Addressing the Linguistic Challenges of Assessing ELs:

A State and Research Organization Partnership

Tracy Noble

TERC

Catherine Bowler

Massachusetts Department of Elementary and Secondary Education

Rachel Kachchaf

Smarter Balanced Assessment Consortium

Stephen G. Sireci

University of Massachusetts, Amherst

Ann Rosebery

TERC

Yang Wang

Education Analytics

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

Abstract

This paper provides an overview of research undertaken in a collaboration between researchers at TERC and test developers at the Massachusetts Department of Elementary and Secondary Education (MA DESE) to investigate how the language of 5th grade multiple-choice state science test items affects the opportunities for English Learners (ELs) to demonstrate what they know about science when answering these science test items.

Addressing the Linguistic Challenges of Assessing ELs:

A State and Research Organization Partnership

The majority of state science tests are given in English, with limited accommodations available to ELs (Wolf et al., 2008), and as a result, ELs' test scores may not always reflect their knowledge and skills in science. Researchers have explored the specific linguistic features of science test items that interfere most with ELs' comprehension of test item texts, and hypothesized that the language of test items in science and in mathematics may contribute to construct-irrelevant variance (CIV) in the test scores of ELs (See Noble, Kachchaf, and Rosebery, 2015, for a brief review). 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 ELs on science tests are significantly influenced by 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 for Educational and Psychological Testing 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).

Assessment research has identified multiple ways in which the language of test items in science and mathematics contributes to CIV in the test scores of ELs, that is, interferes with ELs' demonstrating what they know about science and mathematics when answering test items (See Noble et al., 2015 for a review). However, the majority of research in this area has focused on

mathematics rather than science assessments (Abedi & Lord, 2001; Hofstetter, 2003; Lord, Abedi, & Poosuthasee, 2000; Martiniello, 2008, 2009; Sato, Rabinowitz, Gallagher, & Huang, 2010; Wolf, Kim, & Kao, 2012). While many aspects of mathematics and science assessments are similar, language is used in different ways in science and mathematics practice and pedagogy (Halliday & Martin, 1993; Schleppegrell, 2007), and thus can be expected to have different effects on student performance across the two content areas. Further research specific to science assessments for ELs is needed, and in particular, research that identifies the linguistic features of science test items that influence the performance of ELs (Penfield & Lee, 2010).

We identify ELs as students who are identified by their schools and districts as Limited English Proficient based upon a home language survey and a test of English proficiency. We also recognize the limitations of the classification of students as ELs, due to multiple factors, including 1) the diversity of the backgrounds, languages, language proficiencies, and life experiences that can be masked by grouping students together under the category ELs and 2) the limitations of past and current testing practices in accurately assessing students' English proficiency and assigning students into or out of EL status (Proctor & Silverman, 2011). We nonetheless recognize that the category of English Learner (EL) has significance within the research and practice communities and identifies for schools, districts, and states, a group of students who have historically not received adequate attention from the research, assessment, and policy communities. For this reason, we will use the category EL for the students whom we describe in this paper, while also recognizing its limitations.

In addition to the need for research in on the language of science test items and ELs, there is a need for a conversation between researchers and practitioners about the relationship between the findings of research and the processes of test development. This conversation needs to

recognize that during the time it takes to get a single research article published, states and test developers are continually developing new test items and new tests, and are responding to changes in state and national assessment policy and practice, such as the publication of the Next Generation Science Standards (NGSS Lead States, 2013) and the passage of the Every Student Succeeds Act (2016). As a result, research-practitioner partnerships need to be built around ongoing and shared learning during the course of research.

To describe the ongoing conversation between practitioners and researchers in the course of the projects described herein, we use Goodwin's (1994) concept of professional vision to explore the ways in which practitioners and researchers shaped each others' ways of perceiving science assessment items. In a discussion of the work of archaeologists and the work of trial lawyers, Goodwin (1994) develops the idea that members of professions such as these use various techniques to share the ways in which they see the world through the lenses of their professional training. In the collaboration of researchers and practitioners described in this paper, each group came to incorporate aspects of the professional vision of the other group in order to develop overlapping ways of viewing science test items. The development of this shared professional vision was essential to the collaboration at the heart of this research.

Collaboration

Researchers from TERC and MA DESE science test developers have collaborated on the work of the English Learners and Science Tests project (IES grant #R305A110122) from its inception, in 2010, to the present. In this collaboration, we built upon prior research to explore ways to improve the fairness of the 5th grade Science and Technology/Engineering (STE) Massachusetts Comprehensive Assessment System (MCAS) for ELs and the validity of interpreting ELs' scores on that test as representations of their knowledge and skills in science.

In this paper, we describe the nature of the collaboration, two studies that we conducted in collaboration, and the outcomes of these studies for the researchers and test developers.

This collaboration has had multiple elements critical to its success, and these include 1) sharing information and 2) participation in each others' work. Sharing information has involved researchers sharing research findings with practitioners at all stages of the research process. Similarly, test developers have shared with researchers details of the test development process in MA and data essential to the progress of the research. Participation in each others' work has involved test developers serving as advisors to the research project and providing essential feedback on the development of coding schemes for test items and the creation of test item modifications. In addition, researchers have participated in item development review meetings with test developers to address concerns about science items, and to share expertise with science teachers, science coordinators, and the staff of the state's assessment development contractor.

From the researchers' perspective, additional critical elements of the collaboration have included: 1) the state policy of publicly releasing test items, 2) the state policy of developing data sharing agreements permitting the sharing of existing state data with researchers, and finally, and most importantly, 3) the commitment of state test developers to improving the item development process and to promoting and assisting with research with this aim.

From the state test developers' perspective, additional critical elements of the collaboration have included: 1) researchers' willingness to participate with test developers in discussions about test items under development, for example, discussions of field test results for items, 2) researchers' interest in focusing on the MA state tests and sharing findings with test developers and other practitioners, and 3) researchers' willingness to work collaboratively to create products of research that can be applied by test developers and are not intended only for

other researchers.

One critical element of this collaboration that makes it unusual among other states is that Massachusetts has several science test developers with science content expertise on staff. Their primary roles are to review and edit all STE test items and materials developed by the testing contractor. This practice is not common in other states, and allows the MA DESE test developers to be responsive to the findings of research in their test development process.

State Assessment

The focus of this work was the multiple-choice science test items from the 5th grade STE MCAS. The STE MCAS is administered every year to MA students in 5th grade and 8th grade. Discipline-specific science MCAS tests are administered to students in high school. These STE MCAS tests fulfill federal and state accountability requirements for science testing. The 5th grade STE MCAS currently includes 38 multiple-choice and 4 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 multiple-choice test item is worth one raw score point, and each open-response test item is worth 4 raw score points. The STE MCAS is aligned with the MA STE Curriculum Frameworks (Massachusetts Department of Education, 2006) and the included 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 (MA DESE, 2015). The state currently releases half of the test items to the public each year, and prior to 2009, all test items were released each year. The focus of our study is on the approximately 32 multiple-

[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.

choice test items associated with the Earth and Space Science, Life Science, and Physical Sciences strands in the 5th grade STE MCAS each year.

State test developers and researchers collaborated on two primary studies investigating the effect of specific linguistic features of 5th grade STE MCAS science test items on the performance of ELs. For this paper, we will call them Study 1 and Study 2. They will be described briefly in the sections that follow.

Study 1: Correlation

Study 1 was designed to investigate the relationship between linguistic features of released 5th grade multiple-choice science MCAS items and the performance of ELs compared to non-ELs on those test items. Similar studies have uncovered patterns in students' performance on state mathematics test items (Lord, et al., 2000; Martiniello, 2009) and in combined data from mathematics and science tests (Abedi et al., 2000/2005; Wolf & Leon, 2009). We have not encountered a study of this kind focused exclusively on science test items, and we wished to understand whether the linguistic features correlated with ELs' performance on science test items may differ from those correlated with ELs' performance on math test items. In addition, we sought to explore which of the linguistic features identified in the literature as correlated with differences in performance between ELs and non-ELs for state and national tests would similarly occur in and affect ELs' performance on the science items on the MCAS. Study 1 is reported in in greater detail in by Kachchaf et al. (2015).

Our analysis of 5th grade STE MCAS test items focused on the science test items, which make up approximately 85% of multiple-choice test items in each year. In this study, we analyzed 162 released multiple-choice science items from the 2004-2010 STE MCAS tests for the presence of 13 linguistic features hypothesized based on prior research and an extensive

review of the literature (Noble et al., 2015) to affect the performance of ELs. We examined the frequency and distribution of these linguistic features and their relationship to Differential Item Functioning (DIF) levels comparing EL and non-EL performance. We calculated DIF using the Standardization method (Dorans & Kulick, 1986), due to its sensitivity to small differences in DIF for test items and its use by the MA DESE for other purposes. In the Standardization method, the average difference between the item p-value (percent correct) for ELs and for non-ELs on an individual item is estimated, when students are matched for the total score on the remainder of the items, and the percentage of ELs at each score level is used as a weighting factor. In this method, performance on individual items is compared for ELs and non-ELs scoring similarly on the test as a whole.

Participants

In this study, we analyzed the test scores of Grade 5 students classified as ELs or non-ELs by their schools. We categorized the students classified as Limited English Proficiency (LEP) by their schools as ELs and those students classified as neither LEP nor Formerly Limited English Proficient (FLEP) as non-ELs. For any given year from 2004-2010, this statewide study included 52,694 – 56,991 non-ELs and 2,645 – 3,804 ELs, from all school districts in MA. The EL student populations in the state represented over 68 first languages, with Spanish reported as the first language by the largest percentage of both EL students (55.8%) and Portuguese the second most frequent first language (7.3%). No other first language made up more than 6.0% of either LEP or FLEP student populations.

Results of Study 1

The findings of Study 1 are shown in Table 1.

Table 1

Correlations between Linguistic Features of Test Items and Item DIF Values

Features	LEP
Word	
Low Frequency Vocabulary	.030
General Academic Vocabulary	-.045
Technical Science Terms	-.206**
Context-Specific Science Terms	.018
Low Frequency Non-Technical Vocabulary	.155*
Sentence	
Relative Clause	.048
Sub. Adverbial Clause	-.049
Relative and Subordinate Adverbial Clause	-.071
Number of Nouns in Answers Options	.110
Number of Verbs in Answers Options	.142
Item	
Forced Comparison	.194*
Reference Back	.192*
Visual	-.276**

* $p < .05$; ** $p < .01$.

These findings indicate that five specific linguistic and visual features of Grade 5 multiple-choice science test items from the STE MCAS 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 the *best, most likely*, etc., that is, that is at an extreme on some 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 Study 1. Each feature will be more fully defined in the section describing Study 2.

Study 2: Test Item Modification

In Study 2, we used the findings of Study 1 to do a test item modification study to explore the effects of test item modifications targeting the specific linguistic and visual features of MCAS science test items identified in Study 1. The goal of this study was to identify specific linguistic features of MCAS science test items that interact with ELs' levels of English proficiency to lead to CIV in student test scores. 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 DIF favoring non-ELs over ELs and 2) the presence of one or more

of the three features found to lead to higher levels of DIF favoring non-ELs (i.e., FC, RB, and LFNT vocabulary), 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 found to lead to higher relative performance for ELs. Very few items could accommodate the addition of Technical vocabulary, and as a result, we did not use this modification type. A more detailed description of each of these modification types is given in the sections that follow and in Appendix A.

Visual modification. 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[2], 2) adding visuals to each of the four answer choices, or 3) maximizing a visual already present in the item. The added visuals were created to illustrate objects rather than processes and 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, 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 other aspects of the FC feature, including the question format *Which of the following,* and the noun or verb associated with the extreme value descriptor.

[2] 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.

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 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 experimented with our modification process. 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 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 students' performance on the test items and comprehension of the questions being asked by the 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. Individual feature modifications were tested for the FC, RB, and LFNT features and were found to be insufficient to improve students' comprehension of the items.

Interview-based modification. In this modification, multiple features from the set (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 altered 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. More detail about each of the modifications is provided in Appendix A: Features Modified.

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 the 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 with expertise in linguistics, STE assessment design, and EL education, including MA DESE test developers, reviewed our original and modified test items to judge the effectiveness of the modifications, their consistency with the modification methods, and the consistency of the modifications with state test design standards.

Participants

Four public, urban MA school districts agreed to participate in the study. Districts were recruited in 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. In two districts, all schools serving 5th grade students participated in the study. In another two districts, principal volunteers were sought by the district, and resulting in two participating schools in one district and three in the other district. 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.

Study 2 Research design

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 version consisted of twenty experimental test items, consisting of 10 test items in their

original form, 10 test items in their linguistically modified form, and six anchor items that were common to both test forms, for a total of 26 items on each test form. The anchor items were items selected based on their lack of linguistic complexity and negligible levels of DIF favoring non-ELs over ELs. All experimental test items 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 test forms with 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 developers 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 early were asked to read a previously selected book.

Data Analysis and Results of Study 2

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. 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. The details of these analyses and results are given by Noble et al. (2016). A summary of the results will be provided here.

Effects of modifications by item set. The initial analyses were designed to determine the effects of each set of 10 item modifications on student scores for ELs and for non-ELs. As described earlier, each student answered 10 original items, 10 modified items, and six anchor items. Thus, students who received one set of 10 original items received a different set of 10 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. 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 in Table 2.

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)

The results of these analyses indicated that the students as a whole scored statistically significantly higher on the modified items in Item Set 1, which included a small improvement in scores for ELs, and a smaller improvement for non-ELs. However, there was no statistically significant interaction between student group (EL or non-EL) and item format (original or modified). For Item Set 2, there was no statistically significant effect of these modifications on students' performance as a whole. For this reason, we chose to explore the effects of the specific modification types in Item Set 1, and the individual item modifications in this item set, to better understand why the modifications of the items in Item Set 1 were more successful than those in Item Set 2.

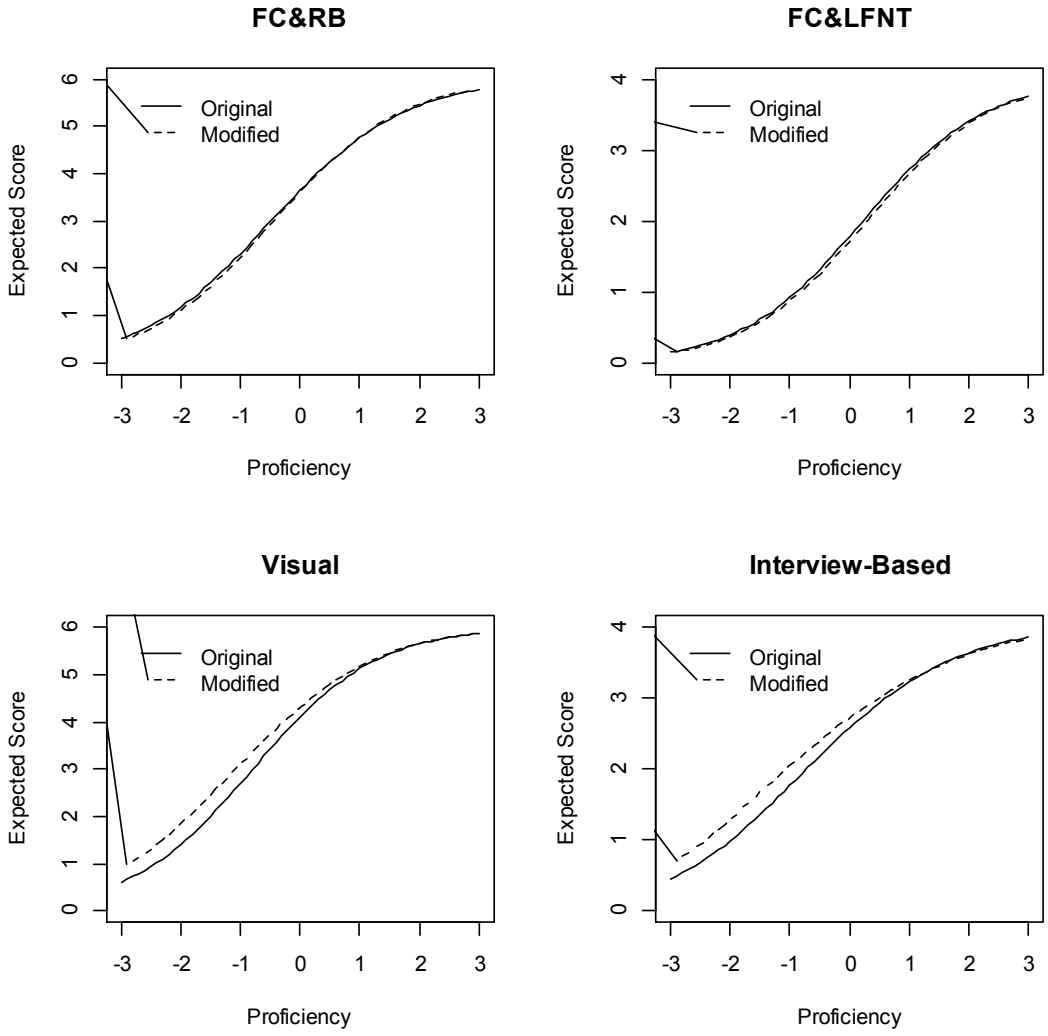
Effects of modifications by modification type. To evaluate whether modification effects were linked to one of the four specific modification types (i.e., FC and RB, FC and LFNT, Visual, and Interview-based), we used IRT to calibrate all the test items so that "mini-test characteristic curves" (TCCs) 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.

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

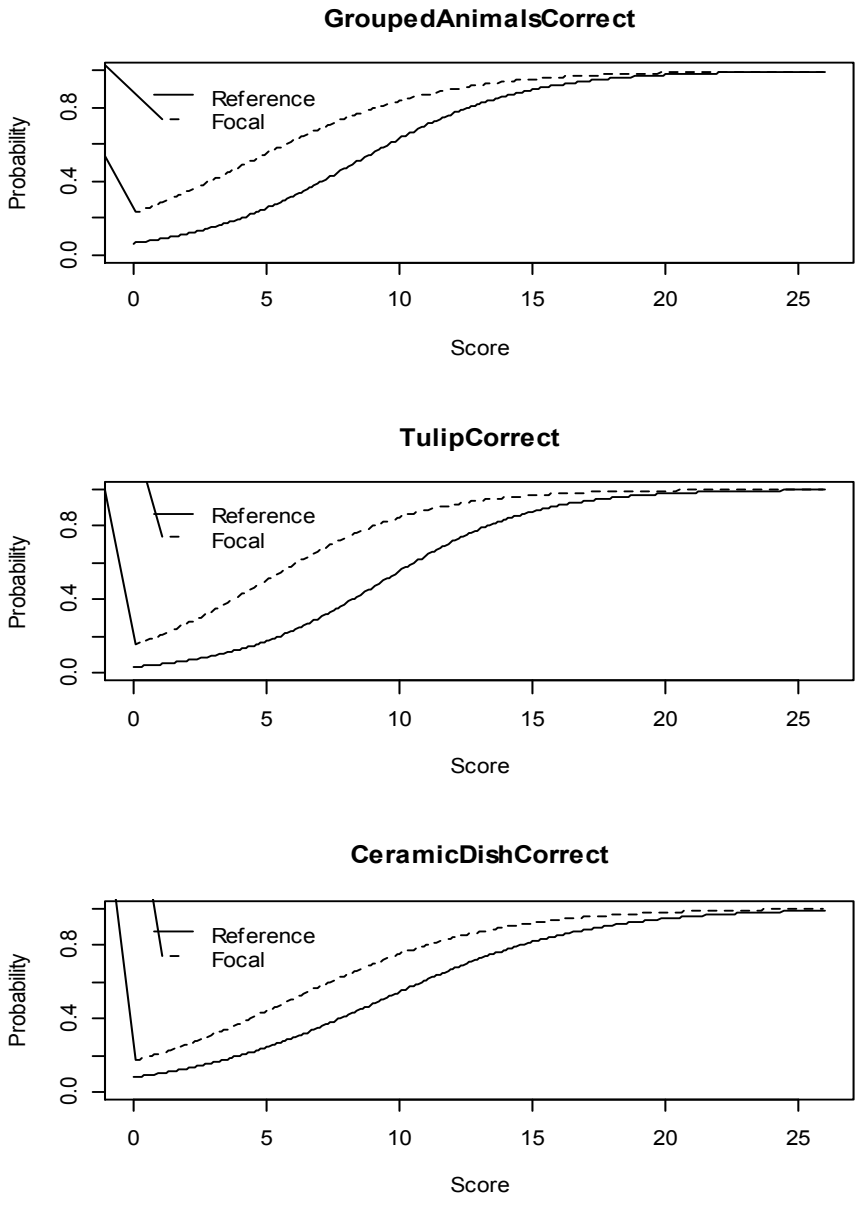


Effects of modifications by individual test item. We used DIF analyses in the present study to determine whether the modified versions of the items were more or less difficult than their original counterparts. Unlike more typical uses of DIF analysis, we compared scores on two different *forms* of an 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 report the logistic regression method results here.

Three test items 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 2 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 2
Logistic Curves for Three Items Flagged for Large DIF (EL Group)



Discussion

The findings of Studies 1 and 2 suggest that some of the language-focused modifications (FC and LFNT, FC and RB) were not sufficient to change ELs' scores on test items, but that Visual and Interview-based modifications were. Furthermore, the analyses identified three test items for which ELs scored statistically significantly higher on the modified

than the original test items, when matched for scores on the remainder of the items. The modified versions of all three items (Grouped Animals, Tulip, and Ceramic Dish) included the addition of visuals to the answer choices. That is, a visual image was added to each answer choice to illustrate the meaning of the words in these answer choices. The original and modified versions of these three items are provided in Appendix B. These findings suggest an alternative route for illustration of test items in order to clarify content for ELs: the illustration of answer choices. These findings also suggest that the language of answer choices may be more important for the experience of ELs than had been previously realized.

The collaboration sustained throughout Studies 1 and 2 offered researchers and test developers multiple opportunities to share their perspectives on test items and to develop their professional vision regarding the features of test items that are problematic for ELs and the requirements of test item writing that are faced by the state. The process of developing test item modifications in this collaboration was particularly important in the development of shared perspectives on test item design. The technique of highlighting, described by Goodwin (1994) as a method for developing professional vision, was used in the process of developing shared perspectives on test items. For example, the findings of Study 1 allowed researchers to highlight specific features of STE MCAS test items showing importance for their role in ELs' performance on these items. In addition, the process of developing test item modifications in Study 2, and asking state test developers to review possible modifications, highlighted for researchers the requirements of test item development from the state's perspective, including alignment to state standards and formatting requirements.

The results of the development of professional vision for test developers include a growing sensitivity to the particular features of test items identified in Study 1 as correlated with

DIF favoring non-ELs over ELs: FC, RB, LFNT vocabulary, and an increased recognition of the importance of using Technical vocabulary and Visuals in test items as supports for ELs. Previously, item development had not focused in particular on the issues of comprehensibility of test items for ELs.

The results of the development of professional vision for researchers include an increased awareness of the structures and policies that guide, and at times constrain, test design. This has led researchers to strive to suggest test item modifications that would be practical for test developers to implement in future test items. In addition, researchers have collaborated with test developers to create a handbook for test item writers summarizing the findings of the project in a way that is usable for test item development and review. The handbook is intended not only for state test developers, but also for testing contractor staff and school and district leaders who design and develop assessments.

Ongoing Collaboration

Finally, the work of this project has led to an ongoing collaboration extending the work of the English Learners and Science Tests project with a new IES project, Learning about Open-Response Science Test Items and English Learners. In this new project, we will be collaborating to learn and draw conclusions about the reading and writing demands of the Grade 5 STE MCAS open-response items on ELs. This new work will build upon our research on the language of multiple-choice items for this test.

Significance

This collaboration has facilitated research that is informed by the larger field of assessment research, but also focused specifically on one state's tests. As a result, the findings of this work have been directly applicable to the challenges faced by state test developers in

improving their science test items. In addition, the collaboration has both informed and facilitated the work of both researchers and test developers, leading to an increasing overlap in perspectives on test items in these two groups, and an ongoing research partnership.

Author Note

The research reported here 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 Carrie Conaway, Darin LaSota, Craig S. Wells, Beth Warren, and Mary Catherine O'Connor for their contributions to the research reported herein. The authors would also like to thank all of the district and school administrators, teachers, and students who participated in this research. This paper is dedicated to them.

References

Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2000/2005). *The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation from Three Perspectives* (CSE Report No. 663). Retrieved from University of California, National Center for Research on Evaluation, Standards, and Student Testing website: http://www.cse.ucla.edu/products/reports.asp

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.

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.

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.

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

Goodwin, C. (1994). Professional Vision. *American Anthropologist, 96*(3), 606-633.

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.

Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English-language learners. *Applied Measurement in Education, 16*(2), 159-188. doi: 10.1207/S15324818AME1602_4

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*.

Lord, C., Abedi, J., & Poosuthasee, N. (2000). *Language difficulty and assessment accommodations for English language learners*: Study Commissioned by the Delaware Department of Education (https://www.doe.k12.de.us/).

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

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review (*http://her.hepg.org*/), 78*(2), 333-368.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and Differential Item Functioning for English Language Learners in math tests. *Educational Assessment, 14*(3), 160-179. doi: 10.1080/10627190903422906

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.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States* (Vol. 1). Washington, DC: The National Academies Press.

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*.

Penfield, R. D., & Lee, O. (2010). Test-based accountability: Potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching, 47*(1), 6-24.

Proctor, C. P., & Silverman, R. D. (2011). Confounds in assessing the associations between biliteracy and English language proficiency. *Educational Researcher*, *40*(2), 62-64.

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

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

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

Wolf, M. K., Kim, J., & Kao, J. (2012). The effects of glossary and read-aloud accommodations on English language learners' performance on mathematics assessment. *Applied Measurement in Education, 25*(4), 347-374.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*(3), 139-159. doi: 10.1080/10627190903425883

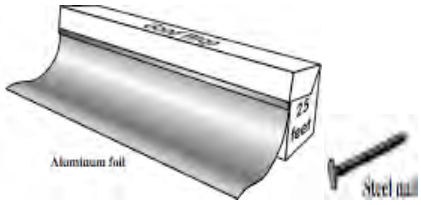
Zeno, S., Ivens, S. H., Millard, R. T., & Rothkopf, E. Z. (1995). *The educator's word frequency guide.* Brewster, NJ: Touchstone Applied Science Associates.

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?

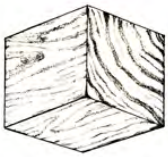
*A. They are metals.
B. They rust rapidly.
C. They weigh the same.
D. They are the same color.

Figure 2. Flexible Object 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

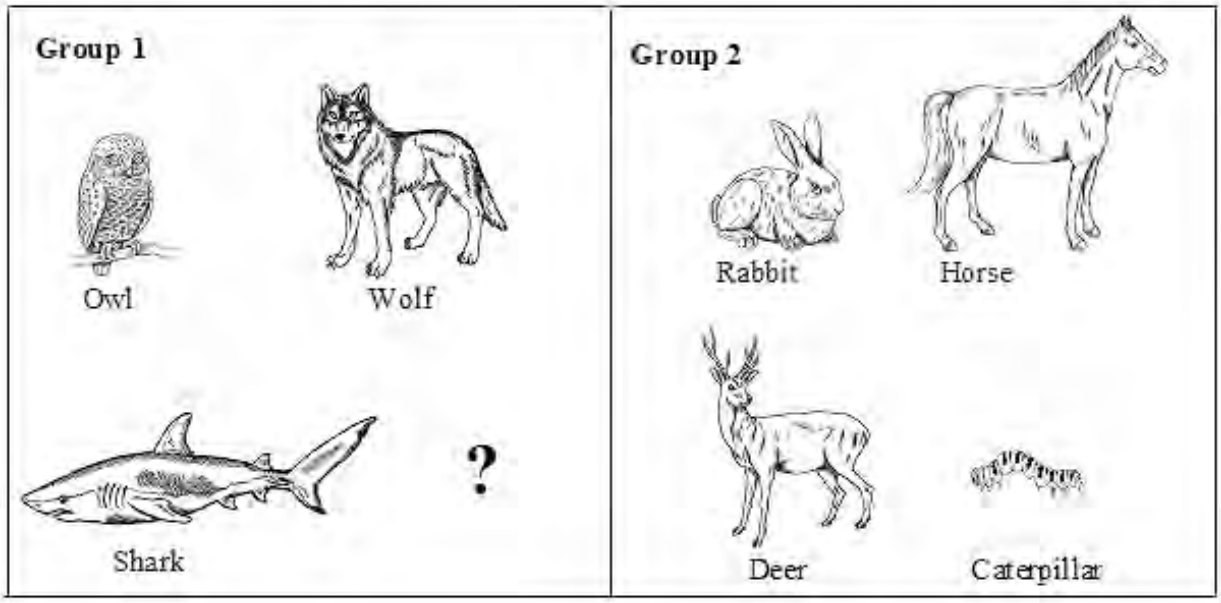
Grouped Animals Original
The lists below show animals separated into two different groups.

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

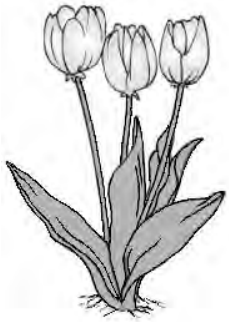
goat

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

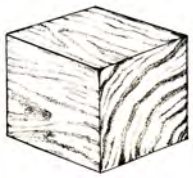
Ceramic Dish Modified (Visual Added)

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