ELL Corner: How to Modify Test Items for ELLs: What Research Says (Part 2 of 3)

Brooke Norval, Miami University

Abstract: *The following is the second installment in a three-part series exploring ways to modify mathematics tasks to make them more equitable for English language learners (ELLs). In this second installment, the author summarizes the existing research literature, uncovering strategies for modifying mathematics test items linguistically.*

Keywords: *English language learners, National Assessment for Educational Progress (NAEP), test items*

1 Introduction

In last issue's ELL Corner article (Norval, 2019), I introduced the reader to English language learners, the importance of thoughtfully considering ELLs in the classroom, and the effectiveness of making changes in the language of mathematics test items so that they become more equitable to ELLs. In this issue, I will explore the available literature to see what researchers have found regarding exactly how to linguistically modify mathematics test items.

2 Constructing Linguistically Modified Mathematics Test Items for ELLs

2.1 Inclusive achievement testing for linguistically and culturally diverse test takers (Fairbairn and Fox, 2009)

Fairbairn and Fox (2009) created a list of recommendations for test developers in order to more accurately measure ELL test scores. They suggest that the English language used on standardized tests is often different and more complex than the language used in everyday conversational English, causing confusion for ELLs (Fairbairn & Fox, 2009). For instance, many tests use low-frequency words such as "simultaneously," a word that an ELL student may not have encountered before, leading to a lack of understanding over an aspect of the problem that is not related to mathematics itself. One example of a low-frequency word leading to inaccurate student response was when students were prompted to write what they think is the most important modern invention. Many of the ELLs tested did not respond satisfactorily because they did not recognize the word "invention." Once the word was explained, however, all of the students wrote on-topic responses (Fairbairn & Fox, 2009). This illustrates the importance of word choice when giving tests to ELLs, as student responses cannot be accurately measured if students do not understand a key word of the test item.

Fairbairn and Fox also discussed how graphics, tables, and visuals aid ELLs during mathematics tests (Fairbairn & Fox, 2009). They state, "Train schedules and charts and tables with well-organized headings appear to support ELL test performance. ELLs draw support from highly cued text (e.g., captions or headlines in bold-face, larger font). ELLs also use picture support" (p. 15). In addition, ELLs spend more time looking at figures and graphics than non-ELLs during tests, likely because they must rely more on visuals to fill in the blanks of what they don't understand. In addition, in a pilot study, Fairbairn (2006) found that ELLs perform better on multiple-choice science test items which include visuals, suggesting the utility of visuals on test questions for ELLs.

2.2 The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities (Shaftel et al., 2006)

Shaftel et al. (2006) evaluated and measured the impact of mathematics test item language on student performance using items from the Kansas general mathematics assessments administered to students in 4th, 7th, and 10th grade. The test items fell into one of four categories, including number and computation, algebra, geometry, and data. The study included a wide variety of items from different grade levels—208 items from the 4th grade assessment, 203 from the 7th grade assessment, and 183 items from the 10th grade assessment. All items were multiple-choice word problems (Shaftel et al., 2006).

Test items were not linguistically modified; however, the linguistic features in the test items were reviewed by a team of professionals that included math teachers, math assessment specialists, and a speech pathologist specializing in ELLs (Shaftel et al., 2006). This team documented and counted various linguistic features in the test items, including "total number of words, sentences, and clauses in each item; syntactic features such as complex verbs, passive voice, and pronoun use; and vocabulary in terms of both mathematics vocabulary and ambiguous words" (p. 111). Criteria for determining which linguistic features to analyze were drawn from both existing literature recommendations and input from the team itself. For example, the team decided that references to American culture and holidays would be considered difficult for non-native English speakers. Two experienced teachers independently reviewed every test item and and coded the items for linguistic characteristics (Shaftel et al., 2006).

Data was drawn from a random sample of 8,000 students who took the Spring 2000 large-scale mathematics assessment (Shaftel et al., 2006). To avoid small sample sizes, researchers included data from all of the ELL students who took the assessment. ELL students who were also categorized as students with disabilities were removed from the ELL group to avoid extra variables. To be included in the study, linguistic features had to occur in at least 10 items in a grade level. Linguistic features that occurred less frequently than this were eliminated. This ensured that each feature analyzed would be sufficiently represented and researchers would be able to draw meaningful conclusions from them. Only 16 linguistic features occurred frequently enough across all grade levels to be included (Shaftel et al., 2006).

The linguistic complexity of each problem was analyzed using a checklist of 17 features (Shaftel et al., 2006). Researchers wrote the number of times each linguistic feature occurred for each test item. A new checklist was used for every test item. Researchers determined that both the total number of words per item and the number of sentences per item increased linguistic complexity. Word length was also a linguistic feature that was analyzed; words with 7 or more letters were considered long and were tallied. Features such as prepositions, pronouns, and relative pronouns were also considered. In addition, researchers counted "slang, idiomatic, ambiguous, or multiple-meaning words or phrases" (p. 126) in the same category for analysis. Homophones were separate from this category. There were a multitude of different verb categories that were considered linguistically complex; these included complex verb forms with 3 or more words (such as "would have been"), infinitive verb phrases, and passive voice. American culture was also taken into consideration; one category for references to American holidays was included, as well as a category for American cultural events or situations that could be unfamiliar for native English speakers. Words such as "picnic" and "dormitory" were included here. Conditional constructions (such as "if-then" statements) and clauses were also included as linguistic features. Finally, two categories were included that were relevant to the content being assessed: difficult or unusual mathematics vocabulary words, and comparative constructions such as "less than" and "greater than." Although these categories were included in the study, researchers recognized that their inclusion could muddy the results of the analysis; therefore, researchers performed another analysis excluding these two categories (Shaftel et al., 2006).

For the first analysis, the linguistic features studied had a significant effect on the students' scores, with the largest impact on Grade 4 scores, a moderate impact on Grade 7 scores, and a smaller impact on Grade 10 scores (Shaftel et al., 2006). In terms of which specific linguistic features had the greatest impact, the category including ambiguous words or words with multiple meanings had a strong effect on students in Grade 4. In addition, researchers found that "Words that are unclear, colloquial, or slang, or that have multiple meanings depending on context for interpretation, may be challenging and their use in any test items should be examined carefully" (p. 120).

A second analysis (Shaftel et al., 2006) excluded the two categories of difficult mathematics vocabulary words and comparative construction words. Researchers found different results for each grade level, as follows:

> For items at Grade 4, the results were similar: Prepositions, ambiguous words, and pronouns were still identified as important contributors to item difficulty levels though complex verbs, which were marginally influential when all variables were included, were not. For Grade 10 items, no additional variables were identified when math vocabulary and comparatives were excluded. For Grade 7 items, the variable of number of words in the item became statistically significant as a predictor (pp. 120–121).

Thus, in lower grades, linguistic features seem to be a source of great difficulty for students, and as students get older, linguistic features become less challenging. Moreover, complex linguistic features have an effect on ELL test scores.

2.3 Language background as a variable in NAEP mathematics performance (Abedi, 1995)

Jamal Abedi (1995) took research on linguistic modifications for ELLs one step further, conducting a two-phase study in which NAEP mathematics test items were linguistically modified for students with limited English proficiency. In the first phase, NAEP data from the 1990 and 1992 8th-grade mathematics main assessment item sets were examined. His analysis suggested that students who favored a language other than English at home scored significantly lower on NAEP math items than students who spoke only English in the home. Next, Abedi rated each test item based on the length of the item and the item's linguistic complexity. Abedi considered items to be "long" if the question was two or more lines, or if the answer choices were more than one line. Items were considered "short" if the question and the answer choices were each less than one line. Medium-length items were discarded as they were considered arbitrary. Linguistic complexity was determined based on "difficulty of vocabulary, abstract or culture-specific content, and number of complex structures in a sentence" (Abedi, 1995, p. 22). Abedi created two composites of items, one with greater complexity and one with lesser complexity.

The results from this analysis revealed that the gap between students who favored a language other than English at home and students who spoke only English at home was greatest for items of high linguistic complexity and long item length (Abedi, 1995). Finally, Abedi reviewed how many test items were omitted or not reached by the end of the test. He found that students who always spoke a language other than English at home had a far greater number of omitted or not-reached items than students who only spoke English at home. The results of Phase 1 of this study corroborate the notion that ELLs score much lower than native English speakers on NAEP mathematics test items, and that linguistic complexity and item length play a significant part in this uneven scoring.

In Phase 2 of Abedi's study, a subset of NAEP data from the 1992 8th-grade mathematics test items was examined for linguistic complexity (Abedi, 1995). Linguistic features chosen for this phase of the study were limited to features that actually occurred in the small subset of data used; these linguistic features included "familiarity/frequency of non-math vocabulary, voice of the verb phrase, length of nominals (noun phrases), conditional clauses, relative clauses, question phrases, and abstract or impersonal presentations" (p. 32). Abedi looked to the literature and anecdotal evidence to determine which linguistic features to analyze (Abedi, 1995). The seven linguistic features include (Abedi, 1995) the following.

- **The frequency and familiarity of words.** Researchers used reference books and staff judgment to determine which words to replace and use in problem statements. For example, some word frequency reference books used in the study listed "census" as more commonly used than the word "video"; however, project staff judged that students were more likely to be familiar with the word "video" than the word "census."

- **The use of passive voice.** According to Abedi, passive voice is more difficult to understand and occurs less frequently in conversational English than active voice. ELLs may have had less exposure to passive voice, and thus, it should be avoided when creating test items.

- **Length of nominals.** For example, "last year's class vice president" was replaced by simply "vice president" (Abedi, 1995, p. 35). ELLs may take more time to process long nominals, and these long nominals often contain unnecessary information irrelevant to the mathematics construct being assessed. As such, they should be shortened.

- **Conditional clauses.** The example Abedi uses is that the conditional phrase "If x represents the number of newspapers that Lee delivers each day ..." should be replaced with "Lee delivers x newspapers each day" (Abedi, 1995, p. 36). Separate sentences as opposed to conditional phrases may be easier to understand.

- **Relative clauses.** Abedi suggests replacing "A report that contains 64 sheets of paper ..." with "He needs 64 sheets of paper for each report" (Abedi, 1995, p. 37). ELLs may be less likely to encounter relative clauses because they are more common in written English than spoken English. Also, depending on a particular ELL student's native language, relative clauses could add an additional layer of confusion; for instance, in Chinese and Japanese, relative clauses come before the noun, while in English, relative clauses come after the noun. This difference in language structure could cause unnecessary confusion for ELLs.

- **Length of question phrases.** The question phrases used in tests can often be convoluted, unnecessarily long, and unnecessarily confusing. Abedi recommends shortening these phrases for clarity. For example, "At which of the following times...?" was replaced with "When ...?" (Abedi, 1995, p. 38). The revised version is also used more frequently in typical English conversation, making the question easier for ELLs to understand.

- **Abstract or impersonal presentations.** The sentence "The weights of three objects were compared using a pan balance. Two comparisons were made. . . " was replaced with "Sandra compared the weights of three objects using a pan balance. She made two comparisons . . . " (Abedi, 1995, p. 38). This removes the passive voice of the problem statement and replaces it with active voice, making the problem statement more narrative and story-like, which may be more easily remembered and understood.

Abedi also cited cultural relevance as a source of confusion for ELLs as they complete mathematics test items (Abedi, 1995). Davison & Schindler (1988) interviewed Native American students whose first language was Crow about mathematics vocabulary used during school. Students stated that much of mathematics vocabulary taught in English is only relevant to school activities and that out-of-school activities did not use this vocabulary. They noted that the mathematics problems had little cultural relevance to them, and that this interfered with students' ability to complete them.

Once researchers used the seven linguistic features to develop linguistically modified problem sets, two experts in mathematics education reviewed the problems to ensure that the math constructs being assessed were the same in each original and modified item (Abedi, 1995). Further revisions were made after taking expert opinion into account to ensure that both problem sets were assessing the same math constructs. Next, researchers interviewed a group of 38 eighth-grade students in the greater Los Angeles area using a structured interview format to gauge student perception on both the original and modified test item sets. Students strongly preferred modified test items. Students felt that the revised items were easier to understand and would take less time to complete. In addition, when asked to read the items out loud, students often stumbled over more complicated words in the original test items, or would make substitutions, such as using active voice when the test item used passive voice. The students read the modified test items as they were written without stumbling or substitution. It is noted that "Student preference for the revised items gave support to the notion that the math items could be linguistically simplified in meaningful ways for the test taker" (Abedi, 1995, pp. 43–44).

Finally, both the original and modified tests were administered to a group of 39 8th-grade classes for a total of 1031 students, with an overrepresentation of ELL students (Abedi, 1995). Sixty-one percent of students spoke a language other than English at home. Sixty percent of these students spoke Spanish as a first language. The test items were sorted into two booklets. Booklet A contained 10 original items and 10 different modified items, while Booklet B contained the 10 original items and the 10 modified items that were not in Booklet A. Each booklet also contained five additional original NAEP items employing simple language with little potential for confusion. These questions were used as a control. When creating the test booklets, researchers distributed various aspects of the problems equitably—for example, each booklet had an equal number of problems containing passive voice, algebra, visual aids, and so on. Each booklet also had average difficulty (Abedi, 1995).

When results were analyzed, researchers found that ELLs in low-level and average-level math classes showed the greatest improvement with the linguistically modified test item sets, while the effect was less pronounced for students in intermediate- to high-level math classes (Abedi, 1995). Abedi explains this difference in performance:

> Since language ability is, in general, a predictor of math performance, it is possible that the language simplifications had little effect on the algebra and honors students' performance because these high-performing students also had strong language ability and had no problem understanding the original items. Although the original items were longer and more complex linguistically, they did not slow down the top students. If the students in low and average math classes had correspondingly low or average

> language comprehension skills, the small changes in the revised items could well have led to greater comprehension and relatively greater improvement in their scores. (p. 61)

Overall, it is clear from the results of Abedi's (1995) study that linguistic modifications do have a positive impact on ELLs, with this impact being the greatest in low- and average-level math classes.

2.4 Accommodations for English language learner students: The effect of linguistic modification of math test item sets (Sato et al., 2010).

Sato et al. (2010) conducted a study similar to Abedi (1995). They modified NAEP test items and California state standardized test items, gave examinations to students, and analyzed results (Sato et al., 2010). Students were arranged into three subgroups: English language learners (often referred to as ELs in this study), non-ELLs who were not proficient in English, and non-ELLs who were proficient in English. If a student had a CELDT (California English Language Development Test) score, they were classified as an ELL student, as only ELLs take this state test. Students who scored at the lowest proficiency level of the CELDT were not included in the study. Non-ELLs were placed into the "English proficient" subgroup if they scored at or above the proficiency cutoff for state standardized testing, and those who did not meet this score were placed in the "non-English proficient" subgroup. In total, 4,617 seventh and eighth grade students from 13 middle schools across five school districts in California were included in the study. All students in the ELL subgroup spoke Spanish. The original and linguistically modified test booklets were randomly distributed to students. In order to ensure that each of the two test booklets were distributed equally across all three subgroups, researchers examined completion frequencies during the analysis process (Sato et al., 2010).

In selecting the test item sets and creating the corresponding linguistic modifications, a work group consisting of "the core study team (senior researchers) and experts in mathematics, linguistics, measurement, curriculum and instruction, and the EL student population" (Sato et al., 2010, p. 22) was assembled. Public NAEP eighth grade test items along with seventh grade test items from the California Standards Test were collected and reviewed by the work group members. Content specialists ensured that each item selected was aligned to a state standard and that the group of test items contained a wide variety of item types, item complexities, and math content areas. To make the linguistic modifications, researchers used a wide variety of criteria (Sato et al., 2010). For example, researchers sought to make the test items familiar to all students without cultural bias by using high-frequency words and avoiding references to American culture. When using familiar vocabulary, words were kept precise; researchers did not use more common vocabulary if doing so reduced the clarity of the problem. In addition, researchers kept vocabulary which was not relevant to the construct being assessed at or below grade level. The complexity level of the sentence structure was kept at or below grade level as well (Sato et al., 2010).

Long or compound problem statements were shortened or separated into several sentences to make the items more readable and less overwhelming (Sato et al., 2010). Conditional (if-then) sentences were separated into two sentences. Researchers also removed any irrelevant words, phrases, and sentences so that students would not waste any time reading unnecessary parts of the test item. Questions that had a negative structure were reframed to have a positive structure; for example, the question "Which one of the following choices is not correct?" was changed to "Which one of the following choices is correct?" Relative clauses were also modified. Statements like "A box that holds ten oranges. . . " was replaced with "Juan needs 10 oranges for each box" (Sato et al., 2010).

Certain verb forms were avoided and modified. For example, passive voice was changed to active voice (Sato et al., 2010). Verbs in past, future, and conditional tenses were changed to present tense. Certain categories of words were avoided, including ambiguous words, words that have more than one meaning, words that are spelled irregularly, proper nouns that are not construct-relevant, words that can function as both nouns and verbs, compound words, and gerunds.

Researchers also strived to provide context for test items, stating, "Context that facilitates access for English language learner students is expressed in concrete language, illustrative language, and illustrations/graphics" (Sato et al., 2010, p. 86). Context can help ELLs make sense of test items, allowing students to focus on the construct. One way to provide context is to use graphics to reduce language complexity from test items. Graphics should be used to clarify math constructs, context of the math constructs, mathematical operations, and so on. If a graphic contains construct-irrelevant information, it may need to be removed or changed. Researchers also suggested using bulleted lists when it would aid in processing larger amounts of text. For example, some paragraphs may be able to be meaningfully broken down into shorter bullet points.

When making the linguistic modifications, all changes and explanations for specific linguistic modification strategies were documented (Sato et al., 2010). Mathematics content experts and test development specialists verified that each item's mathematical constructs were not changed. Sato et al. explains the importance of this process, stating, "Linguistic modification may remove nonessential language to make an item less linguistically dense or complex, but it should not alter the math knowledge and procedures required to solve the problem. For test results with linguistically modified items to yield valid interpretations, the math content of a linguistically modified item must be comparable to that of the original item" (p. 8). To further demonstrate that the constructs did not change, researchers confirmed in a secondary analysis that the linguistic modifications used in the study did not alter the mathematics constructs being assessed. This secondary analysis process was done through three analyses—differential item functioning (DIF) analysis, exploratory factor analyses (EFA), and a correlational analysis. All confirmed that the math constructs were not changed (Sato et al., 2010).

Next, after the linguistically modified math item sets were created, cognitive interviews were conducted with nine 7th and 8th graders, including 5 ELLs and 4 non-ELLs. The objective of the cognitive interviews was to "ensure that the test item set represented a range of knowledge and skills, levels of cognitive complexity, and linguistic modification strategies" (Sato et al., 2010, p. 28). After the cognitive interviews, the item pool was whittled down to 30 items and their linguistically modified counterparts. These items were used in pilot testing. For pilot testing, the tests were administered to 64 ELL students and 48 non-ELL students in 7th and 8th grade. A research team comprised of people with "expertise in assessment, applied linguistics, math content, and the EL population" (p. 29) analyzed the results of the pilot test, including test administrators' observations. The team made suggestions on how to further change the items and decide which items should be removed from the study. The team considered "item format, item content, and performance data" (p. 29). After pilot testing was conducted, 5 of the 30 items were eliminated, leaving 25 remaining to be used for the main portion of the study (Sato et al., 2010).

Results showed that the difference in score between the original item set and the linguistically modified item set was greatest for ELLs, followed by native English speakers who were not proficient in English (Sato et al., 2010). For native English speakers who were proficient in English, the difference in scores between the original and modified item sets was close to zero and statistically insignificant. This shows that the linguistic modifications chosen for this study benefited ELLs without increasing scores for all students.

From the four articles reviewed above (Abedi, 1995; Fairbairn & Fox, 2009; Sato et al., 2010; Shaftel et al., 2011), myriad linguistic modification criteria were selected for the paper. These criteria were commonly mentioned across the literature reviewed and were also highly effective based on the analyses conducted in the studies. These linguistic modification recommendations are provided in Table 1 in the appendix.

3 Conclusion

From the research, it is clear that there are many important linguistic aspects to keep in mind when creating mathematics test items to ensure that they are equitable to ELLs. However, we have not yet investigated actual examples of what a typical math test item looks like before and after it is modified, or how different researchers would modify the same math test item. We will explore actual examples of modified test items in the next issue of the ELL Corner of the *Ohio Journal of School Mathematics*.

References

Abedi, J. (1995, July). Language background as a variable in NAEP mathematics performance. *NAEP TRP Task 3d—Language background study. Final deliverable*. (Rep.). Retrieved `https://eric.ed.gov/?id=ED404176` (ERIC Document Reproduction Service No. ED404176).

Davison, D. M., & Schindler, S. E. (1988). Mathematics and the Indian student. In Reyhner, J. (Ed.). *Teaching the Indian child: A bilingual/multicultural approach*. Billings, MT: Bilingual Education Program.

Fairbairn, S. (2006). *English language learners' performance on modified science test item formats: A pilot study*. Dissertation Abstracts International. DAI-A 68/01. (Publication No. AAT 3248008).

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

Norval, B. (2019). ELL Corner: Can we change mathematics test items to be more equitable to ELLs? (Part 1 of 3). *Ohio Journal of School Mathematics, 82*, 29–35.

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C. (2010, June). *Accommodations for English language learner students: The effect of linguistic modification of math test item sets* (Rep. No. NCEE 2009-4079). Retrieved from `https://eric.ed.gov/?id=ED510556` (ERIC Document Reproduction Service)

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(**2**), 105–126.

Brooke Norval, `norvalbe@miamioh.edu`, is currently a Graduate Associate in the Department of Teacher Education at Miami University with an undergraduate degree in Mechanical Engineering from The Ohio State University. Brooke's research interests include international perspectives of mathematics education and teaching and learning with English language learners.

Table 1. Linguistic modification recommendations (Abedi, 1995; Fairbairn Fox, 2009; Sato et al., 2010; Shaftel et al., 2011).

1. Reduce sentence length and complexity (Abedi, 1995; Fairbairn & Fox, 2009; Sato et al., 2010; Shaftel et al., 2006)
2. Change past, conditional, or future tense verbs to present tense when possible (Abedi, 1995; Fairbairn & Fox, 2009; Sato et al., 2010)
3. Construct sentences using active voice instead of passive voice (Abedi, 1995; Fairbairn & Fox, 2009; Sato et al., 2010; Shaftel et al., 2006)
4. Avoid irrelevant words or phrases (Sato et al., 2010)
5. Vocabulary should be at or below grade level (Fairbairn & Fox, 2009; Sato et al., 2010)
6. Use high-frequency words (Abedi, 1995; Fairbairn & Fox, 2009; Sato et al., 2010; Shaftel et al., 2006)
7. Avoid ambiguous words, unnecessary words, or words that have multiple meanings (Sato et al., 2010; Shaftel et al., 2006)
8. Avoid proper nouns that are construct-irrelevant (Sato et al., 2010)
9. Avoid words that function as both nouns and verbs (Sato et al., 2010)
10. Avoid hyphenated words and compound words (Sato et al., 2010)
11. Avoid gerunds (Sato et al., 2010)
12. Remove unnecessary introductory phrases (Abedi, 1995; Sato et al., 2010)
13. Break compound sentences into two separate sentences (Abedi, 1995; Sato et al., 2010)
14. Break conditional clauses into separate sentences (Abedi, 1995; Sato et al., 2010)
15. Use bulleted lists (Sato et al., 2010)
16. Avoid references to American culture or holidays (Sato et al., 2010; Shaftel et al., 2006)
17. Change complex question phrases to simple question words (Abedi, 1995)
18. Reduce the number of linguistic elements for lower item difficulty (Shaftel et al., 2006)
19. Avoid relative pronouns that do not have a clear antecedent (Sato et al., 2010)
20. Use visuals that mirror the wording of the text (Fairbairn & Fox, 2009)
21. Avoid colloquialisms or slang (Fairbairn & Fox, 2009; Shaftel et al., 2006)
22. Rephrase negative questions into positive questions (Sato et al., 2010)
23. Shorten unnecessarily long nominals (Abedi, 1995; Sato et al., 2010)
24. Remove relative clauses (Abedi, 1995; Sato et al., 2010)
25. Remove abstract or impersonal presentations (Abedi, 1995)