

Summary of ANOVA Results for ELs for Item Set 2 Score

Item Format	Mean (SD)	F	η^2
Original	5.97 (2.23)	1.95	n/a
Modified	5.65 (2.25)		

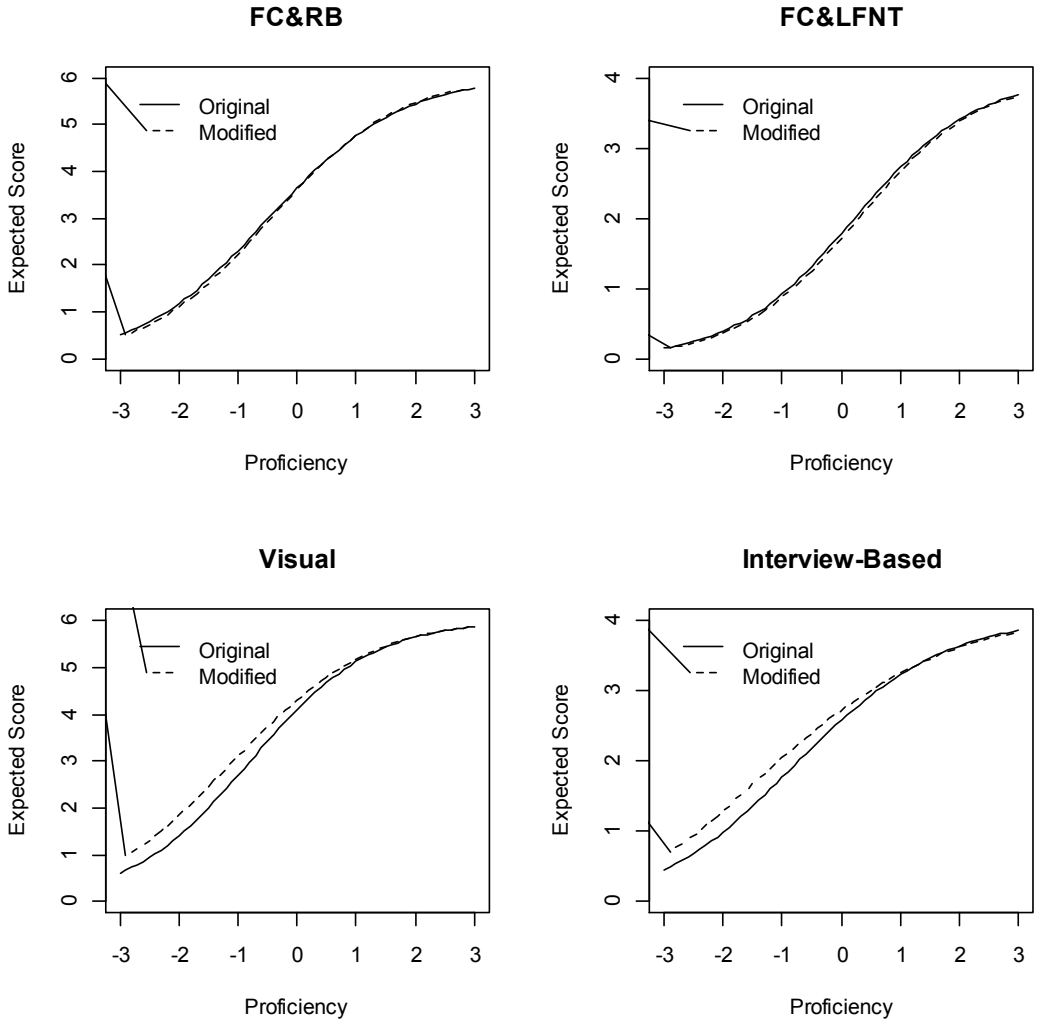
Comparison of Modification-type Test Characteristic Curves

As described in the method section, the item characteristic curves (ICCs) for each version of each test item were summed within each modification type to form "modification type" test characteristic curves (TCCs). Separate TCCs were computed for the ELs, BP non-ELs and PA non-ELs, and the original and modified TCCs for each modification type were plotted together to evaluate whether items undergoing the specific modification were differentially easier (as a group of items) for a specified student group, relative to their unmodified counterparts.

Test Characteristic Curves for ELs. Figure 1 provides the TCCs for each modification type based on the modified and original items and the data from the ELs in the sample. The *x*-axis in each graph represents the IRT proficiency scale and the *y*-axis represents the expected score. The TCCs for the modification types FC and RB and FC and LFNT were nearly identical for the original and modified items, indicating that the difficulty of the original and modified items were nearly identical for the ELs in the sample, across the entire range of proficiency. However, the TCCs for the modification types Visual and Interview-based were higher for the modified items, indicating that the modifications resulted in higher expected scores for the ELs.

Figure 1

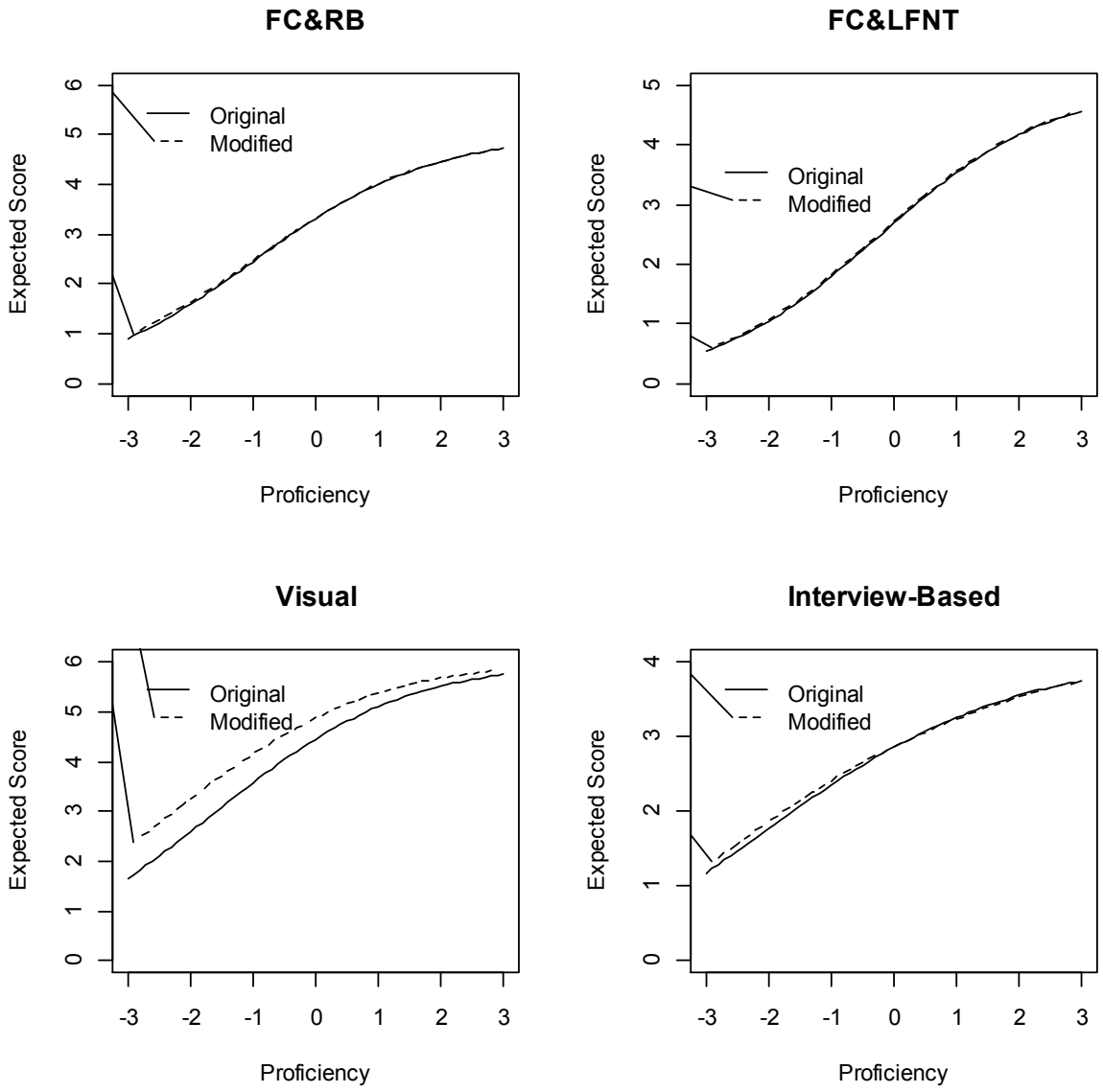
Modification Type TCCs for ELs



Test Characteristic Curves for BP Non-ELs. The TCCs for the Below-Proficient non-ELs are presented in Figure 2. The TCCs for the modification types FC&RB, FC&LFNT, and interview-based were nearly identical for the original and modified items. However, the TCC for the visual modification type was higher for the modified items, indicating that the modification resulted in higher expected scores for the BP non-ELs in our sample.

Figure 2

Modification Type TCCs for Below Proficient Non-ELs

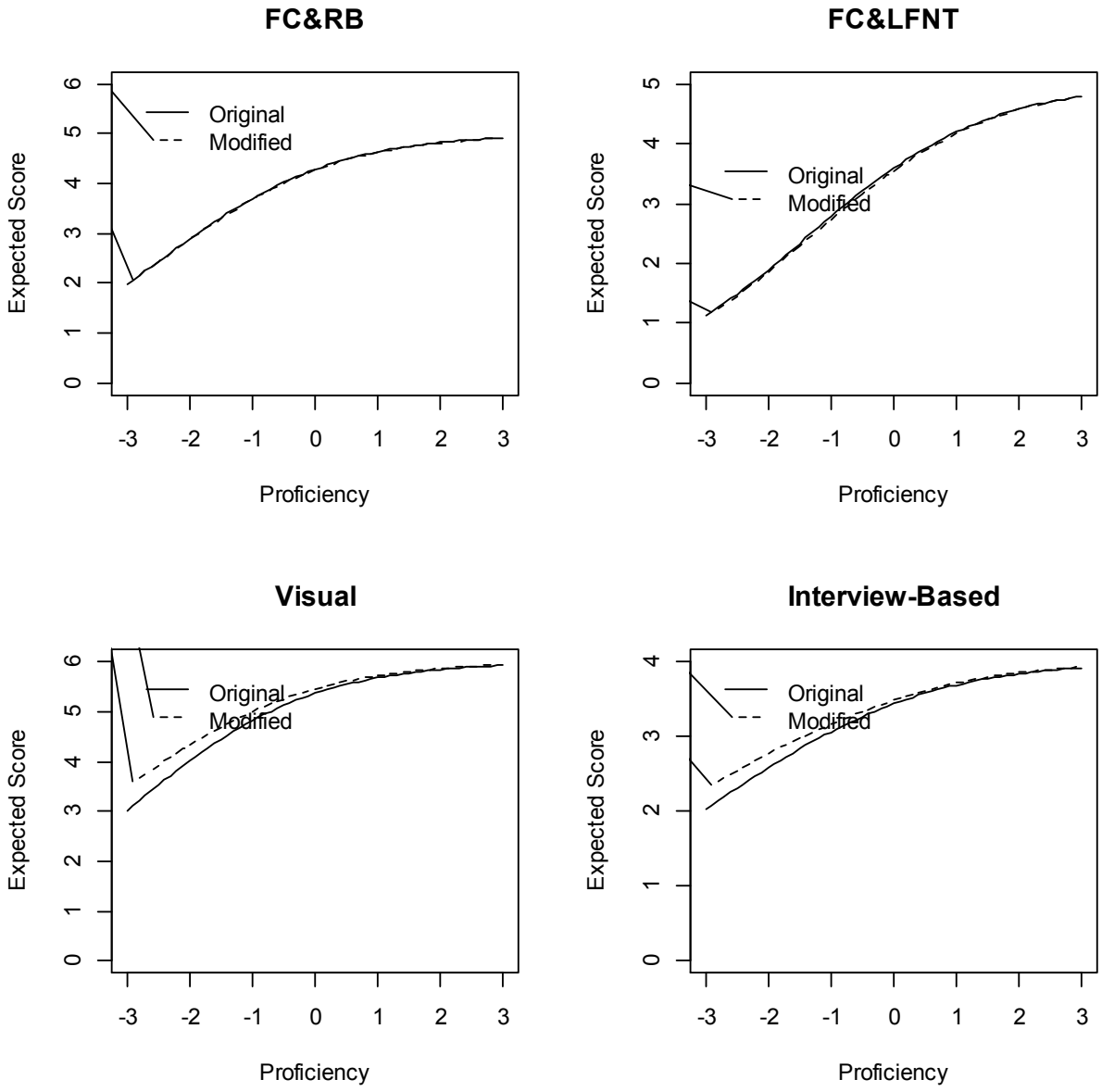


Test Characteristic Curves for PA Non-ELs. For the non-ELs who were Proficient or above, the TCCs for the modification types FC&RB and FC&LFNT were nearly identical for the original and modified items (see Figure 3). However, the TCCs for the modification types Visual and Interview-based were higher for the modified items indicating that the modification resulted

in higher expected scores, but only for those students scoring at the lower end of the proficiency scale.

Figure 3

Modification Type TCCs for Proficient or Above Non-ELs

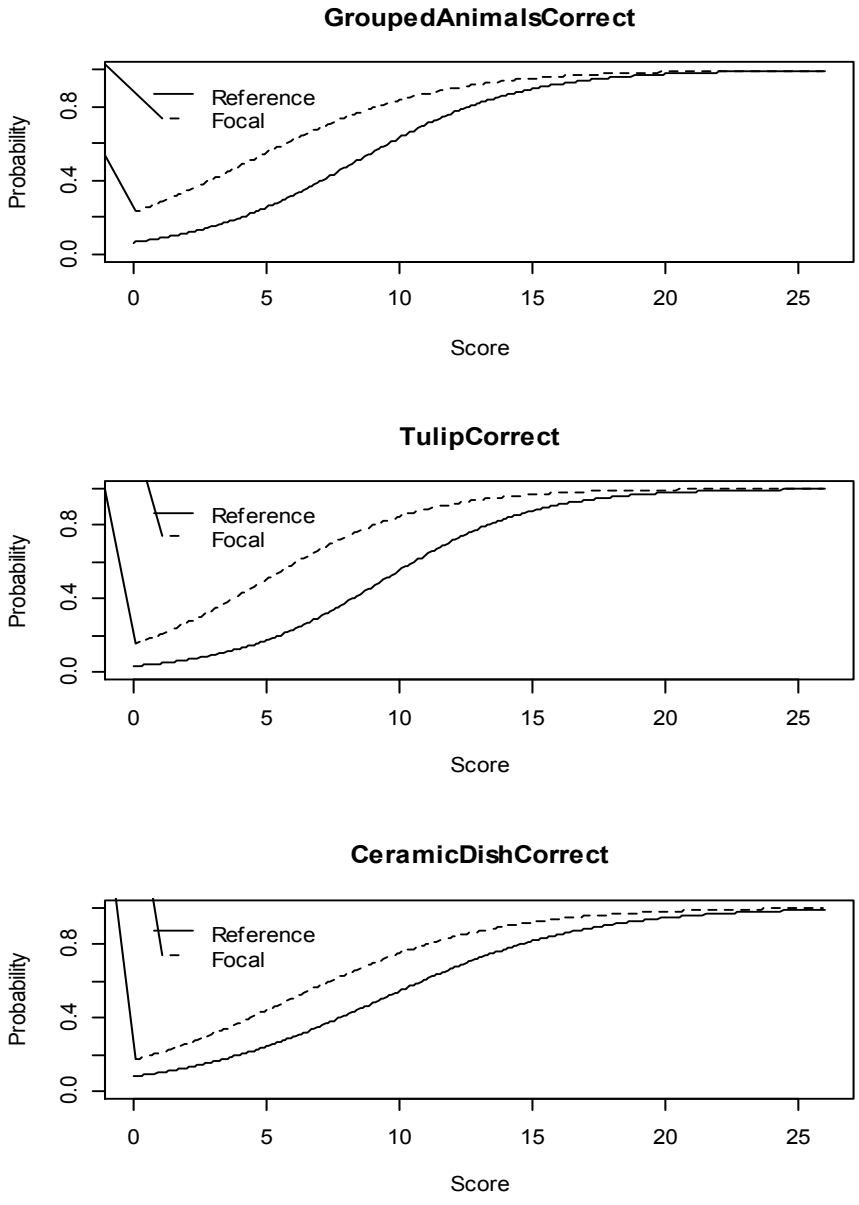


DIF Analyses

The first set of DIF analyses focused on ELs. Logistic regression flagged four items for statistically significant DIF. Of the four items, three had non-negligible effect sizes, all of which classified them as large DIF (i.e., *R*-squared values greater than 0.07). For all three items, the item was easier for the ELs in its modified condition. Figure 4 presents the logistic curves for each of the large DIF items. The *x*-axis represents the raw score and the *y*-axis represents the probability of a correct response. In each case, the logistic curve for the modified items (labeled focal group) was higher than the curve for the original items indicating that the items were easier for the modified condition.

Figure 4

Logistic Curves for Three Items Flagged for Large DIF (EL Group)



DIF analyses were also conducted for the two non-EL subgroups, and neither of these analyses flagged any of the three items listed above for non-negligible DIF. For the non-ELs who scored below proficient on the MCAS ELA test, only one item was flagged for non-negligible DIF, and

it had a large effect size. For the non-ELs who were proficient or above on the MCAS ELA test, two items were flagged—the one flagged in the analysis for BP non-ELs and one additional item. For both items, the modified version of the item was easier than the original version. These findings warrant further investigation of the reasons for the non-negligible DIF for these two test items for non-ELs. However, the findings also confirm that for the Grouped Animals, Tulip, and Ceramic Dish test items, the score improvements were specific to ELs.

Discussion

The findings of these analyses highlight the importance and the challenges of identifying the specific features of test items that have the greatest impact on the performance of ELs. Results showed that one set of item modifications, Item Set 1, had a greater positive impact on ELs' scores than the other, Item Set 2. These results suggest that the effects of modifications may be dependent upon the particular item set chosen for modification. Further analyses of groups of items within Item Set 1 based on IRT-based TCC analyses and DIF analyses showed that the Visual and Interview-based modifications were the most successful for ELs. Visual and Interview-based modifications also seemed to help non-ELs scoring at lower levels on the state ELA test, consistent with findings of other test item modification studies (e.g., Sato et al., 2010). Surprisingly, the modifications of Forced Comparison, Reference Back, and Low Frequency Non-Technical vocabulary features were not found to raise ELs' scores by statistically significant amounts, despite interview study findings demonstrating the problematic nature of these features for ELs (Kachchaf et al., April, 2014; Noble et al., April, 2014).

The analyses of the effects of individual test items on the performance of ELs and non-ELs has provided further information about which types of Visual and Interview-based modifications led to the largest improvements in the performance of ELs. DIF analyses

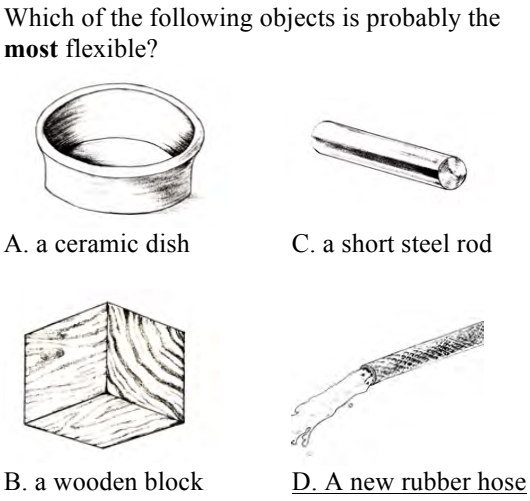
comparing student scores on original and modified versions of test items demonstrated that two of the Visual modifications and one of the Interview-based modifications led to differences in ELs' performance on the modified vs. the original versions of the items that were measurable as large DIF levels. These item modifications did not show any non-negligible DIF for either BP non-ELs or PA non-ELs, suggesting that these effects of modifications were specific to ELs, and supporting the conclusion that the tested science content was not altered for these items.

The modified versions of all three items flagged in the EL DIF analysis (Ceramic Dish, Grouped Animals, and Tulip, See Appendix B: Three Item Modifications) included the addition of visuals to the answer choices. These three items were the *only* items in the set of Visual and Interview-based modifications in which visuals were added to the answer choices, which helps to explain why the Visual modifications in Item Set 2 were not similarly successful. While the modified version of the Ceramic Dish item involved only the addition of visuals to the answer choices, the modified version of the Grouped Animals item also included the addition of visuals to the stem of the item, and the modified version of the Tulip item included small changes to the wording of the item to remove one Low Frequency Non-Technical word (*tulip*) and to remove the Forced Comparison feature by removing an additional word (*first*). It is striking that the three modifications showing the greatest impact on the performance of ELs all involve the addition of visuals, and furthermore, that these are the only items modified in this way.

To better understand how the addition of visuals to the answer choices affected the performance of ELs in this study, we returned to our interviews with ELs regarding the original versions of these test items (Kachchaf et al., April, 2014; Noble et al., April, 2014), with a focus on how students understood the words in the answer choices prior to the addition of visuals. In the Ceramic Dish item (See Figure 5), the words in the answer choices were a series of short

noun phrases: “A. a ceramic dish, B. a wooden block, C. a short steel rod, and D. a new rubber hose”.

Figure 5. Ceramic Dish Modified (Original test item from MA DOE, 2005)



We chose to illustrate the answer choices in this item in part because some of the words in the answer choices (ceramic, hose) were LFNT words. Our interviews with ELs regarding the original test item (without illustrations) confirmed that students had difficulty defining some of the words in the answer choices. For example, only one out of the five students asked to define the word *ceramic* gave a correct definition (acceptable answers included: like a plate, like a dish, what plates are made of, and similar answers) and seven out of eight students asked to define the word *rod* did not give the intended definition (acceptable answers included: like a stick). Of these seven students, three defined *rod* as a fishing rod, which is, by design, more flexible than a short steel rod. This definition led one interviewee to choose *a short steel rod* it as the answer, instead of the correct answer: *a new rubber hose*, because a fishing rod can be highly flexible. Some of the EL students whom we interviewed also did not give correct definitions for the words *steel, dish*, and *hose*. We found similar results in interviews with students about the Tulip test

item. None of the eight students interviewed about the Tulip test item was able to define the word *wilt*, a LFNT word, and a critical part of the correct answer choice: "The leaves would wilt". Similarly, for the Grouped Animals test item, one out of the three students asked to define the LFNT word *hawk*, the correct answer choice, was unable to define it.

This brief account of some of the findings of the interview study suggests one possible explanation for the increase in student scores on the modified items that included visuals in the answer choices. Some of the English LFNT words in the answer choices for these items were challenging to define for a number of the EL students whom we interviewed. The interviews demonstrate that these challenges made the selection of the correct answer choice much more difficult. The findings of the test item modification study indicate that the difficulty presented by these items is not due to lack of science content knowledge, but simply difficulty comprehending specific LFNT words in English. When visuals are provided, clarifying the meaning of the words in the answer choices, ELs perform significantly better on these items. These findings indicate that the use of LFNT vocabulary in answer choices for test items may lead EL students to select an incorrect answer, even when they know the science content targeted by the test item. Thus, such LFNT vocabulary in test item answer choices acts as a source of CIV in ELs' test scores, as demonstrated by the increase in scores for ELs when visuals were added to test items to clarify the meanings of answer choices containing these LFNT words.

It is notable that the specific challenges of low frequency vocabulary in answer choices has not been highlighted in previous research on ELLs and assessments in science and mathematics (Noble et al., 2016). Given that the majority of this research has concerned mathematics test items, it is likely that the issue of the vocabulary in answer choices has not been noted due to the predominance of numbers and symbols in the answer choices in mathematics

test items, which reduces the number of low frequency words found in mathematics test item answer choices[6]. The contrasting predominance of words over numerals, symbols, and visuals in the answer choices of science test items means that the language of the answer choices in science test items can present a specific challenge for ELLs, particularly when the language of the answer choices includes low frequency words unlikely to be taught in science class. The challenge of reading answer choices can have significant consequences for students who are still learning English, given that one's knowledge of a single English word in an answer choice (e.g., "wilt") may determine whether one can identify the correct answer or not.

Focusing on specific linguistic and visual features of science test items in our modifications and in our analyses has allowed us to identify the item modifications that were more (and less) successful at improving scores for ELs. We infer from the success of the modifications involving the addition of visuals to answer choices that these types of visuals can have a substantial impact on students' comprehension of the meaning of individual answer choices, and thus, their ability to select the correct answer choice. Identifying opportunities to use visuals to illustrate answer choices was an effective strategy for reducing the effect of test item language on test scores of ELs for the test items and the student population in our study. We hypothesize that clarifying the meaning of the words in the answer choices, particularly LFNT words, was an important benefit of the visuals added to the answer choices. The effectiveness of any form of test item modification is shaped by the content of the test items, the context of testing, and the characteristics of the students, among other factors. As a result, further research is needed to determine whether item modifications that were effective for our study are effective

[6] In the years 2010-2015, the Grade 5 Mathematics MCAS test has included multiple-choice test items without numbers, symbols, or visuals in the answer choices for only 12% of released items, whereas in 2010–2015, the Grade 5 STE MCAS test has included answer choices containing only words (no numbers, symbols, or visuals) for 85% of multiple-choice test items.

for other tests. We propose that across tests, student populations, and testing contexts, it is important to recognize the critical importance of language in answer choices, particularly for science test items in which knowledge of a single English word may determine one's ability to recognize the correct answer.

Directions for Future Research

Our study also reveals the benefit of using multiple sources of data, and multiple data analyses, to best understand the complex issues of how to write science test items for ELs and non-ELs that will allow valid inferences to be made about ELs' test scores. The interview data helped us to modify test items and to interpret the results of these modifications. While the analyses of total test scores on original and modified items showed no or small effects, the modification-type-level and item-level and analyses suggest that the effects of item modifications were being masked at the test score level. Thus, our findings suggest that researchers should continue to use multiple methods, at various levels of analyses (i.e., item, modification type, and total test score levels) to best understand interactions of ELs with test items.

The finding of this study regarding the positive impact of adding visuals to answer choices is promising, but further research in this area is needed. Extensive research has explored the use of visuals to illustrate item stems and described the characteristics of visuals that are more successful for ELs (Solano-Flores, et al., 2014), but there has not been a similar detailed analysis of the use of visuals to illustrate answer choices. Given the comparative success of illustrating answer choices in reducing the impact of the use of LFNT vocabulary in answer choices and improving the performance of ELs, it is important for assessment researchers to

better understand how answer choice illustration may be used as a tool to improve science assessments for ELs.

Authors' Note

This research was supported by the Institute of Education Sciences, U.S. Department of Education through Grant # R305A110122. The opinions expressed herein are those of the authors and do not reflect the opinions of the funding agency. The authors would like to thank the Massachusetts Department of Elementary and Secondary Education for their commitment to research and their collaboration in this project. The authors would like to thank, in particular, Catherine Bowler and Carrie Conaway. In addition, the authors would like to thank Beth Warren, Mary Catherine O'Connor, Carol Lord, Guillermo Solano-Flores, Richard Durán, Joel Webb, Gary Goldstein, Jack Ridge, Gillian Puttick, Catherine Suarez, Heidi Fessenden, and many others too numerous to mention for their contributions to the research reported herein and feedback on earlier versions of this paper. Finally, the authors would like to thank all of the district and school administrators, teachers, and students who participated in this research. This paper is dedicated to them.

Correspondence should be directed to Tracy Noble, TERC, 2067 Massachusetts Ave., Cambridge, MA 02140. E-mail: Tracy_Noble@terc.edu

References

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 377-398). Mahwah, NJ: Lawrence Erlbaum (http://www.taylorandfrancis.com).

Abedi, J. (2008). Utilizing accommodations in assessment. In N. H. Hornberger (Ed.), *Encyclopedia of Language and Education* (pp. 341-347). New York: Springer (http://www.springer.com/).

Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and Validity of Accommodations for English Language Learners in Large-Scale Assessments* (CSE Report 608). Retrieved from University of California, National Center for Research on Evaluation, Standards, and Student Testing website: http://www.cse.ucla.edu/products/rsearch.asp

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language Accommodations for English Language Learners in Large-Scale Assessments: Bilingual Dictionaries and Linguistic Modification* (CSE Report 666). Retrieved from University of California, National Center for Research on Evaluation, Standards, and Student Testing website: http://www.cse.ucla.edu/products/reports.php

Abedi, J., & Linquanti, R. (2012). *Issues and opportunities in improving the quality of large scale assessment systems for ELLs*. Paper presented at the Understanding Language conference, Palo Alto, CA. http://ell.stanford.edu/publication/issues-and-opportunities-improving-quality-large-scale-assessment-systems-ells

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda*. National Academies Press.

Basterra, M. d. R., Trumbull, E., & Solano-Flores, G. (Eds.). (2011). *Cultural Validity in Assessment: Addressing Linguistic and Cultural Diversity*. New York: Routledge (http://www.routledge.com/).

Butler, F. A., Lord, C., Stevens, R., Borrego, M., & Bailey, A. L. (2004). *An approach to operationalizing academic language for language test development purposes: Evidence from fifth-grade science and math*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Caldas, S. J. (2013). Assessment of Academic Performance: The Impact of No Child Left Behind Policies on Bilingual Education: A Ten Year Retrospective. In V. C. Mueller Gathercole (Ed.), *Issues in the Assessment of Bilinguals* (pp. 205-231). Bristol, UK: Multilingual Matters.

Center on Education Policy. (2011). *State Profiles for Assessment Policies through 2010-11*. Retrieved from website: http://www.cep-dc.org/page.cfm?FloatingPageID=23

Chudowsky, N., & Chudowsky, V. (2010). State Test Score Trends Through 2007-08, Part 6: Has Progress Been Made in Raising Achievement for English Language Learners? Washington, DC: Center on Education Policy.

Chudowsky, N., & Chudowsky, V. (2011). *State test score trends through 2008-2009, Part 3: Student achievement at 8th Grade*. Retrieved from website: http://www.cep-dc.org/publications/index.cfm?selectedYear=2011

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Duncombe, W., Lukemeyer, A., & Yinger, J. (2008). The No Child Left Behind Act Have Federal Funds Been Left Behind? *Public Finance Review, 36*(4), 381-407.

Every Student Succeeds Act, 20 U.S.C, §1177 (2016)

Fritzberg, G. J. (2004). No Child Left Behind?: Assessing President Bush's Assessment Law. *Educational Foundations, 18*, 7-24.

Gee, J. P. (1996). *Social Linguistics and Literacies: Ideology in Discourses* (2nd ed.). London, U.K.: Falmer.

Gee, J. P., Michaels, S., & O'Connor, M. C. (1992). Discourse analysis. In M. D. LeCompte, W. L. Millroy & J. Preissle (Eds.), *The Handbook of Qualitative Research in Education* (pp. 227-282). New York: Academic Press (http://store.elsevier.com/Academic-Press/IMP_5/).

Gutiérrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher, 32*(5), 19-25.

Hakuta, K., & Beatty, A. (2000). *Testing English-language learners in U.S. schools: Report and workshop summary*. Washington, D.C.: National Academy Press.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. doi: http://dx.doi.org/10.1111/j.1745-3992.2004.tb00149.x

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*: University of Pittsburgh Press.

Hemphill, F. C., & Vannerman, A. (2011). *Achievement Gaps: How HIspanic and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress* (NCES 2011-459). Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Kachchaf, R. R., Noble, T., Rosebery, A., Warren, B., O'Connor, M. C., & Wang, Y. (2015). A closer look at linguistic complexity: Pinpointing individual linguistic features of science muliple-choice items associated with English language learner performance. *Manuscript submitted for publication*.

Hopkins, M., Thompson, K. D., Linquanti, R., Hakuta, K., & August, D. (2013). Fully accounting for English learner performance a key issue in ESEA reauthorization. *Educational Researcher*, 0013189X12471426.

Huddleston, A. P. (2014). Achievement at whose expense? A literature review of test-based grade retention policies in US school. *education policy analysis archives, 22*, 18.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Kachchaf, R. R., Noble, T., Rosebery, A., Wang, Y., Warren, B., & O'Connor, M. C. (April, 2014). *The impact of discourse features of science test Items on ELL performance*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia.

Kieffer, M. J., Lesaux, N. K., 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*, 1168-1201. doi: 10.3102/0034654309332490

Kober, N., Chudowsky, N., & Chudowsky, V. (2010). State test score trends through 2008-09, Part 2: Slow and uneven progress in narrowing gaps. Washington, DC: Center on Education Policy.

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources state NAEP 1990–2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis, 34*(2), 209-231. doi: 10.3102/0162373711431604

Li, H., & Suen, H. K. (2012). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education, 25*(4), 327-346. doi: 10.1080/08957347.2012.714690

Linquanti, R., & Cook, H. G. (2013). Toward a "common definition of English learner": Guidance for states and state assessment consortia in defining and addressing policy and technical issues and options. *Washington, DC: Council of Chief State School Officers.*

MA DESE (2015). Spring 2015 MCAS Tests: Summary of State Results. Malden, MA: MA DESE.

MA DESE. (2015). Blueprints and Reporting Categories for Grades 5 and 8 Science and Technology/Engineering MCAS Tests. Malden, MA: MA DESE.

Massachusetts Department of Education. (2006). *Massachusetts Science and Technology/Engineering Curriculum Framework*. Massachusetts Department of Education Retrieved from http://www.doe.mass.edu/frameworks/current.html.

National Research Council. (2014a). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.

National Research Council. (2014b). *Literacy for Science: exploring the Intersection of the Next Generation Science Standars and the Common Core for ELA Standards: A Workshop Summary*. Washington, D.C.: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States* (Vol. 2). Washington, DC: The National Academies Press.

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

No Child Left Behind Act of 2001: State Plans, 20 U.S.C, §1111 (2002).

No Child Left Behind Act of 2001: Academic Assessment and Local Educational Agency and School Improvement, 20 U.S.C. §1116 (2002).

Noble, T., Kachchaf, R. R., & Rosebery, A. (2015). Assessment and English language learners: Synthesizing research on linguistic features and construct- irrelevant variance. *Manuscript submitted for publication*.

Noble, T., Kachchaf, R. R., Rosebery, A., Warren, B., O'Connor, M. C., & Wang, Y. (April, 2014). *Do linguistic features of science test items prevent English Language Learners from demonstrating their knowledge?* Paper presented at the annual meeting of the National Association of Research on Science Teaching, Pittsburgh.

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education, 27*(4), 248-260. doi: 10.1080/08957347.2014.944309

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale

science tests and what they know about science. *Journal of Research in Science Teaching, 49*(6), 778-803. doi: 10.1002/tea.21026

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

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 Practice, 30*(3), 10-28. doi: 10.1111/j.1745-3992.2011.00207.x

Pennsylvania Clearinghouse for Education. (March, 2013). Issue Brief: School Closings Policy. Retrieved from Research for Action website: http://bit.ly/13CAUuN& - 8221

Rivera, C., Collum, E., & Shafner Willner, L. (Eds.). (2006). *State assessment policy and practice for English language learners: A national perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C.-W. (2010). *Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets* (NCEE Report Number 2009-4079). Retrieved from Institute of Education Sciences National Center for Education Evaluation and Regional Assistance website: http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=REL20094079

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly, 23*, 139-159. doi: 10.1080/10573560601158461

Sims, D. P. (2013). Can failure succeed? Using racial subgroup rules to analyze the effect of school accountability failure on student performance. *Economics of Education Review, 32*, 262-274.

Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education, 39*(1), 215-252.

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The Effects of Test Accommodation on Test Performance: A Review of the Literature*. Retrieved from University of Massachusetts, Amherst, School of Education website: http://www.education.umn.edu/NCEO/OnlinePubs/

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English Language Learners. *Educational Researcher, 37*(4), 189-199.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553-573. doi: 10.1002/tea.1018

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. doi: 10.3102/0013189X032002003

Solano-Flores, G., Wang, C., Kachchaf, R., Soltero-Gonzalez, L., & Nguyen-Le, K. (2014). Developing testing accommodations for English language learners: Illustrations as visual supports for item accessibility. *Educational Assessment, 19*(4), 267-283. doi: 10.1080/10627197.2014.964116

Solórzano, R. W. (2008). High Stakes Testing: Issues, Implications, and Remedies for English Language Learners. *Review of Educational Research, 78*(2), 260-329.

WIDA. (2014). ACCESS for ELLs Summative Assessment Retrieved Oct. 2, 2015, 2015, from https://www.wida.us/assessment/access/

Wolf, M. K., Kao, J. C., Griffin, N., Herman, J. L., Bachman, P. L., Chang, S. M., & Farnsworth, T. (2008). *Issues in Assessing English Language Learners: English Language proficiency Measures and Accommodation Uses: Practice Review (Part 2 of 3)* (CRESST Report 732). Retrieved from University of California, Center for Research on Evaluation, Standards, and Student Testing website: http://www.cse.ucla.edu/products/policy.html

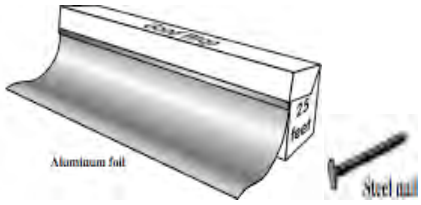
Zeno, S., Ivens, S. H., Millard, R. T., & Rothkopf, E. Z. (1995). *The educator's word frequency guide.* Brewster, NJ: Touchstone Applied Science Associates.

Appendix A: Features Modified

1. Visuals were added to test items when possible, because they were correlated with lower levels of DIF disfavoring ELs. Visuals include any non-linguistic information found in test items such as pictures, tables, charts, and diagrams. Visuals can be found in either the question portion of an item (as shown in Example B) or in the answer choices (as shown in Example C).

Figure 1. Classifying Objects Test Item (MA DOE, 2004):

The picture below shows three objects that can be classified in the same group.



Copper penny

Which of the following statements is true for all three of these objects?

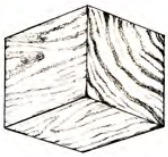
*A. They are metals.
B. They rust rapidly.
C. They weigh the same.
D. They are the same color.

Figure 2. Ceramic Dish Modified Test Item (modification of MA DOE item from 2005 that did not originally include visuals.)

Which of the following objects is probably the **most** flexible?

A. a ceramic dish

C. a short steel rod



B. a wooden block

*D. A new rubber hose

2. The Forced Comparison (FC) feature was removed in test item modifications due to its correlation with increased levels of DIF disfavoring ELs. This feature occurs in test items requiring students to compare all four answer choices and identify the one that expresses the correct **extreme value** of some variable.

- These items use **extreme value** terms like *best*, *most*, *most likely*, *greatest*. The specific meanings of these words depend upon the criteria for judging what is, for example, *best,* and information about the criteria the student should use is not always provided by the item.
- These items often use a verb, noun, or adjective in conjunction with the extreme value word that can have multiple meanings depending upon the context of use, such as *respond, help, cause, explain, important, and effort*.
- These items often use the complex question phrase: *Which of the following.*

An example of an item with the FC feature is shown in Figure 2, below, along with the FC modification of the item. The final sentence is underlined in both versions underlined to indicate the components of the FC feature described above.

Figure 3. Original (MA DOE, 2008) and FC Modification of Marsh Willow Test Item.

Original	FC Modification
The marsh willow herb is a plant native to the northeastern United States. It grows best in damp habitats. Which of the following environmental changes would **most likely** cause a decrease in the marsh willow herb population in an area?	The marsh willow herb is a plant native to the northeastern United States. It grows best in damp habitats. Which environmental change would reduce the marsh willow herb population in an area?
A. a rainstorm lasting several weeks *B. a drought lasting twelve months C. unusually low temperatures during the month of July D. unusually high temperatures during the month of January	A. a rainstorm lasting several weeks *B. a drought lasting twelve months C. unusually low temperatures during the month of July D. unusually high temperatures during the month of January

3. The Reference Back (RB) feature was removed in test item modifications due to its correlation with increased levels of DIF disfavoring ELs. This feature occurs in test items in which the question sentence refers back to information that appeared earlier in the item and that is needed to understand and solve the item. For example, See Figure 4.

Figure 4. Earthworm Test Item (MA DOE, 2004).

An earthworm was placed on top of a thick layer of moist topsoil in a pan. The pan was placed in a room with the lights on. How did the earthworm **most likely** respond to these conditions?

*A. by burrowing under the soil
B. by crawling around in the pan
C. by staying where it was placed
D. by trying to crawl out of the pan

In Example E, the term "these conditions" in the question refers back to the previous two sentences and the student must remember that the conditions for the earthworm include both the presence of a thick layer of moist topsoil in a pan (from the first sentence) as well as the placement of the pan in a room with the lights on (the second sentence). However, the new stimulus that the earthworm is expected to respond to with an aversive behavior is the lights, and thus a modification of the RB feature of this test item would be the following:

Figure 5. Modification of Earthworm Test Item.

An earthworm was placed on top of a thick layer of moist topsoil in a pan. The pan was placed in a room with the lights on. How did the earthworm **most likely** respond to the lights?

*A. by burrowing under the soil
B. by crawling around in the pan
C. by staying where it was placed
D. by trying to crawl out of the pan

4. Low Frequency Non-Technical (LFNT) Words were removed or replaced in modified test items due to their correlation with DIF disfavoring ELs. LFNT words not appear routinely in 5th grade texts, have common non-scientific meanings or uses (hence are non-Technical), and are unlikely to be explicitly taught in science class. In Example G below, LFNT words are *italicized* and their replacements are underlined.

Figure 6. Original (MA DOE, 2008) and LFNT Modification of Marsh Willow Test Item.

Original	Low Frequency not Technical Modification
The marsh willow *herb* is a plant native to the *northeastern* United States. It grows best in damp habitats. Which of the following environmental changes would **most likely** cause a *decrease* in the marsh willow *herb* population in an area?	The marsh willow plant is native to New England. It grows best in damp habitats. Which of the following environmental changes would **most likely** reduce the marsh willow plant population in an area?
A. a *rainstorm* lasting several weeks *B. a *drought* lasting twelve months C. *unusually* low temperatures during the month of July D. *unusually* high temperatures during the month of January	A. rain lasting several weeks *B. a *drought* for twelve months C. very low temperatures during the month of July D. very high temperatures during the month of January

Due to the fact that not all of the low frequency vocabulary identified could be easily replaced with a new word that maintained the meaning, a few additional changes often occurred during the LFNT modification process. For example, the original item stated, *The marsh willow herb is a plant native to*, however, replacing *herb* with *plant* would have resulted in the word *plant* being used twice in the same sentence. Therefore, the modified version replaced *herb* with *plant* and removed the subsequent use of *plant* from the sentence. Similarly, *northeastern* could not be replaced with an equivalent word. The modified version replaced the noun phrase *northeastern United States* with *New England*. This also resulted in the need to remove the article *the,* and the resulting modification changed *the northeastern United States* to *New England.* We acknowledge that these modifications reduce the total number of words in the item, and we attempted to avoid this type of complex modification whenever possible. In addition, *drought* was classified as LFNT but was also judged to be part of the science knowledge needed to answer the item. Therefore, it was determined that *drought* could not be removed. In all other cases, the word identified as low frequent was replaced with an equivalent term that maintained the meaning of the item and was not a low frequency word.

5. Interview-based Modifications

Interviews with Grade 5 EL students informed this modification type. Students' responses to interview questions often revealed that features of test items not previously identified interfered with students' comprehension of test item language. For example, when interviewed about the original version of the Earthworm test item shown in Figure 7, EL students had alternative interpretations of the words *pan*, *thick*, and *respond* that interfered with their comprehension of the item, despite the fact that these are not low frequency words.

In addition, students reported in interviews that they did not notice the second sentence of the original item: *The pan was placed in a room with the lights on*. This sentence contains key information about the stimulus of the lights to which the earthworm is expected to respond by burrowing under the soil, but was presented in a prepositional phrase at the end of the middle sentence of the item, leading some students to ignore it. The modified version states in the first sentence that the earthworms were in a dark room and the second sentence uses active voice rather than passive voice to indicate that there was a change in the amount of light in the room. This modification highlights the stimulus of the lights, which was intended to be noticed and considered important by students, but was not always noticed in its previous form.

Figure 7. Original (MA DOE, 2004) and Interview-based Modification of Earthworm Test Item.

Original	Modification Part 1: Forced Comparison, Reference Back, and Low Frequency non Technical words removed	Interview–based Modification
An earthworm was placed on top of a **thick layer** of moist topsoil in a **pan**. **The pan was placed in a room with the lights on**. How did the earthworm **most likely respond to** these conditions?	An earthworm was placed on top of a **thick layer** of moist soil in a **pan**. **The pan was placed in a room with the lights on**. How would the earthworm **respond** to the lights?	An earthworm was placed on top of **some** moist soil in a **box** in a **dark room**. **Then the lights were turned on.** How would the earthworm **react** to the lights?
*A. by burrowing under the soil B. by crawling around in the pan C. by staying where it was placed D. by trying to crawl out of the pan	*A. by going under the soil B. by crawling around in the pan C. by staying where it was placed D. by trying to crawl out of the pan	*A. by going under the soil B. by crawling around in the pan C. by staying where it was placed D. by trying to crawl out of the pan

- Forced Comparison feature and modification underlined
- Low Frequency non Technical feature and modification in blue
- Reference Back feature and modification in green
- Problematic words and aspects identified by interviews and modifications in bolded black

Appendix B: Three Item Modifications

Grouped Animals Original
The lists below show animals separated into two different groups.

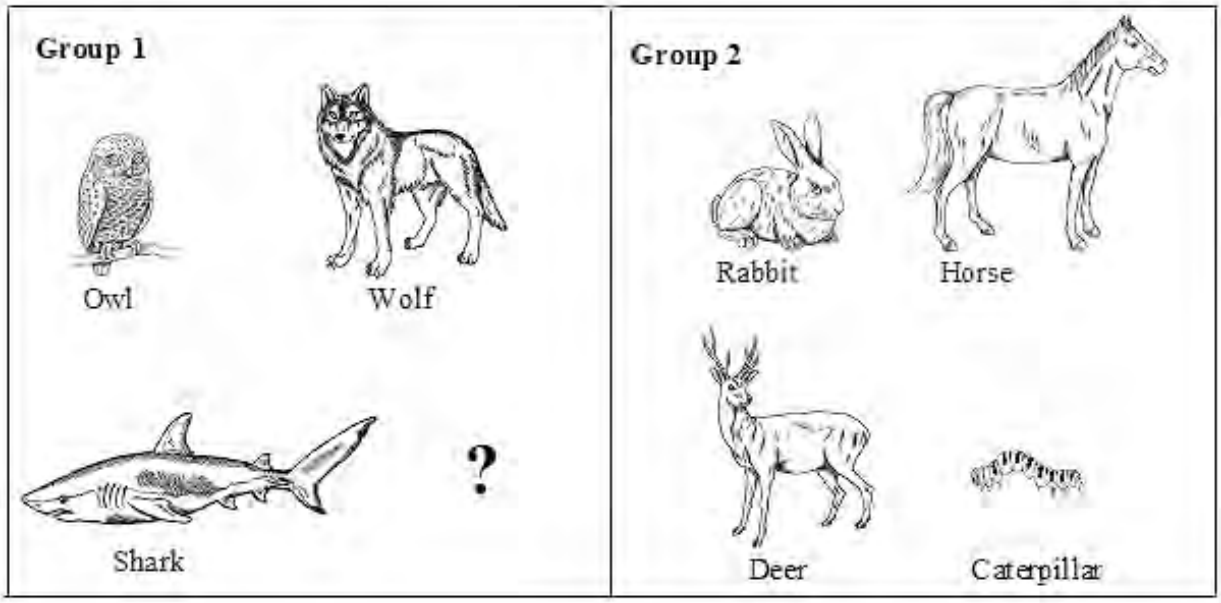
Group 1	Group 2
owl	rabbit
wolf	horse
shark	deer
?	caterpillar

The animals above are grouped by eating habits. Which of the following animals belongs in Group 1?

A. squirrel
B. sheep
C. hawk
D. goat

Grouped Animals Modified (Visual Maximized)

The pictures below show animals separated into two different groups.



The animals above are grouped by eating habits. Which of the following animals belongs in Group 1?

A

squirrel

B

sheep

C

hawk

D

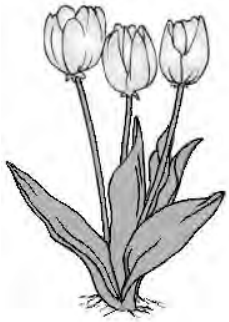
goat

Tulip Original
A healthy red-flowered tulip plant is shown below.

Which of the following would occur first as a result of many days with no rain?

A. The tulip's leaves would wilt.
B. The tulip's flowers would turn blue.
C. The tulip's stems would grow longer.
D. The tulip would produce more flowers.

Tulip Modified (Interview-Based)
A healthy plant with red flowers is shown below.



What would occur as a result of many days with no rain?

A.

The leaves would wilt.

B.

The flowers would turn blue.

C.

The stems would grow longer.

D.

The plant would produce more flowers.

Ceramic Dish Original
Which of the following objects is probably the **most** flexible?

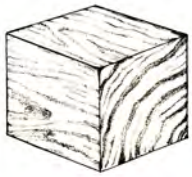
A. a ceramic dish
B. a wooden block
C. a short steel rod
D. a new rubber hose

Ceramic Dish Modified (Visual Added)
Which of the following objects is probably the **most** flexible?

A

a ceramic dish

B



a wooden block

C

a short steel rod

D

a new rubber hose