*American Educational Research Journal*
*October 2020, Vol. 57, No. 5, pp. 2175–2209*
*DOI: 10.3102/0002831220905562*


# Targeted Linguistic Simplification of Science Test Items for English Learners

Tracy Noble
*TERC*
Stephen G. Sireci
Craig S. Wells
*University of Massachusetts Amherst*
Rachel R. Kachchaf
*Smarter Balanced*
Ann S. Rosebery
*Chèche Konnen Center*
Yang Caroline Wang
*Education Analytics*

*In this experimental study, 20 multiple-choice test items from the Massachusetts Grade 5 science test were linguistically simplified, and original and simplified test items were administered to 310 English learners (ELs) and 1,580 non-ELs in four Massachusetts school districts. This study tested the hypothesis that specific linguistic features of test items contributed to construct-irrelevant variance in science test scores of ELs. Simplifications targeted specific linguistic features, to identify those features with the largest impacts on ELs' test performance. Of all the linguistic simplifications used in this study, adding visual representations to answer choices had the largest positive effect on ELs' performance. These findings have significant implications for the design of multiple-choice test items that are fair and valid for ELs.*

KEYWORDS: assessment, bilingual, English learners, linguistic simplification, science

Standards-driven education reform legislation such as the No Child Left Behind Act ([NCLB], 2002) is intended to address persistent gaps in test scores between White middle- and upper-class students who speak English as a first language and students from historically underserved communities. One such underserved group is students who, under the Every Student Succeeds Act ([ESSA], 2015), are classified as English learners (ELs). This classification is based on a set of criteria that typically include a home language survey and a test of English proficiency.

We recognize the limitations of classifying students as ELs, which include (1) the lack of acknowledgment of students' emerging multilingual status (Hopewell & Escamilla, 2014; Ojeda, 2016); (2) the diversity of the backgrounds, languages, language proficiencies, and life experiences that can be masked by grouping students together under the category ELs; and (3) the variations in and limitations of the criteria, including testing practices, used to classify students as ELs (Linquanti & Cook, 2013; Sireci & Faulkner-Bond, 2015; Proctor & Silverman, 2011). We nonetheless recognize that the category ELs identifies for schools, school districts, and states, a group of students who have historically not received adequate attention from the research, assessment, and policy communities, and for this reason, we use the category name, while acknowledging its limitations.

---

Tracy Noble, PhD, is a project leader at TERC, 2067 Massachusetts Avenue, Cambridge, MA 02140; e-mail: *tracy_noble@outlook.com*. Her research focuses on the experiences of English learners with science and mathematics assessments and curricula. The goal of her work is to produce findings that allow educators, curriculum developers, assessment developers, and policymakers to make choices that increase fairness and learning opportunities for English learners in the U.S. education system.

Stephen G. Sireci, PhD, is distinguished university professor and director of the Center for Educational Assessment at the University of Massachusetts Amherst. He earned his PhD in psychometrics from Fordham University and his master's and bachelor's degrees in psychology from Loyola College in Maryland. His research focuses on fairness, efficiency, and accuracy in educational assessment.

Craig S. Wells, PhD, is an associate professor at the University of Massachusetts Amherst in the Research in Educational Measurement and Psychometrics concentration and also serves as associate director in the Center for Educational Assessment. Dr Wells teaches courses in educational statistics and psychometric theory and his research interests include the applications of item response theory (IRT), specifically related to the assessment of model fit, detection of differential item functioning, and applications of nonparametric IRT.

Rachel R. Kachchaf, PhD, is the senior director of Student Supports for the Smarter Balanced Assessment Consortium and oversees the Consortium's efforts to consider the needs of diverse learners, including English learners, students with disabilities, and English learners with disabilities throughout the full assessment system. Her research interests center on improving large-scale assessment for diverse learners.

Ann S. Rosebery, EdD, is co-director of the Chèche Konnen Center. Her research focuses on developing learning environments that support students from nondominant groups to learn complex scientific ideas (e.g., human microbiome) and sociocritical perspectives on those ideas in order to navigate the world in agentic ways. A complementary strand of research conceptualizes teacher professional learning as the cultivation of interpretive power through teacher inquiry that centers issues of race, power, disciplinary canons, and students' sensemaking.

Yang Caroline Wang, PhD, is a research scientist at Education Analytics. Her research focuses on IRT modeling, assessment/survey development and validation, and assessment-related issues in growth models.

Science tests at the state and national levels show persistent test score gaps between ELs and non-ELs (Caldas, 2013; Kober et al., 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 science testing differentially affect ELs and their teachers, schools, and communities (Caldas, 2013; McIntosh, 2011). 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 (Noble et al., 2012; Pellegrino et al., 2001). Despite the stated goals of NCLB to close such test score gaps, they persist on large-scale assessments at the state and national levels (Caldas, 2013; Kober et al., 2010; Lee & Reeves, 2012), resulting in increasingly substantial consequences for EL students, their teachers, and schools (Huddleston, 2014; Pennsylvania Clearinghouse for Education, 2013; Sims, 2013).

Since the passage of NCLB in 2002, there has been an increase in interest from the research and policy communities in mathematics and English language arts testing. However, science testing has received comparatively less attention from the research community (Noble et al., 2018), due in part to its relatively late inclusion, in the 2007–2008 school year, in NCLB-mandated accountability measures. Since 2008, science test scores have been consistently included in high-stakes NCLB-mandated accountability measures and ESSA has maintained similar requirements. More recently, new provisions of ESSA have allowed states greater flexibility in choosing the tests used for annual reporting purposes (Gewertz, 2015), increasing the need for research-based guidelines for judging the fairness and validity of science tests.

### Construct-Irrelevant Variance and Assessment of ELs

The need for attention to science tests is underscored by the work of researchers who have questioned the validity of interpreting ELs' science test scores as measures of ELs' science content knowledge, given that those science tests are written by and for speakers of English as a first language (Abedi, 2002, 2006; Hakuta & Beatty, 2000; Solano-Flores & Gustafson, 2012). Hakuta and Beatty (2000) caution that a test intended to measure students' content knowledge cannot provide valid information about students' knowledge if "a language barrier prevents the students from demonstrating what they know and can do" (p. 20). In addition, the assessment community has expressed concern that ELs' scores on high-stakes science tests include construct-irrelevant variance (Abedi, 2006; Solano-Flores, 2008). Construct-irrelevant variance is systematic variance that arises when an assessment actually measures something *other than the construct the test is intended*

*to measure* (American Educational Research Association [AERA] et al., 2014; Haladyna & Downing, 2004; Sireci & Faulkner-Bond, 2015). 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 et al., 2014, p. 60)

Science tests written in English, particularly when they contain unnecessary linguistic complexity (Abedi, 2006), may be in part measuring ELs' levels of English proficiency, and thus, ELs' scores on such tests may include construct-irrelevant variance.

### Construct-Irrelevant Variance and the Language of Science Test Items

Large-scale studies of ELs' interactions with test items have provided evidence that science and mathematics tests written in English are in part measuring ELs' levels of English proficiency. For example, research on language-focused accommodations such as translation of tests into students' first languages, dual-language test booklets, English or bilingual glossaries of non-content words, and linguistic simplification of test item language, have found that these accommodations can lead to improvements in ELs' test scores (Abedi, 2008; Abedi & Ewers, 2013; Kieffer et al., 2009; Kieffer et al., 2012; Li & Suen, 2012; Pennock-Roman & Rivera, 2011, 2012; Sireci et al., 2003). Testing accommodations are changes to the testing conditions that do not affect the construct being measured, such as knowledge and skills in mathematics and science (Butler & Stevens, 1997). Thus, when ELs who do not receive accommodations score significantly lower than ELs receiving accommodations, these differences are due to English language proficiency of tested students, and not due to students' content knowledge and skills in mathematics and science.

While multiple reviews and meta-analyses agree upon the conclusion that language-based accommodations can lead to improvements in test performance for ELs, they differ about which types of accommodations most help ELs: translation, dual language test booklets, provision of English dictionaries, provision of glossaries with or without Spanish translation, or linguistic simplification (Abedi, 2008; Abedi & Ewers, 2013; Kieffer et al., 2009; Kieffer et al., 2012; Li & Suen, 2012; Pennock-Roman & Rivera, 2011, 2012; Sireci et al., 2003). Reviews and meta-analyses by design compile the findings of multiple different studies involving different test items, student samples, and methods of providing linguistic simplification and other accommodations. For example, in our analysis (Noble et al., 2018) of 10

linguistic simplification studies from 2000 to 2013, we found over 50 different linguistic features (e.g., unfamiliar words, complex grammatical structure, long test items) simplified across studies, with each study focused on a different set of linguistic features of test items. Thus, while these reviews and meta-analyses support the conclusion that the language of science and mathematics test items can interfere with the performance of ELs on those items, they do not agree on how to best address the problem of unnecessary linguistic complexity in test items and construct-irrelevant variance in ELs' test scores.

### The Language of Science Tests

New science tests aligned to the Next Generation Science Standards (NGSS Lead States, 2013) are expected to have increased English language demands, due to the focus of the NGSS on communication in science (National Research Council, 2014a, 2014b). Researchers have expressed concern that such tests may introduce additional English language demands that are not part of the constructs the tests are intended to measure, and that these demands may differentially affect EL test-takers (Abedi & Linquanti, 2012). Others have argued that the English language demands of science tests written in English are part of the construct that science tests are intended to measure, if the language of test items is consistent with the language of the discipline (Avenia-Tapper & Llosa, 2015). While we believe that some forms of language are appropriate targets of instruction, we do not believe that interpreting and producing complex academic language in English is the only way in which students can communicate about science, even though many testing regimes limit the ways students can demonstrate this ability. Furthermore, as the linguistic simplification studies described above have demonstrated, test items in science and mathematics frequently contain unnecessary linguistic complexity that negatively affects ELs' test scores and is neither required to convey the science or mathematics content of the test items, nor part of scientific or mathematical practice (Abedi, 2006).

To address the increasing challenges of assessing ELs' science and mathematics knowledge and skills using valid measures, the field of assessment research and development needs a better understanding of how ELs interact with science and mathematics test items written in English. The main study reported herein focuses on science test items, because they have historically received less attention from assessment researchers than mathematics test items (Noble et al., 2018). The goal of this study is to understand which linguistic features contribute the most unnecessary linguistic complexity to science test items, and thus, most affect ELs' performance on those items. We wish to inform the work of test designers, state departments of education, and assessment consortia, to help them to construct more valid and fair tests for both ELs and non-ELs.

The main study reported herein focuses on the Grade 5 Science, Technology, and Engineering Massachusetts Comprehensive Assessment System (STE MCAS), which is the Grade 5 test that fulfills the science accountability requirements of NCLB and ESSA for Massachusetts (MA Department of Elementary and Secondary Education [MA DESE], 2014a). In preparation for the main study, we sought to identify the linguistic features of multiple-choice test items on the Grade 5 STE MCAS that were most problematic for Grade 5 EL students in MA. The preparatory research we undertook to identify these linguistic features is described in more detail in other publications (Kachchaf et al., 2016; Noble et al., 2018) and briefly summarized in the "Preparatory Work" section that follows.

## Preparatory Work

### *Synthesizing the Literature*

We began our preparation for the study described herein with a synthesis of the literature on science and mathematics testing of ELs (Noble et al., 2018). The goal of this synthesis was to identify reports of research published between 2000 and 2013 showing evidence of the interaction between specific linguistic features of science and mathematics test items and the performance of ELs on those items. Research on mathematics test items was included in the review because it has yielded findings relevant to science test items. From 231 reports identified in our initial searches, we identified a subset that included 16 reports of original studies testing hypotheses regarding the effects of specific linguistic features of science and mathematics test items on ELs' test performance. Over 60 linguistic features were identified across this subset of studies, and we found that 16 of these linguistic features (e.g., unfamiliar words, complex grammatical structures of various kinds, and item length) were each identified by three or more studies as affecting performance of ELs.

### *Correlation of Features With Differential Item Functioning*

To determine how the 16 features identified in the literature synthesis are related to ELs' performance on the Grade 5 STE MCAS, we conducted a correlation study (Kachchaf et al., 2016). We collected all 162 released multiple-choice Grade 5 STE MCAS items from the years 2004 to 2010 that were keyed to the standards within the Earth and Space Science, Life Science, and Physical Sciences topics. We did not include the items keyed to standards within the Technology/Engineering Design topic due to information collected from an informal survey of MA teachers indicating that this content is not consistently taught. We coded each of these 162 items for 11 linguistic features selected, based on their frequency of occurrence in STE MCAS items, from the 16 features identified in the literature synthesis.

We also coded the items for two additional features identified in cognitive interview studies and item analyses as problematic for ELs answering STE MCAS items (Noble et al., 2012), leading to a total of 13 coded linguistic features. We correlated the presence of each of these 13 linguistic features with differential item functioning (DIF) for ELs compared with non-ELs for all 162 test items in the sample, to identify those test items that were differentially harder for ELs compared with non-ELs. Out of the 13 features investigated in that study, 3 features (Low-Frequency Nontechnical Vocabulary, Forced Comparison, and Reference Back) were correlated at a statistically significant level with *higher* levels of DIF favoring non-ELs over ELs. In other words, these three linguistic features appeared more often in test items on which ELs scored lower than non-ELs who scored similarly on the rest of the items on the test. Two other features (Technical Vocabulary and Visuals) were correlated at a statistically significant level with *lower* levels of DIF favoring non-ELs over ELs, that is, appeared to be helpful to ELs. All five of the features found to be correlated with DIF at a statistically significant level are defined in detail in the "Method" section for the main study. The findings of the correlation study provided one indication of how the linguistic features identified in prior research on ELs and science and mathematics testing might affect the test performance of Grade 5 ELs taking the STE MCAS. Further details regarding correlation study methods and findings can be found in Kachchaf et al. (2016).

In the main study, we used the results of this preparatory work to design targeted linguistic simplifications of multiple-choice science test items from the Grade 5 STE. Many prior linguistic simplification studies had the goal of demonstrating the effectiveness of linguistic simplification as an accommodation method, and the role of linguistic complexity, broadly conceived, in EL test performance (e.g., Abedi et al., 2003; Abedi et al., 2005; Abedi & Lord, 2001; Sato et al., 2010). Diverging from this work, we focused our linguistic simplifications on a small number of linguistic features identified in our preparatory work as problematic specifically for Grade 5 ELs taking the STE MCAS, in order to find out which of these features most affect ELs' performance on science test items.

## Method

The primary goal of the main study is to identify the linguistic features of multiple-choice science test items that have the largest effects on the performance of ELs taking the Grade 5 STE MCAS. To evaluate these effects, we performed targeted linguistic simplifications of released Grade 5 science test items from the STE MCAS, we administered original and linguistically simplified versions of the test items to ELs and non-ELs, and we evaluated the effects of linguistically simplified items using a variety of statistical procedures. Our research questions are as follows:

1. What are the effects of targeted linguistic simplifications of multiple-choice Grade 5 STE MCAS test items on the performance of ELs and on the performance of non-ELs who score at varying levels on the MCAS English language arts (ELA) test?
2. How do ELs with varying levels of English proficiency respond to the linguistic simplifications?
3. Which specific linguistic simplification types and individual item linguistic simplifications had the greatest effects on the performance of ELs?

## Measures

The three assessments used in this study were used to fulfill state and federal accountability requirements for MA. They are the Grade 5 STE MCAS, the Grade 5 ELA MCAS, and the MA English proficiency assessment for ELs, the Assessing Comprehension and Communication in English State to State (ACCESS) for English Language Learners (WIDA, 2014). The ELA MCAS and the ACCESS test were used to explore differences in performance across subgroups within our samples. The psychometric properties of these assessments are detailed in the associated Technical Reports (Center for Applied Linguistics, 2015; MA DESE, 2014b), as required under NCLB and ESSA. Table 1 provides details about each assessment used in this study.

## Participants

Students were recruited to participate in this study from four urban school districts in MA. All four school districts had large EL populations and a majority of students were considered economically disadvantaged in three of the four districts. Table 2 provides information about the Grade 5 populations in the four districts in MA compared with the U.S. Grade 5 population (National Center for Education Statistics 2015, 2019). As can be noted in Table 2, MA's EL population has a smaller overall proportion of speakers of Spanish as a first language than the United States as a whole, although some districts within MA have populations of ELs that reflect the Spanish-dominance of the EL population in the country as a whole. As in much of the United States, the districts in this study and MA as a whole reflect the broad linguistic diversity of the country, including speakers of over 100 languages in MA, and speakers of multiple dialects of these languages.

Table 3 provides details regarding our sample of Grade 5 students,[1] including the numbers of students for whom we have scores from the measures described in Table 1.

## Test Item Simplification

The process used to simplify the test items is described in detail elsewhere (Noble et al., 2016), and summarized here. The process involved

*Table 1*
**Assessments Used in the Study**

| Variable | STE MCAS | ELA MCAS | ACCESS |
|---|---|---|---|
| Administered to | All Grade 5 students in state | All Grade 5 students except for EL students in first year in U.S. schools | All Grade 5 EL students in state |
| Multiple-choice items per year | 38 | 36 | 48 |
| Open response items per year | 4 | 4 | 4 |
| Other format items per year | — | — | 3 Scripted interview questions |
| Topics | Earth and Space Science, Life Science, Physical Sciences, and Technology/ Engineering Design | Reading and English Language | Reading, Writing, Listening, and Speaking in English |
| Scaled scores | 200–280 | 200–280 | 100–600 |
| Proficiency | 240 and above | 240 and above | 500 and above |
| Our focus | Multiple-choice items<br>Earth and Space Science, Life Science, Physical Sciences topics | Overall score | Reading score |

*Note.* EL = English learner; STE MCAS = Grade 5 Science, Technology, and Engineering Massachusetts Comprehensive Assessment System; ELA MCAS = English Language Arts Massachusetts Comprehensive Assessment System; ACCESS = Assessing Comprehension and Communication in English State to State.

several steps, including (1) identifying items to simplify, (2) developing item simplification methods and initial simplified test items, (3) review of

*Table 2*
**Size of Grade 5 EL Populations in Study Year**

| Population | Grade 5 Students | Percentage EL Students | Percentage of EL Students With First Language Spanish |
|---|---|---|---|
| United States | 3,710,000 | 9.8 | 76 |
| Massachusetts | 72,048 | 8 | 49 |
| District A | 300–400 | 17 | 80–90 |
| District B | 400–500 | 8 | 40–50 |
| District C | 600–700 | 12 | 40–50 |
| District D | 1,800–2,000 | 18 | 80–90 |

*Note.* EL = English learner.

*Table 3*
**Sample Sizes for Analyses**

| Group | STE MCAS | ACCESS | ELA MCAS |
|---|---|---|---|
| ELs | | | |
| Girls | 165 | 165 | 155 |
| Boys | 145 | 145 | 125 |
| Total | 310 | 310 | 280 |
| Non-ELs | | | |
| Girls | 825 | – | 780 |
| Boys | 755 | – | 704 |
| Total | 1,580 | – | 1,484 |
| ELs first language Spanish | 234 | 234 | 210 |
| ELs other first language | 76 | 76 | 70 |

*Note.* EL = English learner; STE MCAS = Grade 5 Science, Technology, and Engineering Massachusetts Comprehensive Assessment System; ELA MCAS = English Language Arts Massachusetts Comprehensive Assessment System; ACCESS = Assessing Comprehension and Communication in English State to State.

simplified items by expert panels, and (4) refining linguistic simplifications based on cognitive interviews and pilot testing.

### *Identifying Items to Simplify*

Test items to simplify were drawn from the 162 released multiple-choice Grade 5 STE MCAS test items from the years 2004 to 2010 that were coded for the correlation study described in the "Preparatory Work" subsection. From this set of 162 items, 32 items were selected for linguistic simplification based on the presence of the linguistic features that were targeted for simplification and statistical criteria, including DIF values flagging the items as more

difficult for ELs than non-ELs with similar science test scores (Kachchaf et al., 2016).

*Developing Item Simplification Methods and Initial Simplified Items*

Our linguistic simplifications either (1) removed or reduced linguistic features associated with lower performance for ELs or (2) added features associated with higher performance for ELs. The three item features we *removed* were Low-Frequency Nontechnical Vocabulary, Forced Comparison, and Reference Back. The one feature we *added* was a Visual (picture, diagram, or table), which was considered a linguistic simplification, due to evidence that the presence of a visual can mitigate the effects of linguistic complexity of test items on ELs (Martiniello, 2009). The four simplifications we developed based on these features are described in the remainder of this section.

Low-Frequency Nontechnical Vocabulary are words that occur infrequently in fifth-grade texts, and do not have a primarily scientific meaning. Examples of Low-Frequency Nontechnical words in the Grade 5 STE MCAS are "crumble," "hose," "repeatedly," and "unusually." We classified words as low frequency following the procedures of Butler et al. (2004) and excluded from this set words that were identified as technical due to having a primary meaning associated with a scientific discipline or one of the MA state science standards (Kachchaf et al., 2016). Our Low-Frequency Nontechnical simplification involved replacing each Low-Frequency Nontechnical word with a higher frequency synonym, unless the original word was judged to be integral to the science content of the item. For example, in a test item asking: "What can make a rock crack and crumble?" the Low-Frequency Nontechnical word "crumble" could be replaced with the higher frequency word "break" without changing the scientific meaning, leading to the simplification: "What can make a rock crack and break?"

The Forced Comparison feature occurs in test items requiring students to compare all answer choices to select the one that is, in the wording of the item, the "best," or "most likely." That is, the correct answer choice is at an extreme on a scale defined by the test item. For example, an item that asks: "Which of the following is the **most likely** result of heating water in a pan?" has the Forced Comparison feature. A simplification of this test item is "What happens when you heat water in a pan?" To create this and other Forced Comparison simplifications, we removed the extreme value descriptor "most likely" and removed the vague noun associated with it, "result." We also replaced the complex question phrase "Which of the following" with the question word "What."

The Reference Back feature occurs in test items in which the question sentence requires students to refer back to information given earlier in the item. For example, in the question "How would the earthworm respond to this change?" the noun phrase "this change" refers back to prior sentences

of the item. For each test item with the Reference Back feature, we added to the question sentence the information needed from prior sentences (e.g., "How would the earthworm respond to the lights?").

The Visual feature occurs in test items with a visual representation in the form of an image illustrating some object(s) described in the item stem (the part of the item prior to the answer choices) or the answer choices. The Visual simplification involves adding one or more visual representations to an item or enhancing an existing visual representation.[2] For example, Figure 1 shows a test item to which we added visuals to illustrate the answer choices, as shown in Figure 2.

### *Review of Simplifications*

Throughout our development of simplified test items, a panel of experts in linguistics, STE assessment, and EL education reviewed all simplifications for their effectiveness in altering the intended linguistic feature without altering the target science content. In addition, all original and simplified test items were reviewed by experienced school district science coordinators and practicing scientists to verify that the simplifications did not alter the science knowledge or science task targeted by the original test items.

### *Refining Simplifications*

We tested our original set of item simplifications (Low-Frequency Nontechnical words, Forced Comparison, Reference Back, and Visuals) applied to 32 released Grade 5 STE MCAS multiple-choice test items in 16 cognitive interviews with Grade 5 EL students with varying English proficiency levels and first languages. Each student was interviewed about a combination of original and simplified test items. However, no student was interviewed about both the original and the simplified form of the same item. We found that the Visual simplification consistently improved ELs' item comprehension and performance, but that the other simplifications did not. As a result, we were forced to reconsider our simplification method. Our original method was not designed to produce the ideal simplification for each item, but instead to investigate the effects of individual item features on student performance. However, our initial cognitive interview data suggested that simplifying individual item features was unlikely to yield measurable improvements in EL performance on the items. For this reason, we retained the targeted nature of our simplifications, but altered our method as described below.

### *Dual-Feature Simplifications*

We hypothesized that simplifying one language-based feature (Low-Frequency Nontechnical words, Forced Comparison, or Reference Back)

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

*Figure 1.* **Item 200008001 original (MA DOE, 2005).**

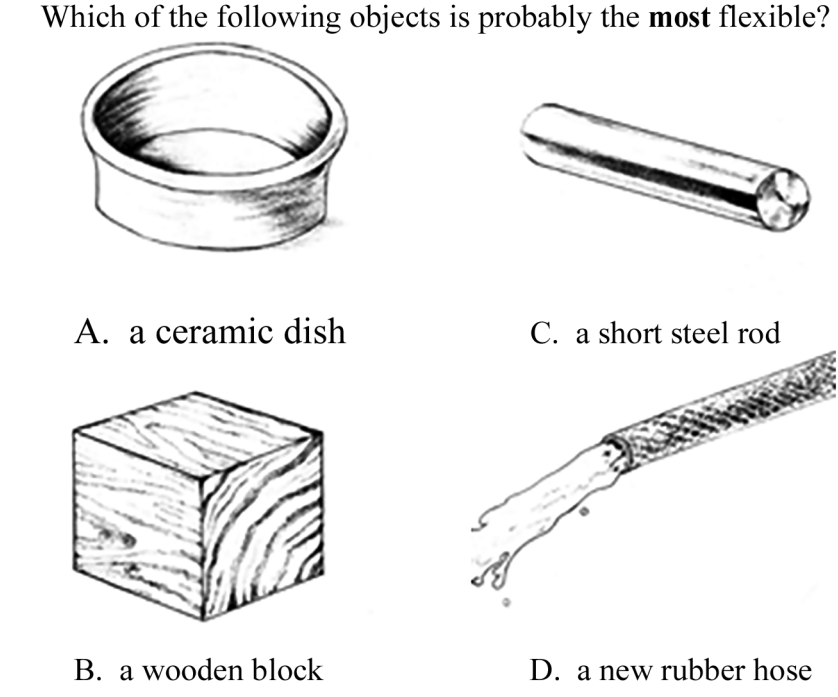


*Figure 2.* **Item 200008001 simplified: Visual added.**

may have been insufficient to improve ELs' item comprehension and performance because of the other interfering features of the items. Thus, we developed a new set of simplifications in which pairs of regularly co-occurring features were simplified for each test item. For example, the Forced Comparison and Reference Back features frequently co-occurred, as in this original test item, which has the Reference Back feature italicized and the Forced Comparison feature bolded: "Ricardo has an igneous rock in his rock collection. Where did *this rock* **most likely** form?" (Massachusetts Department of Education [MA DOE], 2004). The dual-feature Forced Comparison and Reference Back simplification of this item is "Ricardo has an igneous rock in his rock collection. Where do igneous rocks form?" We intended these dual-feature simplifications to sufficiently clarify the language of test items to improve performance for ELs, while at the same time remaining sufficiently targeted to allow us to determine which features had the greatest effect on EL performance.

*Interview-Based Simplifications*

Our initial cognitive interviews surfaced additional features that we had not previously identified as problematic for ELs. For example, in one test item, the word "damp" was found to be unknown to a number of the EL students whom we interviewed, even though it had not been coded as Low-Frequency Nontechnical. As described in previous publications (Kachchaf et al., 2016; Noble et al., 2016), the Low-Frequency Nontechnical classification is based on a word frequency list published in 1995 (Zeno et al., 1995), which was the best resource available to us at the time of our study. However, EL students in MA may have different experiences of written English than those documented by Zeno et al. in 1995. Thus, we chose to use our interviews with ELs to find problematic language in test items that our coding methods had not identified, including additional unfamiliar words in test items, polysemous words with a familiar meaning inconsistent with the meaning intended in the test item, confusing timelines in item stems, and visuals needed in items that had not been targeted for the Visual simplification. Based on this information, we created Interview-based simplifications for some test items. Each Interview-based simplification also included changes to one or more previously identified item features (i.e., Forced Comparison, Reference Back, Low-Frequency Nontechnical words, or addition of Visuals).

As a result of these two changes in our item simplification method, our simplifications became less targeted, but more likely to affect EL performance. In addition, the Interview-based simplification provided the opportunity to see the effects of feature-driven targeted simplifications versus simplifications driven both by item features and by findings of cognitive interviews with EL students from the population for whom the test was intended.

The updated simplifications were tested in a second set of interviews with a similarly diverse group of 36 Grade 5 EL students, five of whom were interviewed twice, for a total of 41 interviews about test items simplified according to four updated simplification types: (1) the Forced Comparison and Reference Back simplification, (2) the Forced Comparison and Low-Frequency Nontechnical simplification, (3) the Visual simplification, and (4) the Interview-based simplification. The findings from these 41 interviews suggested that each of these simplifications improved EL performance, and that 24 of the 60 different test item simplifications we tested were the most effective in improving ELs' item comprehension and performance.

These 24 item simplifications were then pilot-tested with a sample of 50 Grade 5 EL and non-EL students from an urban school district in MA, and reduced to a set of 20 item simplifications when we learned that the pilot test form was too long for some students. The 20 test items we chose to simplify for the full-scale study cover 14 different standards from across Earth and Space Science, Life Science, and Physical Sciences, but are not as comprehensive in coverage as a whole STE MCAS test, which contains 38 test items. Further details about each simplification type and the linguistic simplification process can be found in other writing about this study (Kachchaf et al., 2014; Noble et al., 2014; Kachchaf et al., 2016).

### Research Design and Procedure

To evaluate differences in students' performance across the original (unsimplified) and simplified items, we used an experimental design in which both ELs and non-ELs took tests consisting of both original and simplified items. Two different test versions, Test Version A and Test Version B, were created so that no student would see both the simplified and the original version of the same item. Student participants were randomly assigned one of the test versions. Each test version had a total of 26 multiple-choice items, including 20 experimental test items (consisting of 10 test items in their original form and 10 test items in their simplified form) and 6 "anchor" items, common to both test forms. The anchor items were chosen for their lack of linguistic complexity, negligible levels of EL/non-EL DIF, and coverage of the standards within the Earth and Space Science, Life Science, and Physical Sciences topics. A student's score on the six anchor items was used as an independent measure of the student's science test performance for analysis purposes.

The 10 experimental test items in Item Set 1 were presented in their original form in Test Version A and were presented in their simplified form in Test Version B. The 10 experimental test items in Item Set 2 were presented in their original form in Test Version B and in their simplified form in Test Version A (see Table 4). All participating students saw both Item Sets 1 and 2, but students who saw Item Set 1 in original form saw Item Set 2 in

*Table 4*
**Research Design**

| Test Version | Original Items | Simplified Items | Anchor Items |
|---|---|---|---|
| A | Item Set 1: 10 items | Item Set 2: 10 items | 6 items |
| B | Item Set 2: 10 items | Item Set 1: 10 items | 6 items |

*Note*. "Original" refers to the unsimplified versions of the items.

simplified form, and vice versa. As a result, no student saw both the original and simplified version of any test item. The anchor items were the same in both test versions, in randomly determined positions that were kept fixed in all tests. The experimental (nonanchor) items were presented in two different, counterbalanced orders in Test Version A and in Test Version B, to control for fatigue effects. Table 4 presents the research design for the study.

The tests were administered between February 3 and March 7, 2014. 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 primarily retired teachers and school administrators. A maximum of one hour was allowed for students to complete the test, to accommodate school schedules.

Four percent of the students (71) did not fill in answers for between one and three test items. Half of 1% of students (8) did not fill in answers for four or more test items. Given that these unanswered items were distributed throughout students' test booklets, rather than being grouped at the end of students' tests, we considered unanswered items to be items that students were unable to answer rather than items students had not had an opportunity to see, and thus, scored them as incorrect responses. Tests were electronically scored and data were entered into spreadsheets for data analysis. Demographic data provided by the MA DESE and merged with our data set included students' gender and first language information, which was used to characterize our sample. We also collected information about students receiving special education services, which was used to exclude this student group from our main analysis, due to the potential for confounding of variables.

### Data Analysis

To evaluate the effects of the linguistic simplifications, we conducted analyses at the item set level, the linguistic simplification type level, and the individual item level. Each analysis is discussed in the sections that follow.

#### *Item Set–Level Analysis*

All students saw a test including both Item Sets 1 and 2. However, since each student saw one item set in original form and the other item set in

simplified form, we analyzed the findings for each item set separately. At the item set level, we used analysis of covariance (ANCOVA) to compare the raw scores of ELs and non-ELs on the original versus simplified versions of the items, using students' raw scores on the six anchor items as a covariate. The ANCOVA involved comparing three student groups—ELs, non-ELs who scored below "proficient" on the ELA MCAS (hereafter referred to as Non-ELs–Low), and non-ELs who scored at or above "proficient" on the ELA MCAS (hereafter referred to as Non-ELs–High). The dependent variable in these analyses was the total score on the 10-item set (Item Set 1 or Item Set 2). Each item was scored 0 for an incorrect or missing answer choice, and 1 for a correct answer choice. Language status was one independent variable with three levels (EL, Non-EL–Low, Non-EL–High), and format of the 10-item set (original or simplified) was the other independent variable. Total score on the six anchor items was the covariate.

We conducted two separate ANCOVAs: one for Item Set 1 and one for Item Set 2. These ANCOVA analyses allowed us to evaluate whether the simplified versions of the items were differentially easier for ELs, relative to their non-EL counterparts.[3] We expected to see ELs improve more on the simplified items than non-ELs, consistent with the differential boost hypothesis (Fuchs et al., 2000). In addition, we expected to see the greatest difference in improvement between ELs and Non-EL–High students, due to findings from prior research (e.g., Sato et al., 2010) indicating that linguistic simplification can benefit non-ELs with lower levels of English language proficiency, such as our Non-EL–Low group.

To investigate the effects of linguistic simplifications at the item set level on ELs at varying levels of English proficiency, we conducted an ANCOVA using data from ELs only, with reading subscores from the ACCESS English proficiency test as an additional covariate. In this analysis, we again compared how ELs did on the original versus simplified versions of the items, but we used two covariates: English language reading proficiency score and score on the anchor items. The English language reading proficiency score was used, rather than a score involving writing, speaking, or listening, to isolate the skills most needed to interact with multiple-choice test items written in English. This analysis allowed us to evaluate whether the linguistic simplification effect interacted with English reading proficiency (e.g., do the simplifications differentially affect ELs of lower or higher English reading proficiency?).

### *Evaluating Simplification Types and Simplified Items*

In addition to looking at overall differences in mean performance across student groups on original versus simplified item sets, we were also interested in exploring whether specific simplification types led to improved performance, and if so, which specific test items contributed to the success of

specific simplification types. Such effects could be overshadowed at the item set level. Therefore, we used an item response theory (IRT) approach that grouped together items based on how they were simplified to evaluate groups of items defined by simplification type. To follow up on this analysis, we evaluated the effects of individual test item simplifications, using a DIF approach. These procedures are described next.

*Comparison of simplification-type test characteristic curves.* To evaluate whether linguistic simplification effects were linked to one of the four specific simplification types (i.e., Forced Comparison & Reference Back, Forced Comparison & Low-Frequency Nontechnical, Visual, or Interview-based), we used IRT to calibrate all the test items so that simplification type–specific "test characteristic curves" (TCCs) based on items of the same simplification type could be created. Item difficulty and proficiency scores for examinees were estimated using the one-parameter logistic (1PL) IRT model,

$$P_i(\theta_j) = \frac{1}{1 + \exp\left[-D(\theta_j - b_i)\right]}, \quad (1)$$

where $P_i$ is probability of a correct answer on item $i$, $\theta_j$ (theta) is student $j$'s proficiency, $b_i$ is the difficulty parameter for item $i$, $D$ is a scaling factor equal to 1.7, and exp refers to the base of the natural logarithm.

The items were calibrated separately for each group defined by language status (ELs, Non-ELs–Low, and Non-ELs–High). The item difficulty parameter estimates (*b*-parameter estimates) for Test Version B were placed onto the same scale as those for Test Version A using the mean-sigma method. The *b*-parameter estimates were used to create mini-TCCs based on items of the same simplification type, in order to analyze the effects of specific simplification types on the test scores of ELs and non-ELs. To compute these TCCs, the item characteristic curves were summed for proficiency values ranging from −3 to 3 in increments of 0.1 for each item associated with each simplification type. The TCCs for original versus simplified items were then compared for each of the four simplification types.

*Differential item functioning analyses for test items.* DIF refers to a situation where an item is more difficult for one group of students compared with another, when students in the two groups are matched on overall proficiency (typically defined as overall test score). 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* [italics added]" (p. 31). 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 test score measured.

We used DIF analyses in the present study to determine whether the simplified 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 vs. non-ELs). The hypothesis underlying our linguistic simplifications is that the items would be easier for ELs after linguistic simplification, but they would *not* be easier for non-ELs after linguistic simplification. We used both logistic regression and an IRT-based method (Lord's chi-square) to detect DIF, 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., original) and focal (e.g., linguistic simplification) group, the logistic regression model may be specified as

$$P\left(u_{ij}=1\middle|\theta_j\right)=\frac{e^{\left[\tau_0+\tau_1\theta_j+\tau_2 g_j+\tau_3\theta_j g_j\right]}}{1+e^{\left[\tau_0+\tau_1\theta_j+\tau_2 g_j+\tau_3\theta_j g_j\right]}} \tag{2}$$

where $P\left(u_{ij}=1\middle|\theta_j\right)$ represents the probability of correctly answering item $i$ given examinee $j$'s proficiency, denoted $\theta_j$; $g_j$ is a dummy code used to represent whether examinee $j$ is in the reference ($g$ = 0) or focal ($g$ = 1) group; $\tau_0$ represents the intercept; $\tau_1$ represents the overall relationship between proficiency and the probability of a correct response; $\tau_2$ represents the difference between the reference and focal group, controlling for proficiency level; and $\tau_3$ corresponds to the interaction between group and proficiency, denoted $\theta_j\, g_j$. Uniform DIF refers to the situation where an item is differentially difficult for one group of students (or for one format of the item, as in this study) across the entire range of proficiency. Uniform DIF is indicated by $\tau_2 \neq 0$ and $\tau_3 = 0$. Nonuniform DIF refers to the situation where there is a difference in difficulty, but the degree of the difference changes at different levels of proficiency. Nonuniform DIF is represented by $\tau_3 \neq 0$, whether or not $\tau_2 = 0$.

In addition to evaluating original/simplified items for statistically significant DIF, we also used effect size criteria to determine whether any items flagged for DIF had nonnegligible effect sizes. Following Jodoin and Gierl's (2001) effect size guidelines, items with effect sizes greater than .035 (signifying the linguistic simplification status of the item accounted for 3.5% of the variation in item performance) were used to identify items as "medium DIF" (i.e., representing a moderate change in difficulty across original and simplified items) and effect sizes greater than .07 were used to identify items displaying "large DIF" (indicating a large change in difficulty).

The covariate in these logistic regression analyses ($\theta_j$) was based on raw scores. We used a two-stage purification approach, in which the DIF items

*Table 5*
**Means (Standard Deviations) for EL Groups and Linguistic Simplification Conditions**

| | Item Set 1 | | Item Set 2 | |
|---|---|---|---|---|
| Group | Original | Simplified | Original | Simplified |
| Non-EL–High ($n$ = 832) | 8.42 (1.40) | 8.54 (1.27) | 8.03 (1.52) | 8.09 (1.48) |
| Non-EL–Low ($n$ = 655) | 7.01 (1.92) | 7.35 (1.72) | 6.22 (1.97) | 6.40 (1.87) |
| EL ($n$ = 310) | 6.15 (2.26) | 6.71 (2.04) | 5.97 (2.23) | 5.65 (2.25) |

*Note.* "Original" refers to the unsimplified versions of the items. EL = English learner.

were initially identified using all items to represent $\theta_j$. A second analysis was conducted in which $\theta_j$ was based on only the non-DIF items. An $\alpha$ level of .05 was used to flag DIF items and the change in $R^2$ 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 results at the item set level using the ANCOVAs based on overall scores on the original and simplified item sets, then reporting the results at the simplification type level, and finally the item level.

### Analyses of Effects of Linguistic Simplifications on Test Scores

The results of the ANCOVAs for Item Sets 1 and 2 are summarized in Tables 5 through 7. Table 5 presents the descriptive statistics for all groups, and Tables 6 and 7 summarize the statistical significance tests for Item Sets 1 and 2, respectively. An inspection of the means in Table 5 reveals the largest differences were across Language status groups. Non-EL–High students had the highest means, and ELs had the lowest means for both item sets. This main effect was statistically significant for both item sets (Item Set 1: $F_{(2,\ 1790)} = 71.3$, $p < .001$; Item Set 2: $F_{(2,\ 1790)} = 95.9$, $p < .001$).

For Item Set 1, the pattern of mean differences across the original and simplified items was consistent with our hypotheses, including the differential boost hypothesis—all students performed better on the simplified items, with ELs showing the largest increase, followed by Non-EL–Low, and then Non-EL–High students (see Table 5). As shown in Table 6, the main effect of linguistic simplification was also statistically significant ($F_{(1,\ 1790)}$=12.6, $p < .001$). However, the effect size associated with this main effect was

*Table 6*
**Summary of ANCOVA Across EL and Non-EL Groups: Item Set 1**

| Source | SS | *df* | MS | *F* | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|
| Common items | 1195.7 | 1 | 1195.7 | 549.9 | <.001 | .24 |
| EL group | 310.1 | 2 | 155.1 | 71.3 | <.001 | .07 |
| Linguistic simplification | 27.4 | 1 | 27.4 | 12.6 | <.001 | .01 |
| Group × Linguistic simplification | 7.2 | 2 | 3.6 | 1.7 | .19 | .00 |
| Error | 3892.3 | 1,790 | 2.2 | | | |
| Total | 5432.7 | 1,796 | | | | |

*Note.* ANCOVA = analysis of covariance; EL = English learner; MS = mean square; SS = sum of squares.

*Table 7*
**Summary of ANCOVA Across EL and Non-EL Groups: Item Set 2**

| Source | SS | *df* | MS | *F* | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|
| Common items | 1475.2 | 1 | 1475.2 | 607.6 | <.001 | .25 |
| EL group | 465.4 | 2 | 232.7 | 95.9 | <.001 | .10 |
| Item simplification | 1.3 | 1 | 1.3 | 0.5 | .47 | .00 |
| Group × Linguistic simplification | 7.6 | 2 | 3.8 | 1.6 | .21 | .00 |
| Error | 4345.6 | 1,790 | 2.4 | | | |
| Total | 6295.1 | 1,796 | | | | |

*Note.* ANCOVA = analysis of covariance; EL = English learner; MS = mean square; SS = sum of squares.

negligible ($\eta^2$ = .01), and the EL Group-by-Item simplification interaction was *not* statistically significant ($F_{(2,\ 1790)}$ = 1.65, $p$ = .19).

For Item Set 2, the pattern of mean differences across the original and simplified items did not fit our hypotheses; the EL group did slightly worse on the simplified versions of the items, and the two Non-EL groups did slightly better (see Table 5). As shown in Table 7, the main effect for the linguistic simplification was *not* statistically significant $F_{(1,\ 1790)}$ = 0.5, $p$ = .47). The EL Group-by-Item simplification interaction was also *not* statistically significant $F_{(2,\ 1790)}$ = 1.6, $p$ = .21). Thus, these results did not support the differential boost hypothesis that the ELs would have larger score increases on the simplified versions of the items, relative to non-ELs.

We ran an additional ANCOVA on ELs, controlling for their English reading proficiency score as measured by the ACCESS test. In this analysis, we compared ELs' scores on the original and simplified versions of the items, but we used two covariates: their score on the anchor items as in the

*Table 8*
**Summary of ANCOVA Results for ELs Using English Reading Proficiency as a Covariate: Item Set 1**

| Source | SS | *df* | MS | *F* | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|
| Common items | 84.3 | 1 | 84.3 | 37.2 | <.001 | .11 |
| Reading proficiency | 271.2 | 1 | 271.2 | 119.6 | <.001 | .28 |
| Item simplification | 6.3 | 1 | 6.3 | 2.8 | .10 | .09 |
| Simplification × Reading proficiency | 2.3 | 1 | 2.3 | 1.0 | .32 | .00 |
| Error | 689.1 | 304 | 2.3 | | | |
| Total | 1053.2 | 308 | | | | |

*Note.* ANCOVA = analysis of covariance; EL = English learner; MS = mean square; SS = sum of squares.

*Table 9*
**Summary of ANCOVA Results for ELs Using English Reading Proficiency as a Covariate: Item Set 2**

| Source | SS | *df* | MS | *F* | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|
| Common items | 147.9 | 1 | 147.9 | 60.9 | <.001 | .17 |
| Reading proficiency | 225.2 | 1 | 225.2 | 92.8 | <.001 | .23 |
| Item simplification | 0.1 | 1 | 0.1 | 0.1 | .82 | .00 |
| Simplification × Reading proficiency | 0.0 | 1 | 0.0 | 0.0 | .99 | .00 |
| Error | 738.2 | 304 | 2.4 | | | |
| Total | 1114.4 | 308 | | | | |

*Note.* ANCOVA = analysis of covariance; EL = English learner; MS = mean square; SS = sum of squares.

previous ANCOVAs and their ACCESS reading subscores. As with the previous ANCOVAs, separate analyses were run for Item Sets 1 and 2.

The ANCOVA summary tables for these analyses are presented in Tables 8 and 9 for Item Sets 1 and 2, respectively. For both analyses, neither the main effect for Simplification nor the interaction of Simplification and English Reading Proficiency were statistically significant, although the main effect for Simplification approached statistical significance for Item Set 1 (with an $\eta^2$ of .09); however, that finding was not replicated for Item Set 2. These results indicate that overall, the simplifications did not have a statistically significant effect on the ELs' science test performance, and the lack of effect was consistent across ELs with various levels of English reading proficiency.

### Comparison of Linguistic Simplification-Type Test Characteristic Curves

As described in the Method section, the item characteristic curves for each version of each test item were summed within each linguistic

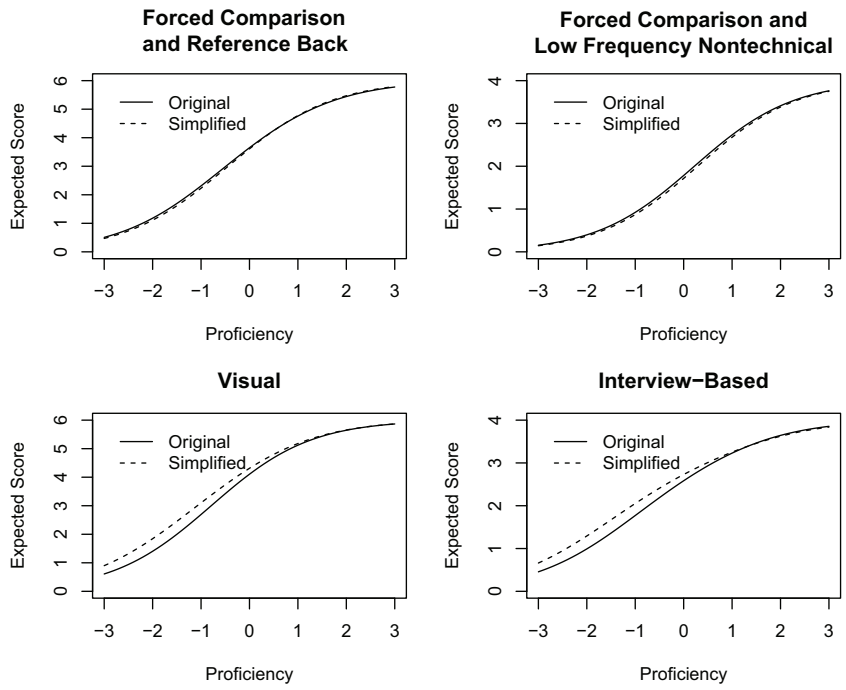


*Figure 3.* **Linguistic simplification type test characteristic curves for English learner students.**

*Note.* "Original" refers to the unsimplified versions of the items.

simplification type to form "linguistic simplification type" TCCs to investigate which specific linguistic simplification types had the greatest effects on the performance of ELs. Separate TCCs were computed for the EL, Non-EL-Low, and Non-EL-High groups. The original and simplified TCCs for each linguistic simplification type were plotted together to evaluate whether items undergoing the specific linguistic simplification were differentially easier (as a group of items) for a specified student group, relative to their original counterparts.

### *Test Characteristic Curves for ELs*

Figure 3 provides the TCCs for each linguistic simplification type based on the simplified 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 linguistic simplification types (1) Forced Comparison and Reference Back and (2) Forced Comparison and Low-Frequency Nontechnical were nearly identical for the original and

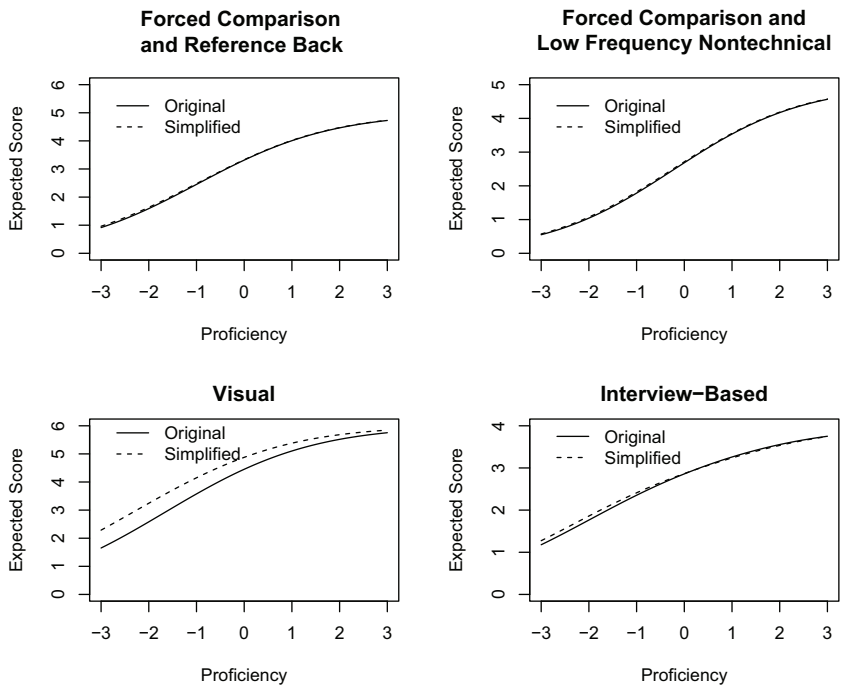


*Figure 4.* **Linguistic simplification type test characteristic curves for Non–English learner–Low students.**

*Note.* "Original" refers to the unsimplified versions of the items.

simplified items, indicating that these simplifications did not markedly change performance of ELs in the sample. However, the TCCs for the linguistic simplification types Visual and Interview-based indicate that the linguistic simplifications resulted in higher expected scores for the ELs, that is, that these linguistic simplifications made the items easier for ELs.

*Test Characteristic Curves for Non-EL-Low Students*

The TCCs for the Non-EL-Low students are presented in Figure 4. The TCCs for the linguistic simplification types (1) Forced Comparison and Reference Back, (2) Forced Comparison and Low-Frequency Nontechnical, and (3) Interview-based were nearly identical for the original and simplified items. However, the TCC for the Visual simplification type was higher for the simplified items, indicating that this linguistic simplification resulted in higher expected scores for the Non-EL-Low students (i.e., the linguistic simplification made the items easier for Non-EL-Low students).

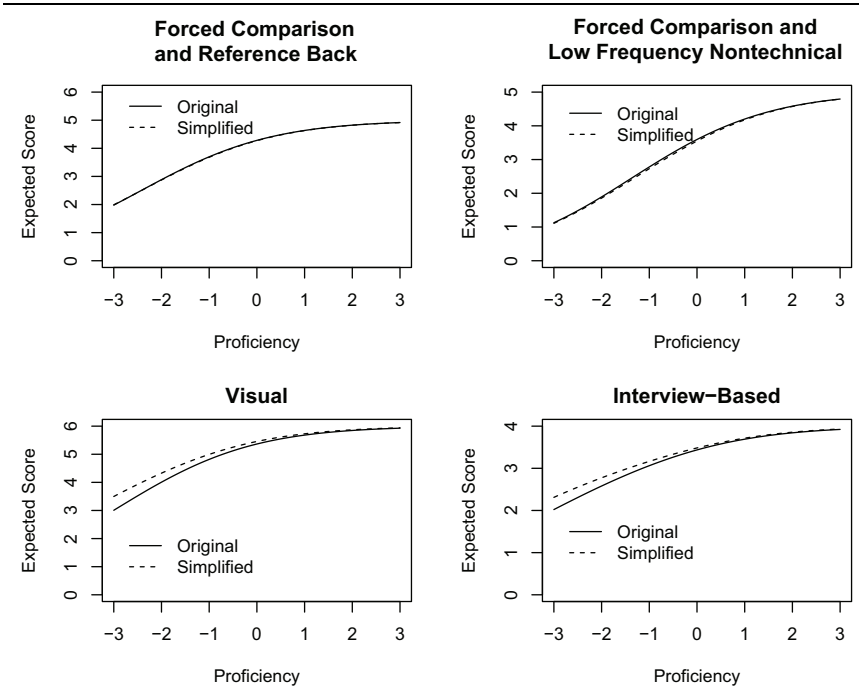


*Figure 5.* **Linguistic simplification type test characteristic curves for Non–English learner–High students.**

*Note.* "Original" refers to the unsimplified versions of the items.

### *Test Characteristic Curves for Non-EL-High Students*

For the Non-EL-High group, the TCCs for the linguistic simplification types (1) Forced Comparison and Reference Back and (2) Forced Comparison and Low-Frequency Nontechnical were nearly identical for the original and simplified items (see Figure 5). However, the TCCs for the linguistic simplification types Visual and Interview-based were higher for the simplified items, indicating that these two linguistic simplifications resulted in higher expected scores, but only for those Non-EL-High students scoring at the lower end of the proficiency scale.

## Differential Item Functioning Analyses

The results thus far have focused on total scores computed across sets of items (original or simplified). The DIF analyses provide a more powerful microscope to use to investigate the effect of *individual item* simplifications on student performance. The first set of DIF analyses focused on ELs. The

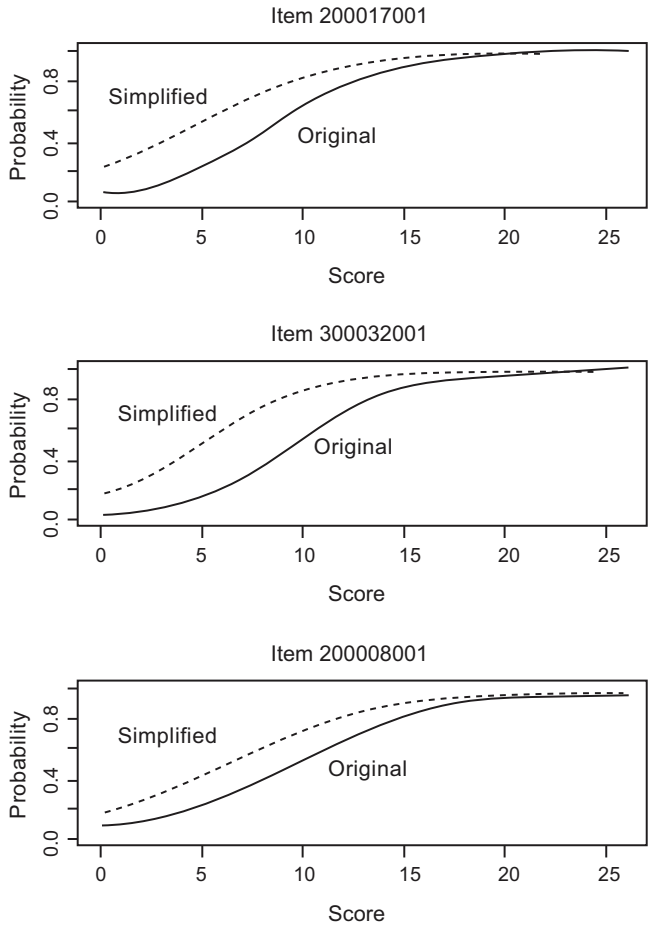


*Figure 6.* **Logistic curves for three items flagged for large differential item functioning (English learners group).**

*Note.* "Original" refers to the unsimplified versions of the items.

logistic regression DIF results flagged four items for statistically significant DIF, indicating that ELs scored very differently on the original versus simplified forms of these four test items. Of the four test items, three had nonnegligible effect sizes, all of which classified them as large DIF (i.e., $R^2$ values greater than .07). For all three items, the simplified item was easier than the original item. Figure 6 presents the logistic curves for each of the three 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

*Table 10*
**Items Flagged for Nonnegligible DIF Indicating Differences Between Scores on Original and Modified Item for the Listed Group**

| Group | Item | Simplification Type | DIF Effect Size | Direction |
|---|---|---|---|---|
| ELs | 200017001 | Visual | Large | Simplified easier |
| ELs | 300032001 | Interview-based | Large | Simplified easier |
| ELs | 200008001 | Visual | Large | Simplified easier |
| Non-EL–Low | 400015001 | Visual | Large | Simplified easier |
| Non-EL–High | 400015001 | Visual | Moderate | Simplified easier |

*Note.* DIF = differential item functioning; EL = English learner.

simplified item was higher than the curve for the original item, indicating that the item was easier in the simplified condition. The linguistic simplification demonstrated the hypothesized effect for these three items.

DIF analyses were also conducted for the two non-EL subgroups (Non-EL-Low and Non-EL-High), and neither of these analyses flagged for nonnegligible DIF any of the three items that were flagged for ELs, supporting the argument that the simplifications did not alter the tested construct. However, for the Non-EL-Low and -High groups, one item (Item 400015001) was flagged for nonnegligible DIF indicating that the simplified item was easier, and it had a large effect size for the Non-EL-Low group and a moderate effect size for the Non-EL-High group. Although EL scores did not improve significantly on the simplified version of this item, these findings warrant further investigation, as they suggest that the test item simplification altered the tested construct for this item. The overall DIF findings confirm that the score improvements for the three items flagged for DIF for ELs, all of which were from Item Set 1, were specific to ELs. Table 10 presents a summary of the DIF results.

## Discussion

In this study, we administered original and simplified science test items to EL and Non-EL Grade 5 students using an experimental design. We used three sets of statistical analyses and several comparisons to evaluate overall linguistic simplification effects, effects of specific linguistic simplification types, and effects associated with specific items. One finding is clear: The effects of linguistic simplifications depend on the specific linguistic simplifications that were performed on specific test items. Although there were no statistically significant differences at the item set score level, we did find that for Item Set 1, the linguistic simplifications improved EL student performance more than the performance of Non-EL-Low and Non-EL-High students, supporting the hypothesis that our test item simplifications affected only the English language demands of the items, and not the science content of the items. However, for

Item Set 2, we did not improve EL student performance. To better understand how these two different sets of items may have affected students differently, we explored effects at the linguistic simplification type and item level.

At the linguistic simplification type level, we observed statistically significant differences in performance of ELs on simplified and original items for the Visual and the Interview-based simplifications. The success of the Visual simplification is consistent with the findings of Solano-Flores et al. (2014), indicating that the addition of a visual representation is uniquely powerful in clarifying the meaning of test items written in English. The success of the Interview-based simplification suggests that more substantial effects are likely to be seen when multiple features are simplified, consistent with findings of other linguistic simplification studies (see Noble et al., 2018, for a review), and that cognitive interviews are an important component of simplification design.

The analyses of the effects of individual test items on the performance of ELs and non-ELs provides further information about which specific Visual and Interview-based linguistic simplifications led to the largest improvements in the performance of ELs. DIF analyses comparing student scores on original versus simplified versions of test items (rather than comparing EL to non-EL scores) demonstrated that two of the Visual simplifications and one of the Interview-based simplifications led to significant improvements in scores for ELs. Non-ELs' scores did not improve on these items, supporting the conclusion that the tested science content was not altered for these three items, and that the improvements in scores for ELs were related to clarification of the item language.

The simplified versions of all three items flagged in the EL DIF analysis included the addition of visuals to the answer choices.[4] One of these items is shown in original and simplified form in Figures 1 and 2, respectively. These three items appeared in simplified form in Item Set 1, which may provide an explanation for the improvements in EL scores on the simplified versions of Item Set 1. In addition, these three items were the *only* items in the set of Visual and Interview-based linguistic simplifications in which visuals were added to the answer choices, as opposed to the stem of the item. We chose to illustrate the answer choices in these three items because some of the words in the answer choices were Low-Frequency Nontechnical words, a type of word found in preparatory work to be correlated with lower test scores for ELs (e.g., *ceramic, hose, wilt*; Kachchaf et al., 2016). Our findings suggest that the visuals added to answer choices may have clarified the meanings of these Low-Frequency Nontechnical words for students. Without the visuals, ELs performed significantly worse on these items. The specific challenge of Low-Frequency Nontechnical vocabulary in *answer choices* has not been highlighted in previous research on ELs and assessments in science and mathematics (Noble et al., 2018). This is likely due in part to the predominance of research on ELs and mathematics

assessments, in which the challenge of unfamiliar words in answer choices may be mitigated by the prevalence of numbers, symbols, and visuals in answer choices on math tests.

### Limitations

There are some limitations of the present study that should be noted for consideration in future research. The goal of our study was to construct targeted linguistic simplifications to isolate the effects of small numbers of linguistic features on ELs' test performance. We wished to provide the field of assessment research and development with evidence regarding the effects of specific linguistic features on ELs' test scores and guidance about which of the more than 60 different problematic linguistic features identified in the literature (Noble et al., 2018) have the largest negative effects on ELs' test performance. Except in the case of the Visual simplification, we found that isolating one or two linguistic features to simplify did not cause a significant change in the performance of ELs compared with non-ELs. Thus, the use of targeted linguistic simplifications to isolate the effects of specific linguistic features on ELs' performance may require larger sample sizes, larger numbers of test items for each simplification type, or both.

### Implications

Our study highlights the benefits of using multiple sources of data and multiple forms of data analysis to understand the complex issues of how to write science test items that allow valid inferences to be made about ELs' test scores. The development of simplified test items was informed and guided by a cognitive interview study with EL students. The analyses of test results at the item set, linguistic simplification type, and individual item levels allowed for investigation of the effects of linguistic simplification types and individual item simplifications that were masked at the item set level. This type of multilevel analysis may be helpful for future research on linguistic simplification and other forms of accommodation.

The finding of this study regarding the positive impact of adding visuals to answer choices is promising and has implications for future assessment research. Extensive research has explored the use of visuals to illustrate the stem of test items and the characteristics of visuals that are most effective for ELs (Solano-Flores et al., 2014). However, there has not been a similarly detailed analysis of the use of visuals to illustrate answer choices. Given the comparative success of illustrating answer choices in this study, it is important for the field of assessment research to further investigate how answer choice illustration may be used as a tool to improve science assessments for ELs.

In the debate over test item accommodations, this study highlights the importance of a careful, comprehensive, and research-based linguistic simplification process. The effectiveness of specific forms of simplification

clearly depends on both the features simplified and the nature of the test item to which the simplification is applied. For example, while we have demonstrated that the addition of visuals to answer choices on science test items can improve performance for ELs, mathematics test items with numerical answer choices may not need visuals added to answer choices, but may benefit from the addition of other forms of visual representation. Due to the limitations of our study, we do not have the evidence to recommend specific forms of linguistic simplification beyond the addition of visual representations. However, our findings suggest that linguistic simplifications are more successful when they include multiple test item features and are informed by cognitive interviews with the students for whom the simplifications are intended. In addition, there is an existing body of research on linguistic simplification that has shown that comprehensive simplification processes involving features across the word, sentence, and item levels can improve EL performance on test items (Noble et al., 2018). The implications of this study for test accommodations apply equally to test design. Selecting and refining test items according to the principles of successful linguistic simplification studies and testing these items in cognitive interviews and field tests with significant numbers of ELs would reduce the need for accommodations and improve the validity of test score interpretations.

## Notes

This research was supported by the Institute of Education Sciences, U.S. Department of Education through Grant No. R305A110122 and the Education Research Collaborative at TERC. The opinions expressed herein are those of the authors and do not reflect the opinions of the funding agency. The authors thank Catherine Bowler and Carrie Conaway of the Massachusetts Department of Elementary and Secondary Education for their partnership with this research effort. The authors also thank all of the students, teachers, and district and school administrators who generously chose to participate in this research. This article is dedicated to them.

[1]Using Cohen's (1988) tables and formulae, power analyses indicated power above .90 for the statistical comparisons made in this study, even when assuming an effect size as low as .20, and a significance level of .01. Therefore, the sample sizes used in this study were considered sufficient for evaluating our hypotheses.

[2]In three out of six visual simplifications, we compared test items that included a visual illustration in the stem when the item was originally administered with test items for which the illustration was removed, to investigate the effect of the illustration on the performance of ELs. In these three cases, the data from the items with the visuals maintained was grouped with the data from the items to which a visual was added.

[3]An IRT approach was also used, and yielded similar results, so is not reported here.

[4]The linguistic simplification of Item 200017001 also included the addition of visuals to the stem of the item, and the Interview-based linguistic simplification of the Item 300032001 included removing the Low-Frequency Nontechnical word "tulip" and removing the Forced Comparison feature by removing the word "first."

## References


Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment*, *8*(3), 231–257. https://doi.org/10.1207/s15326977ea0803_02

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Lawrence Erlbaum. https://doi:/10.4324/9780203874776.ch17

Abedi, J. (2008). Utilizing accommodations in assessment. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (pp. 341–347). Springer. https://doi:/10.1007/978-0-387-30424-3_185

Abedi, J., Courtney, M., & Leon, S. (2003, September). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Report No. 608). National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/R608.pdf

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 No. 666). National Center for Research on Evaluation, Standards, and Student Testing. https://doi.org/10.1037/e645072011-001

Abedi, J., & Ewers, N. (2013, February). *Smarter Balanced Assessment Consortium: Accommodations for English Language Learners and Students with Disabilities: A Research-Based Decision Algorithm*. Smarter Balanced Assessment Consortium. https://portal.smarterbalanced.org/library/en/accommodations-for-english-language-learners-and-students-with-disabilities-a-research-based-decision-algorithm.pdf

Abedi, J., & Linquanti, R. (2012, January 13–14). *Issues and opportunities in improving the quality of large scale assessment systems for ELLs* [Paper presentation]. Stanford University Understanding Language conference, Palo Alto, CA, United States. 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. https://doi.org/10.1207/s15324818ame1403_2

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

Avenia-Tapper, B., & Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English Language Learners' science knowledge. *Educational Assessment*, *20*(2), 95–111. https://doi.org/10.1080/10627197.2015.1028622

Butler, F. A., Bailey, A. L., Stevens, R., Huang, B., & Lord, C. (2004). *Academic English in fifth-grade mathematics, science, and social studies textbooks* (CSE Report No. 642). University of California, National Center for Research on Evaluation, Standards, and Student Testing. http://cresst.org/publications/cresst-publication-3013/

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (*Vol. 448*). National Center for Research on Evaluation, Standards, and Student Testing. https://doi.org/10.1037/e651402011-001

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). De Gruyter. https://doi.org/10.21832/9781783090105-011

Center for Applied Linguistics. (2015). *Annual technical report for ACCESS for ELLs® English Language Proficiency Test Series 302, 2013-2014 Administration* (WIDA Consortium Annual Technical Report No. 10). https://www.ride.ri.gov/Portals/0/Uploads/Documents/Instruction-and-Assessment-World-ClassStandards/Assessment/ACCESS/ACCESS_Technical_Report_2013%E2%80%932014.pdf

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. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum. https://doi.org/10.4324/9780203771587

Every Student Succeeds Act, 20 U.S.C. § 1177 (2015). https://www.ed.gov/essa

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data. *School Psychology Review*, *29*(1), 65–85. https://psycnet.apa.org/record/2000-03323-004

Gewertz, C. (2015, December 18). ESSA's flexibility on assessment elicits qualms from testing experts. *Education Week*, *35*(15), 16–17. http://www.edweek.org/ew/articles/2015/12/21/essas-flexibility-on-assessment-elicits-qualms-from.html

Hakuta, K., & Beatty, A. (2000). *Testing English-language learners in U.S. schools: Report and workshop summary*. National Academies Press. https://doi.org/10.17226/9998

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

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). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pdf/studies/2011459.pdf

Hopewell, S., & Escamilla, K. (2014). Struggling reader or emerging biliterate student? Reevaluating the criteria for labeling emerging bilingual students as low achieving. *Journal of Literacy Research*, *46*(1), 68–89. https://doi.org/10.1177/1086296x13504869

Huddleston, A. P. (2014). Achievement at whose expense? A literature review of test-based grade retention policies in US schools. *Education Policy Analysis Archives*, *22*, Article No. 18. https://doi.org/10.14507/epaa.v22n18.2014

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. https://doi.org/10.1207/s15324818ame1404_2

Kachchaf, R., Noble, T., Rosebery, A., Wang, Y., Warren, B., & O'Connor, M. C. (2014, April 3–7). *The impact of discourse features of science test items on ELL performance* [Paper presentation]. Annual meeting of the American Educational Research Association, Philadelphia, PA, United States. https://external-wiki.terc.edu/display/CKC/Publications+by+Date

Kachchaf, R., Noble, T., Rosebery, A., Warren, B., O'Connor, C., & Wang, Y. C. (2016). A closer look at linguistic complexity: Pinpointing individual linguistic features of science multiple-choice items associated with English language learner

performance. *Bilingual Research Journal*, *39*(2), 152–166. https://doi.org/10.1080/15235882.2016.1169455

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*(3), 1168–1201. https://doi.org/10.3102/0034654309332490

Kieffer, M. J., Rivera, M., & Francis, D. J. (2012). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments*. RMC Research Corporation, Center on Instruction. https://www2.ed.gov/about/inits/ed/lep-partnership/assessments.pdf

Kober, N., Chudowsky, N., & Chudowsky, V. (2010). *State test score trends through 2008-09, Part 2: Slow and uneven progress in narrowing gaps*. Center on Education Policy. https://www.cep-dc.org/cfcontent_file.cfm?Attachment=Kober%5FFullReport%5F2008%2D09%5FPart2%5FGaps%2Epdf

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. https://doi.org/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. https://doi.org/10.1080/08957347.2012.714690

Linquanti, R., & Cook, H. G. (2013, February 1). *Toward a ''common definition of English learner'': Guidance for states and state assessment consortia in defining and addressing policy and technical issues and options*. Council of Chief State School Officers. https://files.eric.ed.gov/fulltext/ED542705.pdf

MA Department of Elementary and Secondary Education. (2014a). *Massachusetts Comprehensive Assessment System: Test Questions*. http://www.doe.mass.edu/mcas/testitems.html

MA Department of Elementary and Secondary Education. (2014b). *2014 MCAS and MCAS-Alt Technical Report*. http://www.doe.mass.edu/mcas/tech/?section=techreports

MA DOE. (2004). *Massachusetts Comprehensive Assessment System: Test questions*. http://www.doe.mass.edu/mcas/testitems.html

MA DOE. (2005). *Massachusetts Comprehensive Assessment System: Test questions*. http://www.doe.mass.edu/mcas/testitems.html

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*(3), 160–179. https://doi.org/10.1080/10627190903422906

McIntosh, S. (2011, December 8). *State high school tests: Changes in state policies and the impact of the college and career readiness movement*. Center on Education Policy. http://www.cep-dc.org/displayDocument.cfm?DocumentID=385

National Center for Education Statistics. (2015, May). *Table 204.27: English language learner (ELL) students enrolled in public elementary and secondary schools, by grade and home language: Selected years, 2008-09 through 2013-14*. https://nces.ed.gov/programs/digest/d15/tables/dt15_204.27.asp

National Center for Education Statistics. (2019, December). *Table 203.10: Enrollment in public elementary and secondary schools, by level and grade: Selected years, fall 1980 through fall 2026*. https://nces.ed.gov/programs/digest/d16/tables/dt16_203.10.asp

National Research Council. (2014a). *Developing assessments for the next generation science standards*. National Academies Press. https://doi.org/10.17226/18409

National Research Council. (2014b). *Literacy for Science: Exploring the Intersection of the next generation science standards and the common core for ELA standards: A workshop summary*. National Academies Press. https://doi.org/10.17226/18803

NGSS Lead States. (2013). *Next generation science standards: For states, by states* (*Vol. 2*). National Academies Press. https://doi.org/10.17226/18290

No Child Left Behind Act of 2001, 20 U.S.C. (2002). https://www2.ed.gov/nclb/overview/intro/guide/index.html

Noble, T., Kachchaf, R. R., & Rosebery, A. S. (2018). Perspectives from research on the linguistic features of mathematics and science test items and the performance of English learners. In D. L. Baker, D. L. Basaraba, & C. Richards-Tutor (Eds.), *Second language acquisition: Methods, perspectives and challenges* (pp. 209–236). Nova Science. https://novapublishers.com/shop/second-language-acquisition-methods-perspectives-and-challenges

Noble, T., Kachchaf, R. R., Rosebery, A., Warren, B., O'Connor, M. C., & Wang, Y. C. (2014, March 30–April 2). *Do linguistic features of science test items prevent English Language Learners from demonstrating their knowledge?* [Paper presentation]. Annual meeting of the National Association of Research on Science Teaching, Pittsburgh, PA, United States. https://external-wiki.terc.edu/display/CKC/Publications+by+Date

Noble, T., Sireci, S. G., Wells, C. S., Kachchaf, R. R., Rosebery, A., & Wang, Y. C. (2016, April 8–12). *Targeted linguistic modifications of science test items for English learners* [Paper presentation]. Annual meeting of the American Educational Research Association, Washington, DC, United States. https://external-wiki.terc.edu/display/CKC/Publications+by+Date

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. https://doi.org/10.1002/tea.21026

Ojeda, R. L. V. (2016). *Supporting emerging multilingual newcomer students and their teachers in California public high schools* (Publication No. 181) [Master's thesis, University of San Francisco]. University of San Francisco Scholarship Repository: http://repository.usfca.edu/thes/181/

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press. https://doi.org/10.17226/10019

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. https://doi.org/10.1111/j.1745-3992.2011.00207.x

Pennock-Roman, M., & Rivera, C. (2012). *Summary of literature on empirical studies of the validity and effectiveness of test accommodations for ELLs: 2005-2012*. Smarter Balanced Assessment Consortium. Retrieved April 7, 2017, from https://portal.smarterbalanced.org/library/en/summary-of-literature-on-empirical-studies-of-the-validity-and-effectiveness-of-test-accommodations-for-ells-2005-2012.pdf

Pennsylvania Clearinghouse for Education. (2013, March). *Issue brief: School closings policy*. Research for Action. https://www.researchforaction.org/publications/school-closings-policy/

Proctor, C. P., & Silverman, R. D. (2011) Confounds in assessing biliteracy and English language proficiency. *Educational Researcher*, *20*(2), 62–64. https://doi.org/10.3102/0013189X11403138

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). Institute of Education Sciences National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=REL20094079

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. https://doi.org/10.1016/j.econedurev.2012.12.003

Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education*, *39*(1), 215–252. https://doi.org/10.3102/0091732x14557003

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature* (Center for Educational Assessment Research Report No. 485). University of Massachusetts, Amherst, School of Education. https://nceo.umn.edu/docs/OnlinePubs/TestAccommLitReview.pdf

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. https://doi.org/10.3102/0013189x08319569

Solano-Flores, G., & Gustafson, M. (2012). Academic assessment of English language learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education* (chap 6). Routledge. https://doi.org/10.4324/9780203154519-6

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. https://doi.org/10.1080/10627197.2014.964116

WIDA. (2014). *ACCESS for ELLs summative assessment*. Retrieved October 2, 2015, from https://www.wida.us/assessment/access/

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

Manuscript received August 30, 2018
Final revision received November 21, 2019
Accepted January 6, 2020