Investigating Two Linguistic Factors Associated with Differential Performance of English Language Learners

Cari F. Herrmann-Abell, George E. DeBoer, Elise Trumbull, Ursula M. Sexton, Sara Glassman, Chun-Wei (Kevin) Huang, & Sharon Nelson-Barber

Paper presented at NARST Annual Conference

March 31- April 3, 2019

Baltimore, MD

Abstract

This study is part of a larger effort to investigate linguistic features of assessment items that might explain differential performance of English learners on science assessments. An analysis of a large set of multiple-choice items suggested two features that held promise in reducing the performance gap: the presence of contrast words and the reduction of 1+ clauses. A subset of 48 items were modified and used in a study with 638 students to investigate the effect of these features. For half of those items, we reduced the number of 1+ clauses; for the other half, we removed contrast words. We analyzed the results from two subsets of students: 358 9th grade native-English and native-Spanish speaking students from across the U.S. and 138 8th grade students from one teacher at a school with 50% Spanish speaking students who agreed to participate in follow-up focus group discussions about the items. We compared student performance on the modified and original items and found little to no support for the selected linguistic features as affecting English learners' scores differentially, suggesting that performance was more strongly dependent on content knowledge than on the linguistic features of the items. The follow-up discussions also supported the idea that students' performance was strongly related to their content knowledge level.

Introduction

Students whose primary language is not English do not perform as well as English-speaking students on tests that assess their understanding of the school curriculum (Abedi, Leon, & Mirocha, 2003; Martiniello, 2008; Sato, Rabinowitz, Gallagher, & Huang, 2010). Although some of this disparity may reflect real differences in students' knowledge, it is also possible that the tests are not fairly evaluating what English learners (ELs) know. Some studies involving interviews with EL students have shown that ELs who can demonstrate an understanding of the science content in an interview setting select incorrect answers on items targeting the same content during testing (Noble, Rosebery, Suarez, Warren, & O'Connor, 2014; Noble, Suarez, Rosebery, O'Connor, Warren, & Hudicourt-Barnes, 2012).

Understanding the factors that affect EL students' performance is becoming even more important with the advent of *Common Core State Standards* and *Next Generation Science Standards* (NGSS). These standards place greater emphasis on the use of language and reasoning skills so that students can engage effectively in science discourse and sense making (Lee, Quinn, & Valdés, 2013), likely increasing the language demands of assessments. Given that ELs are such a rapidly growing segment of the U.S. student population, the underperformance of Els is a

particularly significant issue to address.

This validation study is the capstone of a larger project to analyze a set of 913 multiple-choice assessment items to identify linguistic features and cognitive complexity features that explain differential student performance (DeBoer, Herrmann-Abell, Trumbull, Nelson-Barber, Huang, Sexton, & Glassman, submitted). The goal was to propose strategies for ameliorating those features and use those strategies to revise a set of items. The revised items were then to be used in an empirical study exploring whether the changes made in the items narrow the performance gap between EL and non-EL students. The results of this work will help answer fundamental questions about the nature of and the extent to which linguistic factors can affect EL students' performance on science assessments.

Research on Language Modification of Assessments

Previous studies that examined the effect of language modifications of assessment items have shown mixed results. While Abedi and Lord (2001) found that simplifying the language used in an item helped ELs in general, some studies have shown that modification helps subsets of students such as ELs at certain grade levels (Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005), ELs at intermediate levels of English proficiency (Pennock-Roman & Rivera, 2011), or only the least English-proficient ELs (Kiplinger, Haug, & Abedi, 2000). Other studies have shown that the modifications help non-ELs more than ELs (Young, King, Hauck, et al., 2014) or help both ELs and non-ELs on some items and neither group on others (Young, et al., 2014). Still other studies have found no significant effect of simplifying language on ELs' performance (Li & Suen, 2012; Rivera & Stansfield, 2004).

These varied results stem from at least two factors that complicate research in this area: (1) It is difficult to completely separate students' general facility in a language from their content knowledge. In any assessment, there is some amount of complex language that is needed to convey the academic content and cannot be removed without affecting the construct validity of the assessment (Abedi & Herman, 2010; Kieffer, Lesaux, Rivera, & Francis, 2009; Trumbull & Solano-Flores, 2011). Therefore, language facility is always going to be a factor in how well students do on test items. (2) The sources of linguistic complexity are vast, and they may interact with other factors besides content knowledge, including visual representations used in the items and characteristics of the learners themselves (Solano-Flores, 2010; 2014). Hence, research is unlikely to consistently identify the same sources of linguistic complexity contributing to differential performance between ELs and native-English speakers across different populations and assessments.

Assessment Items Used in this Study

The items used in the validation study were selected from a set of 913 previously developed multiple-choice assessment items closely aligned to targeted learning goals in 17 middle and early high school science topics (American Association for the Advancement of Science [AAAS], n.d.). The topics targeted by the items were chosen based on their frequency of coverage in middle and early high school science and their centrality for providing students with a coherent understanding of science. In addition to testing ideas from the life, physical, and earth sciences, there are also items that test ideas about the control of variables and the use of models in science. Furthermore, the item distractors probe for common student misconceptions (Sadler, 1998). The items focus on a range of cognitive skills, which includes being able to: (1) evaluate the truth of scientifically correct statements, (2) analyze a situation to determine what else must be true about it, given certain critical information, (3) explain natural phenomena in terms of scientific principles, and (4) use scientific principles to make predictions about natural

phenomena.

Item development focused on maximizing construct validity by paying close attention to the alignment of the items to the knowledge being targeted and the comprehensibility of each item to students (DeBoer, Herrmann-Abell, Gogos, Michiels, Regan, & Wilson, 2008). Specific attention was paid to reducing construct-irrelevant features that might disadvantage EL students as suggested by Kopriva (2000). After initial item development, pilot testing was used to obtain written feedback from students on their understanding of why each answer choice was correct or incorrect. About 100 students answered and made comments on about six items each. The results of pilot testing were then used to revise the items. Finally, each of the revised items was then field tested with approximately 2000 middle and high school students from rural, urban, and suburban school districts across the country. During testing, students self-identified whether or not English was their primary language. Data from the field tests showed an average performance gap between non-EL and EL students of approximately seven percentage points.

Linguistic Features of the Items

Linguistic complexity features can be categorized into four groups: sentence complexity, vocabulary complexity, multiword expressions (idiomatic language), and discourse complexity. In this study, we wanted to focus on a small number of features that we could reliably count in the full set of items. We used a natural language processing system developed by ETS called Language Muse (Burstein, Shore, Sabatini, Moulder, & Holtzman, 2012) to electronically count the frequency of 35 linguistic features in the set of 913 items from the AAAS item bank. We determined the reliability of these counts by hand-counting the features for 150 items and comparing the hand-counts to those produced by Language Muse. The results of that reliability study identified two features that were reliably counted by Language Muse and showed promise in narrowing the performance gap based on previous research: 1+ clauses and contrast words. A sentence was flagged as containing 1+ clauses if it included an independent clause plus at least one dependent clause.[1] An item was flagged as containing contrast words if it included words or terms that indicate a contrast relation between text segments, such as "however" or "but."

Research on the effect of using 1+ clauses and contrast words. Regarding 1+ clauses, past research has shown that complex sentences that include dependent clauses present difficulties for non-native English speakers (Botel & Granowsky, 1974; Hunt, 1965, 1977; Wang, 1970; Abedi, 2007; Abedi, Lord, & Plummer 1997). Spanos and colleagues (1988) suggest that the use of separate sentences may be easier for some students to understand. On the other hand, previous research has shown that cohesive devices, such as the use of contrast words, may help EL students interpret sentences (Abedi, Leon, Wolf, & Farnsworth, 2008).

Research questions

Previous findings suggest that using items that have fewer 1+ clauses and more contrast words may reduce the performance gap between non-ELs and ELs. However, none of the past studies rigorously investigated the effect of modifying these linguistic features on student performance on science tests. This study aims to fill this research gap by determining the effect of systematically modifying these two non-construct-related linguistic features of science assessment items on the gap in performance between non-EL and EL students. More specifically, we sought to answer the following questions:

[1] In the following sentence, the first four words are a dependent clause, and the remainder of the sentence is an independent clause: "When warm air cools, it forms clouds."

1. Does the reduction of 1+ clauses in science assessment items increase the performance of either native Spanish speakers or native English speakers?
2. Does the removal of contrast word in science assessment items decrease the performance of either native Spanish speakers or native English speakers?

Methodology

Modification of Items

We selected a subset of 48 items to modify that were chosen from across the full range of topics. Items were chosen that have a difficulty level (percent correct) on the original version of the item between 40% and 60% in order to minimize guessing and ceiling effects. The items were also selected so that the performance difference between EL and non-EL is about seven percentage points, the average for the full set.

Contrast words. We selected 24 items that could be modified by removing contrast words. (Even though contrast words seem to help students, it was easier to remove contrast words than to add them.) Great care was taken to ensure that the meaning of the statements was not changed when contrast words were removed so that the validity of the item as a measure of student understanding of the science ideas was not changed. Revisions for each item were discussed and agreed upon by three researchers with content and linguistic expertise. The original versions of the items selected had a minimum of two contrast words, a maximum of four contrast words, and an average of 2.67 contrast words. The modified versions had no contrast words. The most common modification was replacing the word "but" with "and" as illustrated in Figure 1.

Original version with contrast words	Modified version without contrast words
Two white powders were mixed together. A chemical reaction occurred, and a yellow powder was formed. What is the relationship between the yellow powder and the white powders?	Two white powders were mixed together. A chemical reaction occurred, and a yellow powder was formed. What is the relationship between the yellow powder and the white powders?
A. The yellow powder is made up of the same kinds of atoms as the white powders, *but* the atoms are combined into different molecules.	A. The yellow powder is made up of the same kinds of atoms as the white powders, *and* the atoms are combined into different molecules.
B. The yellow powder is made up of the same kinds of molecules as the white powders, *but* the molecules are a different color.	B. The yellow powder is made up of the same kinds of molecules as the white powders, *and* the molecules are a different color.
C. The yellow powder was released from inside the atoms of the white powders.	C. The yellow powder was released from inside the atoms of the white powders.
D. There is no relationship between the yellow powder and white powders.	D. There is no relationship between the yellow powder and white powders.

Figure 1: An example of an item modification where all the contrast words were removed. Replacing "but" with "and" is one of the simplest and cleanest way to turn a contrast word sentence into a non-contrast word sentence.

1+ clauses. Another 24 items were modified by reducing the number of sentences with 1+ clauses. Items that had a minimum of four 1+ clauses were considered for selection. Because reducing the number of clauses tends to increase the sentence count by creating simpler sentences, we kept track of the number of dependent clauses per sentence for each version of the items. The number of dependent clauses per sentence for the original versions ranged from 0.50 to 8.00 with an average of 1.92. The number of dependent clauses per sentence for the modified versions ranged from zero to 1.33 with an average of 0.57. Figure 2 shows an example item in its original and modified versions.

Original version	Modified version with a reduced number of dependent clauses per sentence
Cool air is blowing from Location 1 toward Location 2, where the air is warmer. When the air that was at Location 1 reaches Location 2, clouds begin to appear. If there is no increase in the amount of water vapor in the air as it moves, what could explain why the clouds appear at Location 2?	Cool air is blowing from Location 1 toward Location 2. The air is warmer at Location 2. The air from Location 1 reaches Location 2, and clouds appear. If there is no increase in the amount of water vapor in the air as it moves, why do clouds appear at Location 2?
A. The cool air pushed the warm air upward, which caused the warm air to become cooler, condense, and form clouds.	A. The cool air pushed the warm air upward. This caused the warm air to become cooler, condense, and form clouds.
B. The cool air pushed the warm air upward, which caused the warm air to become warmer, condense, and form clouds.	B. The cool air pushed the warm air upward. This caused the warm air to become warmer, condense, and form clouds.
C. The warm air pushed the cool air upward, which caused the cool air to become warmer, condense, and form clouds.	C. The warm air pushed the cool air upward. This caused the cool air to become warmer, condense, and form clouds.
D. The clouds must have moved from somewhere else because air without clouds cannot turn into air with clouds.	D. The clouds must have moved from somewhere else. Air without clouds cannot turn into air with clouds.

Figure 2: An example of an item modification where the number of dependent clauses was reduced.

Design of Instruments for the Validation Study

For the validation study, we developed four forms of an instrument to compare student performance on 48 revised items against student performance on the same unrevised items. As described above, we reduced the number of 1+ clauses for half the items and removed contrast words from the other half. Linking items on each form were used so that we could give different test forms to different students and then combine the data. Each participant in the study completed one of the four forms. Each form was made up of 30 items in all: six items with a reduced number of 1+ clauses, six items with the original number of 1+ clauses, six items with no contrast words, six items with the original number of contrast words, and six linking items. Items were clustered by topic (e.g. physical science, life science, and earth science). All linking

items were unrevised items selected on the basis of their psychometric features (appropriate level of difficulty, reliability, lack of bias).

Participants

Teachers from around the country who had a significant number of Spanish-speaking students in their classes were recruited by email to participate in the study. Students were asked to report their gender, grade in school, whether the first language they learned was English or Spanish, and their race and ethnicity. Native-Spanish-speaking students who participated in the study were considered, on average, to be moderately proficient in English. Although English was not their native language, their teachers felt they were proficient enough in English to participate in a study like this.

Students in each classroom were randomly assigned one of four test forms so that approximately one quarter of the students received each form of the test. The test was administered using our online testing utility. Students took the test during their regularly-scheduled science class and were given one class period to respond to the items.

A total of 638 students in 7th through 12th grade participated in the study in the winter and spring of 2018. The results presented here focus on two subsets of the data. The first includes data from 358 9th grade students from across the U.S. who indicated that either English or Spanish was their first language. The second includes data from 138 8th grade students from one teacher in a school in the western U.S. who indicated that either English or Spanish was their first language. Below we describe the subsets and the reasons for focusing on them.

Ninth-grade sample. The 9th grade sample is made up of students from 10 teachers in eight states across the U.S and in Puerto Rico. The student demographics are presented in Table 1. We focused on 9th grade students because it is reasonable to assume that they have had exposure to the full set of topics targeted by the items at the time of testing. Additionally, there were approximately equal numbers of native-English and native-Spanish students in the 9th grade sample, which was not the case for the other grades.

Eighth-grade sample. The 8th grade sample is made up of students from one teacher whose classes were approximately 50% native-Spanish speakers and 50% native-English speakers, and whose school allowed a member of the research team to conduct focus group discussions with a subset of students after testing. Table 1 summarizes the demographics of the 8th grade sample. About 85% of the native-Spanish speakers were proficient in English and 15% were less than proficient. The community in which the school is located is agrarian with 45% immigrant Hispanic/Latino population. The school enrollment is just under 840 students, out of which 472 are Hispanic/Latino (National Center for Education Statistics, n.d.). Twenty-three of these students had parental permission to participate in the follow-up interviews. These students took the tests at the end of May to ensure that they would have had exposure to the full set of topics at the time of testing.

Table 1. *Student Demographics*

	9th grade sample		8th grade sample	
	English (N = 177)	Spanish (N=181)	English (N = 77)	Spanish (N=61)
Gender				
Female	36%	50%	39%	44%
Male	59%	46%	61%	56%
Race/Ethnicity				
White	46%	1%	48%	0%
Black	5%	0%	4%	0%
Hispanic	28%	93%	25%	80%
Asian	2%	0%	0%	2%
Pacific Islander	3%	0%	0%	0%
American Indian	2%	1%	0%	3%
2 or more/Other	8%	2%	22%	5%

Student Focus Groups

Thirteen native-Spanish-speaking students and 10 native-English-speaking students from the 8th grade sample participated in follow-up focus group discussions with a native-Spanish-speaking member of the research team. In order to have productive discussions, six focus groups of two to five students were selected. The focus groups took place within eight days of testing and each lasted 35 to 45 minutes. All students preferred to speak in English during the protocol but, to lower the threshold of anxiety for native-Spanish speakers, informal greetings and an introductory portion of the protocol were given both in English and in Spanish. Each focus group discussed the original and modified versions of three to four items, with the original item presented first.

Results

Item Difficulty

For each item, we calculated the percentages of correct responses from the native-English speakers and native-Spanish speakers in each sample. Table 2 summarizes these results by item type (i.e. linking items, contrast words items, and 1+clauses items). We expected to see a larger percent correct for the items that had a reduced number of clauses compared to the original versions and a smaller percent correct for the items with the contrast words removed compared to the original versions. Instead, we found that in no case was the difference in percent correct on the original and modified items statistically significant for any of the comparisons. Sometimes the results were as expected, and sometimes they were contrary to expectations, but none of these apparent differences was statistically significant. For example, English speakers in the 8th grade sample showed a small difference in the expected direction for both types of items (a three percentage point decrease for contrast words items and a five percentage point increase for 1+ clauses items) but the differences in the percentage of correct responses were not statistically significant. Contrary to expectations, on average, the 8th grade Spanish speakers performed better on the modified contrast words items (a five percentage point increase) and the same on the modified 1+ clauses items. But, again, the difference was not statistically significant.

In the 9th grade sample, we saw the expected trend for the Spanish speakers on the contrast words items although there was only a one percentage point difference. The 9th grade English speakers performed equally well, on average, on the original and modified contrast words items.

Overall, both groups of 9th graders performed worse on the modified versions of the 1+ clauses items (a four percentage point decrease for English speakers and a one percentage point decrease for the Spanish speakers), contrary to what was expected, but those differences were not statistically significant. Chi-square statistics were calculated for each comparison and none was found to be statistically significant on the .05 level.

Table 2. *Average percentage of correct responses by item type*

	8th grade sample		9th grade sample	
	English	Spanish	English	Spanish
Linking items	41%	38%	43%	34%
Contrast words items				
Original items	40%	39%	43%	37%
Modified items	37%*	44%	43%	36%*
1+ clauses items				
Original items	33%	34%	41%	35%
Modified items	38%*	34%	37%	34%
Total	37%	39%	42%	35%

* indicates the items exhibited the predicted trend; however, when Chi-square statistics were calculated for each comparison, none was found to be statistically significant on the .05 level.

Item-level analysis of contrast words items. We looked at the individual items that were modified by removing the contrast words to try to identify patterns in the item pairs that performed as predicted (modified version had a lower percent correct). In the 8th grade sample, for the native-Spanish speakers, only eight out of the 24 modified contrast word items had a lower percentage of correct responses. For the native-English speakers, 11 out of the 24 modified contrast word items had a lower percentage of correct responses. Five of these items showed the predicted trend for both English and Spanish speakers.

In the 9th grade sample, for the native-Spanish speakers, 11 out of the 24 modified contrast word items had a lower percent correct, five of which overlap with the 8th grade sample. For the native-English speakers, 10 out of the 24 modified contrast word items showed the predicted trend, six of which overlap with the 8th grade sample. Four items showed the expected trend for 9th grade English and Spanish speakers. An analysis of the item pairs that did and did not perform as predicted showed no patterns. Sometimes the substitution of "and" for "but" decreased performance, and other times the substitution made the item easier. Nothing in the structure or content of the items suggested why that might be.

Item-level analysis of 1+ clauses items. We repeated the analysis with the items that were modified to reduce the number of 1+ clauses. In the 8th grade sample, for the native-Spanish speakers, 11 out of the 24 modified 1+ clauses items had the predicted higher percentage of correct responses. For the native-English speakers, eight out of the 24 modified 1+ clauses items had a higher percentage of correct responses. Four items showed the expected trend for both English and Spanish speakers.

In the 9th grade sample, for the native-Spanish speakers, 11 out of the 24 modified 1+ clauses items showed the predicted trend, six of which overlapped with the 8th grade sample. For the native-English speakers, 11 out of the 24 modified 1+ clauses items had a higher percentage of correct responses, five of which overlapped with the 8th grade sample. Eight items showed the

expected trend for both English and Spanish speakers. Again, no pattern was found to explain the performance of the 1+ clause item pairs.

To see if the number of dependent clauses *per sentence* is related to percent correct, we calculated correlation coefficients between those two variables across all items, both original and modified. The correlation coefficients between the number of dependent clauses per sentence and student percent correct for native-English speakers were 0.17 for the 8th graders and 0.14 for the 9th graders. The correlation coefficients for the native-Spanish speakers were even smaller (0.06 for the 8th graders and 0.07 for the 9th graders). None of these correlations was statistically significant, indicating no relationship between item difficulty and the number of dependent clauses in an item.

Student Focus Groups

Overall, the students who participated in follow-up focus groups found both the original and modified items comprehensible, and it was clear they were using their science content knowledge to answer the questions. For the 1+ clause items, several students commented that the modified versions having the simpler sentences were easier to read, but they said it rarely affected what answer choice they selected. Past research by Young et al. (2014) also found that even when ELs reported that modified items were easier to read, their scores did not change.

The results for the contrast items were mixed. For four out of the 12 contrast-word items used during the focus groups, students preferred the version without the contrast words. For two items, there was one student who preferred the original versions with the contrast words. For one item, two students preferred the original version and one preferred the modified version. For five contrast words items, none of the students thought the modifications made a difference.

Overall Percent Correct

We also looked at the overall performance of English and Spanish speakers in both grades. As reported in Table 2, the native-English speakers in the 9th grade sample outperformed the native-Spanish speakers by seven percentage points ($\chi^2 = 20.24$, $p < .001$). This performance difference is comparable to the difference found between EL and non-EL students in the full set of 913 items based on previous national field testing. In the 8th grade single-school sample, however, where native-English and native-Spanish speakers were taught in the same classroom, the native-English speakers and native-Spanish speakers performed similarly (37% correct for English speakers and 39% for Spanish speakers; $\chi^2 = 0.83$, n.s.).

Discussion

Linguistic Complexity

Although researchers have demonstrated that certain linguistic features may be responsible for the performance gap between EL and non-EL students on science assessments, we were unable to find support for the modification of two non-construct-related linguistic complexity features to reduce this gap. Our study investigated the effect of removing contrast words and reducing the number of sentences with 1+ clauses.

Removal of contrast words. Discourse transition words such as contrast words have been shown to improve ELs' performance on assessments, so we expected that removing them from the items would lead to lower performance. We did not consistently see the expected trend for the native-English speakers or for the native-Spanish speakers in either the 8th grade or 9th grade

sample. During the focus group discussions, students were split on whether or not the contrast words helped or not, and most seemed to interpret the sentences in the same way whether the sentences contained a "but" or an "and."

Reduction of 1+ clauses. Past research has suggested that the presence of complex sentences with dependent and independent clauses in assessment items disadvantage EL students. Therefore, we expected to see increased performance on the items that had been modified to reduce the number of dependent clauses per sentence. However, we did not see this pattern consistently across the items for either the native-English speakers or the native-Spanish speakers. We also found no correlation between student performance and the number of dependent clauses per sentence. Even though students' performance on the items was not affected, students preferred reading the modified versions with fewer clauses per sentence.

Opportunity to Learn

To explore the effect of opportunity to learn, we compared the overall performance of 8th and 9th grade samples. Because the native-English and native-Spanish speakers in the 8th grade sample all had the same science teacher and the classrooms had a mix of native-English and native-Spanish students, we can assume that they received the same science instruction (i.e., the same opportunity to learn the science content). The 9th grade students, on the other hand, were from many schools across the country and, therefore, we can assume received different instruction and had different opportunities to learn. Results from across the full set of original and modified items showed that there was very little difference in the performance of the 8th grade English speakers and Spanish speakers, but there was a seven percentage point difference in the performance of the 9th grade English speakers compared to the Spanish speakers. These results suggest that the performance gap may be predominantly explained by differences in opportunity to learn.

Study Limitations

We would like to acknowledge the limitations of this study. In the 9th grade sample, we had to rely on students' self-identification of whether English or Spanish was the first language they had learned, and we were unsuccessful at collecting information about students' proficiency with English. In the 8th grade sample, rather than relying on students' self-reports of their primary language, we were able to gather this information from the teacher. The teacher was also able to provide information on students' English proficiency level based on the California English Language Development Test (CELDT); most of the students were classified as Redesignated Fluent English Proficient (R-FEP). Our study would have been improved if we were able to obtain more proficiency data from the 9th grade sample and if we were able to include students with a wider range of English proficiency. This would have allowed us to explore whether students' level of proficiency was associated with differential impact of the item modifications.

Additionally, we were unable to collect specific information about the science instruction the students in our study had received. While we tried to increase the likelihood that the students had been introduced to the targeted science content by including 9th graders who had completed middle school and 8th graders s at the end of the school year, we do not know the extent to which these efforts were effective. Collecting specific information about the classroom instruction the students received would improve the validity of the study.

Finally, regarding the contrast word items, most of the modifications entailed replacing "but" with "and." We chose that substitution because it was the simplest way to change the sentence without changing its meaning in terms of the science content. "But" is also the most frequently

used contrast word in the item set. Because the two words are the same length and their difference in meaning is subtle, most students interpreted them to mean the same thing. The point of using any contrast word is to focus on the differences between two things, but apparently students could see that a comparison was being made even without the explicit use of a contrast word. Because we cannot be sure whether the presence or absence of a different contrast word such as "yet," "however," "nonetheless," or "in contrast" would have a different effect, the generalizability of our results to other contrast words may not be justified.

Conclusions

The analysis of linguistic features of a set of 913 science assessment items, combined with results from previously published research, suggests that the presence of contrast words and the reduction of clauses are the most promising features, among a larger set that we considered, for narrowing the performance gap between English learners and native-English speakers. To more rigorously test the promise of modifying these features, we removed all the contrast words from a set of items with the expectation that this would make the items more difficult for English learners. We also reduced the number of clauses in a different set of items with the expectation that this would make the items easier for English learners. Then we randomly assigned students to respond to either the original or modified versions of the items.

Overall, our analyses of the data from 8th and 9th grade students indicated that the modifications had little to no effect on the performance of native-English speakers and native-Spanish speakers. Student focus group discussions showed that, in some cases, students preferred to read sentences with fewer clauses, but when clauses were present in clearly stated complex sentences, they did not negatively affect students' answer choice selection. From these results, we conclude that performance differences on these assessment items are more the result of content knowledge differences than these specific linguistic features. This conclusion was supported by a comparative analysis of the overall percent correct of the 8th and 9th grade students. This analysis revealed that the performance gap in the 8th grade sample where students had the same instruction by the same teacher was smaller than the performance gap in the 9th grade sample, where students experienced different instruction from different teachers and where, most likely, students had different opportunities to learn science. This outcome reinforces the belief that with equal opportunity to learn, the impact of language differences among students is minimized.

While we acknowledge that the linguistic features of test items play an important role in assessment, our results suggest that the level of opportunity to learn may play a larger role. This does not mean that linguistic features are not important. Attention to linguistic features that have been shown in previous studies to penalize or benefit ELs is, we believe, still very much warranted. Clear, straightforward language in assessment items should be as important a goal for educators and test developers as accuracy of content or appropriateness of cognitive demand. The relationship between content knowledge and language skill is particularly important as new science standards call for instruction that expects students to make greater use of science practices such as providing evidence-based explanations as they make sense of phenomena and solve engineering problems (Lee, Quinn, & Valdés, 2013). These new standards and associated assessments call for high levels of language use, so differences in English proficiency may be magnified and affect EL performance in as yet undefined ways (Lee et al., 2013).

Funding

This material is based upon work supported by the National Science Foundation under Grant No. 1348622. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Abedi, J. (2007). Language factors in the assessment of English language learners: The theory and principles underlying the linguistic modification approach. In J. Abedi and E. Sato. *Linguistic modification*. (Report prepared for the U.S. Department of Education: LEP Partnership). Retrieved from https://ncela.ed.gov/files/uploads/11/abedi_sato.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). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://cresst.org/wp-content/uploads/R666.pdf

Abedi, J., & Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and concerns. *Teachers College Record*, *112*(3), 723-746. Retrieved from http://www.ncaase.com/docs/Abedi_OTL_TRC_2010.pdf

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of students' language background on content-based assessment: Analyses of extant data* (CSE Report No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://cresst.org/wp-content/uploads/R603.pdf

Abedi, J., Leon, S., Wolf, M. K., & Farnsworth, T. (2008). Detecting test items differentially impacting the performance of ELL students. In M. K. Wolf, J. L. Herman, & J. Kim, J. (Eds.). *Providing validity evidence to improve the assessment of English Language Learners* (pp. 55-80). (Report. No. 738). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://files.eric.ed.gov/fulltext/ED502627.pdf

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219-234. Retrieved from http://jwilson.coe.uga.edu/EMAT7050/articles/Abedi.Lord.pdf

Abedi, J., Lord, C., & Plummer, J. (1997). *Final Report of language background as a variable in NAEP mathematics performance* (CSE Technical Report. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from https://cresst.org/wp-content/uploads/TECH429.pdf

American Association for the Advancement of Science. (n.d.). *AAAS Science Assessment Website*. Retrieved from http://assessment.aaas.org

Botel, M., & Granowsky, A. (1972). A formula for measuring syntactic complexity: A directional effort. *Elementary English, 49*, 513-516.

Burstein, J., Shore, J., Sabatini, J., Moulder, B., & Holtzman, S. (2012, April). *The Language Muse[SM] System: Linguistically-focused instructional authoring*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, BC.

DeBoer, G. E., Herrmann-Abell, C. F., Gogos, A., Michiels, A., Regan, T., & Wilson, P. (2008). Assessment linked to science learning goals: Probing student thinking through assessment. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing student learning: Perspectives from research and practice* (pp. 231-252). Arlington, VA: NSTA Press.

DeBoer, G.E., Herrmann-Abell, C.F., Trumbull, E., Nelson-Barber, S., Huang, C.W., Glassman, S., & Sexton, U.M. (2019). Linguistic and cognitive factors affecting English learners' performance on science tests. Manuscript submitted for publication.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. (NCTE Research Report No. 3). Champaign, IL: National Council of Teachers of English. Retrieved from https://files.eric.ed.gov/fulltext/ED113735.pdf

Hunt, K. W. (1977). Early blooming and late blooming syntactic structures. In C.R Cooper & L. Odell (Eds.), *Evaluating writing*. Urbana, Ill.: National Council of Teachers of English. Retrieved from https://files.eric.ed.gov/fulltext/ED143020.pdf

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*, 1168-1201.doi.org/10.3102/0034654309332490

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). *Measuring math—not reading—on a math assessment: A language accommodations study of English language learners and other special populations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kopriva, R. J. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English Language Arts and Mathematics. *Educational Researcher*, *42*(4), 223-233.

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

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

National Center for Education Statistics. (n.d.). *Common core of data: Search for public schools* (Public school data 2015-2016, 2017-2018 school years). Retrieved from nces.ed.gov/ccd/schoolsearch/

Noble. T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education*, *27*(4), 1-19.

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.

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*, 10-28. doi.org/10.1111/j.1745-3992.2011.00207.x

Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment*, *9*(3–4), 79–105.

Sadler, P.M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*(3), 265-296.

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: Final report* (NCEE 2009-4079). Retrieved from the U.S. Department of Education website: http://files.eric.ed.gov/fulltext/ED510556.pdf

Solano-Flores, G. (2010, April-May). *Vignette illustrations as a form of testing accommodation for English language learners: A design methodology for use in large-scale science assessment*. Paper presented at the Annual Conference of the National Council of Measurement in Education, Denver, Colorado.

Solano-Flores, G. (2014). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. *Applied Measurement in Education, 27*(4), 236-247.

Spanos, G., Rhodes, N., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 221-240). Hillsdale, NJ: Erlbaum.

Trumbull, E., & Solano-Flores, G. (2011). The role of language in assessment. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.). *Cultural validity in assessment* (pp. 22-45). New York: Routledge.

Wang, M.D. (1970). The role of syntactic complexity as a determiner of comprehensibility. *Journal of Verbal Learning and Verbal Behavior 9*, 398-404.

Young, J. W., King, T. C., Hauck, M. C., Ginsburgh, M., Kotloff, L., Cabrera, J., & Cavalie, C. (2014). *Improving content assessment for English language learners: Studies of the linguistic modification of test item*s (ETS Research Report No. 14-23). Retrieved from the Educational Testing Service website: https://www.ets.org/research/policy_research_reports/publications/report/2014/jtby