EDUCATIONAL ASSESSMENT, *11*(2), 105–126

The Impact of Language Characteristics in Mathematics Test Items on the Performance of English Language Learners and Students With Disabilities

Julia Shaftel
Center for Psychoeducational Services
University of Kansas

Evelyn Belton-Kocher
St. Paul Public Schools
St. Paul, Minnesota

Douglas Glasnapp and John Poggio
Center for Educational Testing and Evaluation
University of Kansas

The impact of language characteristics in mathematics test items on student performance was evaluated for students with disabilities (SWD) and English language learners (ELL) as well as a large general student sample. Relationships were examined for test items and students at 4th, 7th, and 10th grades. The individual test item was the unit of analysis. Student performance was represented by item difficulty, or the probability of answering the item correctly. Regression analyses were conducted to examine relationships between item linguistic characteristics as independent variables with item difficulty as the dependent variable. Language characteristics had moderate effects on item difficulty at 4th grade, dropping to small-to-medium effects at 10th grade. ELL and SWD groups were not disproportionately affected by language characteristics in these test items. Difficult mathematics vocabulary had a consistent effect on performance for all students at all grades. Ambiguous or multiple-meaning words increased item difficulty at 4th grade.

Correspondence should be addressed to Julia Shaftel, Center for Psychoeducational Services, University of Kansas, 1122 West Campus Road, Room 130, Lawrence, KS 66045. E-mail: jshaftel@ku.edu

Recent federal legislation has increased requirements for the inclusion of English language learners (ELL) and special education students in large-scale assessments for purposes of accountability, and publication of performance data by these groups is now required (No Child Left Behind Act of 2001 [2002]; Pitoniak & Royer, 2001). One of the major concerns in all assessments is the extent to which any test contains irrelevant sources of score variance, such as the impact of reading difficulty on mathematics test items, that might unfairly impact performance for some students. The complexity of the English used in mathematics items is theorized to have a disproportionate impact on ELL students and students with disabilities (SWD) due to lower English proficiency or general language skills (Abedi, 2004; Johnson & Monroe, 2004). Simplified English assessments have been proposed as one way to address the language factor in large-scale assessments.

Sources such as Kopriva (1999) and Hanson, Hayes, Schriver, LeMahieu, and Brown (1998) have provided guidelines for the development of simplified English assessments. Examples of linguistic simplifications include reducing the total number of words, avoiding passive voice and complex sentences, minimizing difficult vocabulary, and avoiding ambiguous or multiple-meaning words. However, there is little research on which specific language features have the greatest impact on the performance of vulnerable student populations.

Several studies have investigated whether linguistic simplification of mathematics items improves the performance of ELL students. A series of studies on this issue using National Assessment of Educational Progress (NAEP) items by the Center for Research on Evaluation, Standards, and Student Testing (CRESST) produced mixed results. In a 2001 study (Abedi, Hofstetter, Baker, & Lord, 2001) using 1996 NAEP mathematics items, the impact of several modifications, including linguistic simplification, was addressed with eighth-grade Limited English Proficient (LEP) and non-LEP students. In this often-cited study, Abedi et al. stated that modified language reduced the score difference from 5.49 to 3.31 points between LEP (ELL) and non-LEP (English proficient and non-ELL) students on a 35-item test. However, a large proportion of the reduced score difference was due to unexpected lower performance by English proficient students on the modified English version (mean 15.94) as compared to the original English version (mean 17.56). The score improvement of ELL students on the modified English version was slightly more than one half of a point (12.63 vs. 12.07). In two earlier studies (Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, & Plummer, 1997), modified wording increased the performance of non-ELL students but did not impact the performance of ELL students. In contrast, Abedi and Lord (2001) found that simplifying language improved performance for almost all students tested on modified eighth-grade NAEP items when compared to original items, including a 3.7% improvement for ELL students and 6.7% improvement for students in low math classes. Only students in the highest math classes were not helped by the simplified English math test items. Rivera and Stansfield (2001) and Brown (1999) studied the impact of linguistic modification on items from the Delaware state mathematics and science assessments. Unfortunately, the small size of the ELL samples in these studies rendered the results inconclusive regarding ELL students. However, Brown also found that linguistic simplification improved the performance of the non-ELL students in the study.

Three studies included special education students in simplified language investigations. Kiplinger, Haug, and Abedi (2000) attempted to replicate the Abedi et al. (2001) study with fourth-grade students in Colorado. For the highest performing schools, a glossary accommodation was helpful to all but the most English proficient students, and simplified wording also helped the least English proficient group. Students with disabilities performed equally well on the original version and the version with glossary accommodation but more poorly on the simplified version of the test. No differences were found by accommodation at the lower performing schools, perhaps due to overall test difficulty. Johnson and Monroe (2004) evaluated responses from 1,232 seventh-grade students, including 138 SWD and 34 ELL students, to two test forms containing half original-language and half modified-language math items in a counterbalanced design. Not surprisingly, ELL and SWD scores were statistically significantly lower than scores of the general education group overall, but the SWD group performed slightly better on the simplified language test, whereas ELL and general students performed slightly better on the original version. Though effect sizes were not reported, all mean differences were less than one half point out of 12 total points. Tindal, Anderson, Helwig, Miller, and Glasgow (2000) assessed alternate forms of a math test using standard and simplified language with 48 seventh-grade general (32) and special education (16) students. No statistically significant differences were found in performance for either group on either form of the test. However, they discovered that simplifying test items was not a straightforward process as many of their simplified items became more difficult than the original versions.

Shaftel, Belton-Kocher, Glasnapp, and Poggio (2002) evaluated linguistically modified state mathematics assessment items with general education students, ELL students, and SWD in three ways. First, they compared the performance of general education students at three grades randomly assigned by whole classes to test forms containing modified and original test items in a counterbalanced design. No statistically significant differences in performance between original and linguistically simplified items were found. Second, responses for ELL students on original items in the mathematics assessment from Spring 2000 were compared to responses from a new group of ELL students taking the same test items in modified format during the Spring 2001 test administration. Responses to a set of unchanged test items occurring on both test forms and administered to both groups served as a covariate. Grade 7 ELL students performed slightly better on the original items whereas Grade 4 and Grade 10 students performed slightly better on the modified test items, with no practical differences in overall mean scores. Third, three-parameter IRT methodology was used to estimate item difficulty parameters for all items, original and modified, using the common set of items as an anchor block. The results showed that the modified or plain English items were slightly more difficult than the original test items at Grade 7 and equivalent at Grades 4 and 10, with slightly lower item discrimination values at all grades. These three studies provided no evidence that linguistic simplifications provide a boost for ELL students, and the conflicting results at different grades are consistent with other studies on plain English as a test modification.

Solano-Flores, Trumbull, and Kwon (2003) have provided a conceptual framework based on perspectives from multiple disciplines to study the linguistic complexity of mathematics items. In this continuing study, the investigators parse items as tree diagrams according to the conventions of structural linguistics and use graph theory to examine language complexity. A series of indicators are computed that summarize the features of the sentences. Preliminary findings show that NAEP mathematics items for students as young as fourth graders are extremely complex, suggesting that item writers did not consider the level of verbal sophistication of young students who cannot process language as adults can. Though this exciting line of research may provide a needed contribution, this program is still in its early stages, and no data have been reported using this approach.

Studies examining variables that affect student performance on mathematics problems or test items have provided some insight into the role of readability and vocabulary. De Corte, Verschaffel, and De Win (1985) and Cummins, Kintsch, Reusser, and Weimer (1988) conducted experiments in which word problems requiring simple computation were written in three formats characterized by either a change in quantity, a combination of quantities, or a comparison of quantities. Different forms of each problem used different wording and different sequences of revealed information required to solve the problem. In studies with early elementary aged children, these researchers found that difficulty with word problems was related to difficulty in comprehending abstract or ambiguous language, including common words, which led to semantic misinterpretation. Apparent mistakes made by students, therefore, were often actually correct answers for the misunderstood problems. Revising the language of the items to make relationships clearer had a positive impact on performance. Hanson et al. (1998) asked students to “think aloud” while solving original and simplified math and science test items. Transcripts showed that the most common problem was encountering difficult or unfamiliar vocabulary, accounting for 33% to 67% of student errors.

Two studies (Larsen, Parker, & Trenholme, 1978; Wheeler & McNutt, 1983) used specially created mathematics assessments at three levels of syntactic complexity for eighth-grade students. In both of these studies, syntactic complexity, including the use of compound and complex sentences with more words per sentence (but not more words per test item), affected the ability of low-achieving eighth-grade students to solve math word problems even when the problems were at or below the students' computational and reading–vocabulary level. Lepik (1990) considered linguistic and structural variables (e.g., the number of known quantities, the number of unknown quantities) and their relationship with two performance measures, proportion of correct strategies and average solving time. Although 14 of the structural variables were good predictors of both performance variables, only one linguistic variable, the average number of words for each relationship, predicted proportion correct, and the best predictor of problem solving time was the number of words in the problem statement.

Whereas the previous studies looked only at the performance of English-speaking general education students, Lean, Clements, and Del Campo (1990) used two samples of students from Australia and Papua New Guinea, one a sample of English-speaking students and one a sample of students whose first language was not English. The English-speaking sample included students in Grades K–6. The students who were not native English speakers were in Grades 4–6. For this study the investigators constructed a special set of mathematics items in English. The purpose of the study was to analyze the effects of several factors, including linguistic factors, on student performance. The major finding of this study was that both groups of students had more difficulty with comparison problems. The investigators proposed that these types of problems are semantically complex and therefore children adopt erroneous strategies for solution.

Lord, Abedi, and Poosuthasee (2000) evaluated several linguistic features of mathematics items from the Delaware Student Testing Program and the Stanford Achievement Test, 9th Edition, to compare the performance of ELL and non-ELL students. No individual item features explained the score gap between ELL and non-ELL students using a weighted scores approach. However, evaluation of an item's disparity quotient and its complexity rating for each language characteristic showed that the score gap was greater at Grade 8 for longer items and for constructed response rather than multiple-choice items. Item length accounted for about 10% of the variance in the score difference between the two groups of students. The researchers found that third-grade items were carefully written with low incidence of potentially problematic language features such as relative clauses and passive voice.

In a recent CRESST study (Abedi, Courtney, & Leon, 2003) investigating other accommodations, NAEP items were rated on a scale of 1 to 5 for linguistic demand using a rubric developed for this study. Data were analyzed using multivariate analysis of covariance. ELL status explained 7.8% of the variance on the linguistically demanding items at Grade 4 and 21.3% of the variance at Grade 8, whereas ELL status accounted for only 4.5% of variance at Grade 4 and 9.3% of the variance at Grade 8 for the less demanding items. This study indicates that although linguistic features have some impact on the performance of ELL students, a large proportion of score variance is explained by other factors.

Results from the previously described research are inconclusive on performance improvement for general education students, SWD, and ELL students for linguistically simplified items versus traditionally written items. In addition, this research has often been hampered by small sample sizes and the absence of special education students in many of these studies. Prior research provides only limited guidance about which linguistic features affect the performance of special populations. Studies with single populations have demonstrated that ambiguous wording, item length, difficult vocabulary, syntactic complexity with longer sentences, and comparison problems may contribute to item difficulty.

This study set out to evaluate the overall impact of test item language on performance with particular attention to which language features have the most effect and which student groups are most vulnerable. Questions addressed by this study include the following:

1. Do linguistic features affect the difficulty of mathematics test items?
2. Do linguistic test item features affect English language learners and students with disabilities disproportionately when compared to a general student sample?
3. Which language features have the greatest impact on student performance?

METHOD

Instruments and Procedures

Mathematics test items. The items that serve as the unit of analysis in this study are all original items in the Kansas general mathematics assessments given at Grades 4, 7, and 10 with four parallel forms of the assessment available at each grade. The items are based on state mathematics curricular standards and include knowledge and application items (Glasnapp, Poggio, & Omar, 2000). The standards and items were written in four mathematical domains: number and computation, algebra, geometry, and data. The item pool comprised of 208 items at 4th grade, 203 items at 7th grade, and 183 items at 10th grade. The mathematics standards used in this assessment (2000) had received outstanding external review regarding their quality and comprehensiveness (Finn & Petrilli, 2000). Items reviewed for this study included the entire pool of scored items from these assessments. All items used a multiple-choice format. Furthermore, all items were presented as word problems, though the number of words per item ranged from 2 words (in six items at 4th grade) to 177 words (in three items at 10th grade), with a mean of 45 words.

Published simplified English guidelines for writing test items from several sources (Abedi et al., 2001; Abedi et al., 1998; Abedi et al., 1997; Hanson et al., 1998; Kopriva, 1999) were initially used to identify a set of potentially important linguistic features, including all of the features deemed to influence item language difficulty by those experts. However, the test items themselves were not written or modified according to plain English guidelines but were the originally worded items from the general mathematics assessments. The item rating rubric for scoring the linguistic features (see Appendix) was reviewed by professionals including mathematics teachers, math assessment specialists, and a speech–language pathologist who specialized in second language learning. The features included the 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. Many of the characteristics were drawn from recommendations for reducing language difficulty but for which there are no empirical data. Others were features that common sense suggests would present obstacles for low achievers or for nonnative English speakers, such as references to American culture and holidays. Two of the linguistic features—mathematics vocabulary and comparative terms such as "greater than"—are clearly pertinent to mathematics and cannot be considered irrelevant language characteristics.

Each item was assessed and coded for linguistic characteristics by two independent raters, both experienced teachers. Interrater reliability was not computed because coding questions and issues raised by the first rater led to redefining some of the criteria for the second rater, with subsequent review and recoding of confusing criteria (e.g., relative pronouns, types of clauses) by the first two authors. The meanings of the confusing criteria were not changed; rather, redefinitions consisted of tightening the descriptions so that the number of instances to be counted was better controlled. For example, large words were redefined to be words of seven letters or more rather than words of six letters or more, which were common. The list of relative pronouns was made comprehensive to eliminate confusion about what a relative pronoun is. Clauses were redefined as dependent or independent clauses with a subject and verb, because some prepositional phrases and adverbial phrases had been erroneously counted as clauses. Rating of mathematics features and all review of linguistic features was accomplished by the first and second authors in consultation with one another.

Student groups. Student data from the large-scale mathematics assessments of Spring 2000 were selected because all four original test forms were administered at each grade level, thus maximizing the pool of items for which data were available. In previous years, only one test form was typically used per grade level. Further, no alternative simplified English test forms were available that year. In subsequent years, simplified English forms were developed for ELLs, and the use of responses to these alternate test forms and items would not necessarily have been comparable with responses to the originally worded items used in 2000.

The scored responses and demographic data for 8,000 students were randomly selected from the total number of students at each grade level who completed the Spring 2000 assessments, providing a sample of more than 20% of the students at each grade level. This large sampling procedure, chosen to establish stable item difficulty estimates, offered item data for approximately 2,000 students for each test form at each of the three grade levels, with a range of 1,949 to 2,074 students per test form. These random general education samples were not subsequently modified in any way: They included students with disabilities and ELL students in the proportions naturally occurring in this state at each grade level.

Next, special education students included in these samples were identified by their disability coding and categorized as a separate group. SWD who were also coded as ELL were removed from the special education groups (but not the overall groups), leaving 650 to 809 special education students per grade level with robust sample sizes of 154 to 235 having data on each of the four test forms per grade level. This was done so that any conclusions about the effects of linguistic item features on students with disabilities would not be confounded with any other student status variables.

Even these large general education samples, however, provided only 16 to 61 ELL students per test form, which were unacceptably small sample sizes. Therefore, the responses for all ELL students who completed the assessments were obtained for each grade level, providing entire populations of 328 (at 10th grade) to 905 students (at 4th grade), with 71 to 241 ELL students per test form. Even the smallest group of 71 is sufficient for item analysis, and because this represented the entire group of ELL students responding to that particular test form, the item difficulties for those test items have been measured exactly. ELL students who also received special education services were removed from the ELL groups to avoid counting their data in both of the smaller groups and to avoid confounding interpretation of the results.

To summarize the sampling procedures, the general student group included SWD and ELL students in the proportions in which they occurred in the general population, including students who might belong to both target groups. The groups of SWD excluded students with ELL status, and the ELL groups excluded SWD, so that those two groups represented only students of the status under investigation, and no student could appear in both groups. The ELL groups included all students of ELL status who completed these tests, and hence item properties are not estimates but have been measured exactly for this population. Adequate group sizes (71–241) were achieved for all subgroups on each test form.

It should be noted that the lowest performing students from each of the ELL and SWD groups had not been required to take these general assessment tests in the first place and so they were not included in these samples. The participation of SWD in large-scale testing is determined by IEP teams who follow state guidelines for determining eligibility for the general tests. Alternate tests were available for SWD who were not being instructed in the curriculum covered on these tests. Similarly, ELL students participated in the general assessments on the basis of time in the district and scores on a standardized test of language proficiency. ELL students who scored below proficiency cutoffs or who were new arrivals to the district were exempted from testing. These guidelines were put in place to reduce participation of students for whom these assessments would not have been appropriate measures of their mathematics knowledge. Therefore, though low-performing students from all three groups were certainly represented in the item data, a systematic attempt was made to exclude students who did not possess minimum levels of language proficiency, cognitive ability, or exposure to the grade-level mathematics curriculum from taking these tests during regularly scheduled state test administration.

Mean item scores, or item difficulties, were computed for each item on each test form for three groups of students: First, the total sample at each grade level; second, the special education students represented within the total sample; and third, the total ELL groups. As the ELL groups consisted of the entire body of ELL students who completed these assessments in 2000, the mean performance for each item is not an estimate of the population mean but is the actual difficulty for each item. The mean item score represents the probability for members of that group to answer that item correctly. It must be noted that the item difficulty parameters used in this study are dependent on each group of students who provided item response data; hence, adequate numbers of students were required to provide this information. However, groups of equivalent size were not necessary as student groups were not compared to one another.

Because the unit of analysis in this project was the individual test item, all student-level information was collapsed into the mean score for each item for each group. These mean item difficulties, therefore, regardless of the size of the group used in their calculation, became a single variable in the analyses conducted for this project. Because students were no longer used as the unit of analysis, the size of each student group had no effect on statistical significance in any of the subsequent item-level analyses. The large number of test items, however, provided ample power to detect the impact of item attributes on student performance at each grade level.

This also meant that all other factors contributing to item difficulty for individual students, such as language proficiency, verbal ability, academic achievement, confidence, motivation, variations in the testing environment, and category of disability within SWD, along with many others, were aggregated across the students who provided data for each item. In the ELL groups, all participating ELL student data were included, so the entire range of these factors was represented. For the other groups, the random selection of individual students should have eliminated any systematic expression of these factors on item difficulties. Therefore, the effects of the linguistic features should be interpretable on their own merits.

The anonymous test data available for analysis did not include several other factors of potential interest. The only other meaningful variable available to researchers was special education category for SWD. There were two reasons for not including this variable in the study, however, in addition to retaining the disaggregated groupings called for by the No Child Left Behind Act of 2001. First, subdividing special education students on the basis of categorical placement would have resulted in numerous groups too small for reliable interpretation. Second, in other research, including special education category as a variable has shown no systematic relationship with performance except for overall ability in the assessed content area (Yang, Shaftel, Glasnapp, & Poggio, 2005). To give a general sense of the characteristics of this sample group, more than half had been identified as having a learning disability (56%), with the next largest subgroup having a speech–language impairment problem (12%). The remaining 32% of the SWD sample was spread across 11 separate categories of disabilities.

Analyses

Two preliminary steps were taken before the actual regression analyses were performed. First, the distribution of item difficulties or *p* values was examined for each student group. A bounded distribution, such as that of item *p* values, may have skewness and/or kurtosis that could affect interpretation of regression analyses using these values as criterion variables. Before proceeding with the analyses, it was necessary to evaluate the characteristics of these distributions.

Second, because a large number of item language features had been identified and counted, it was anticipated that some would occur rarely or not at all in the items of any particular grade level. Therefore, the frequency of each item characteristic was examined at each grade level to identify and remove features that were not sufficiently represented to provide valid information about their effects on item difficulty. Attributes that did not occur in at least 10 items at a grade level, or about 5% of the items, were eliminated. The linguistic feature of reference to American holidays occurred in only two items at Grade 7 and three items at Grade 10, so it was removed from consideration for all analyses. All other linguistic features (16) occurred frequently at all grades and served as a common set of independent variables for all analyses.

The central investigations of the study involved multiple linear regression analyses conducted to examine the relationship between item linguistic characteristic scores as predictor variables with item difficulties for a specific group and grade level as the criterion variable. Within a grade-level set of items, the same set of item characteristic scores was replicated three times, and each replication was linked with the unique item difficulties for one of the three separate student groups, that is, item difficulties for the general student sample, for the ELL student sample, and for the SWD student sample. Dummy coding of items based on the separate group or grade-level item difficulties was used to identify the group and grade-level effects, and interaction variables were computed for each item linguistic feature with the dummy coded group and grade-level variables. To test appropriate effect hypotheses in each analysis, sets of variables were entered into the regression equation in a hierarchical forced entry order. All of the language feature variables were entered at step 1 as a block, then the dummy coded group or grade-level variables were entered at step 2, and then interaction variables for group or grade level with linguistic characteristics were entered as a block in step 3. Statistical significance testing of the change in R^2 values from step 2 to step 3 (interaction variables) was conducted to identify any differential relationships among item linguistic characteristics and item difficulties across groups or grade levels. If the test of the interaction variables contribution was not statistically significant, evidence would exist to indicate that the pattern of linguistic characteristics important to the prediction of item difficulty did not differ across groups or grade levels. Within the separate regression analyses, the beta weights, tests of statistical significance, and semipartial correlation coefficients for individual item attributes were examined to determine which features had independent and unique main or interaction effects on item difficulty.

RESULTS

Descriptive statistics for the combined grade-level distributions of item mean scores are provided in Table 1. The overall mean item difficulty (*p* value) for the general student group for all grades is almost precisely 0.50, whereas the two target subgroups performed considerably worse, with mean item difficulties of approximately 0.36 for each group. The effects of many of the myriad individual variables that influence performance, such as English language proficiency, familiarity with test items and formats, cognitive ability, and exposure to appropriate instruction of grade-level content, are clearly evident in the depressed mean scores for these vulnerable groups. Skewness and kurtosis statistics reveal some deviations from normality, with both of the subgroups' item mean distributions positively skewed, and with the SWD group showing slight positive kurtosis. The overall student group's item mean distribution shows slight negative kurtosis and miniscule positive skewness. With the large number of item means used in the analyses (n = 594 test items), the standard errors of kurtosis and skewness decrease, and statistically significant deviations from normality become likely even when the practical results are negligible (Tabachnick & Fidell, 2001). For these empirical distributions, major deviations from normality are not evident on either visual or statistical inspection. Furthermore, multiple regression is generally considered to be robust regarding violations of distributional assumptions (Pedhazur, 1997).

TABLE 1
Descriptive Statistics for Item Mean Score Distributions for Each Student Group

	General Population	*Students With Disabilities*	*English Language Learners*
Number of test items	594	594	594
Mean item difficulty	.5001	.3604	.3554
Median	.4878	.3349	.3294
SD	.17461	.14858	.15685
Variance	.030	.022	.025
Skewness	.265	.837	.762
Standard error of skewness	.100	.100	.100
Kurtosis	–.634	.427	.217
Standard error of kurtosis	.200	.200	.200

A 3 × 3 univariate analysis of variance (ANOVA) was conducted to evaluate whether there was a grade (4, 7, and 10) by student group (all, SWD, ELL) interaction for overall item means. Though statistically significant, the interaction effect size was extremely small (p = .020, F = 2.921, partial eta squared = .007). As shown in Figure 1, the SWD group performed slightly higher than the ELL group at Grades 4 and 7, and the ELL group scored slightly higher than SWD at Grade 10. However, Scheffé post hoc simple effect tests showed that these differences between the scores of the ELL and SWD groups were not statistically significant at any grade level. Both of these latter group means were statistically significantly lower than the general all-student group at each grade.

Multiple regression analyses conducted to evaluate the interaction of grade level and linguistic features indicated that the effects of language variables differed among the three grades (R^2 change for the set of grade interaction variables = .055). Due to this interaction of grade with linguistic features, as well as the group by grade ANOVA interaction described earlier, all subsequent analyses were conducted within each individual grade level.

Similar multiple regression analyses were conducted to assess the interaction of group with linguistic features within each of the three grades. An outcome of consequence is that no statistically significant interaction effects were found, thus indicating that the pattern of item linguistic characteristics important in contributing to the prediction of item difficulties did not differ across groups. This was evident in the nonsignificant overall changes in R^2 when the group interaction terms were added to the regression equations (R^2 change = .005 for Grade 4 items, .003 for Grade 7 items, and .004 for Grade 10 items) and in the absence of a statistically significant unique effect on item difficulty for any of the individual item linguistic features. This demonstrated that there was no differential impact of test item language for the SWD or ELL groups when compared to the overall group or to each

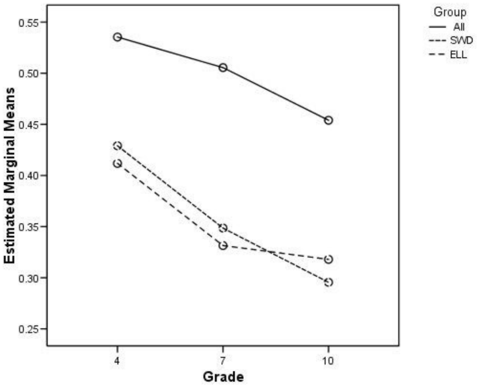

FIGURE 1 Overall item means for each group at each grade level.

other, and that the relationship of linguistic characteristics and item difficulties could be summarized by the pooled group coefficient information from the step 2 analyses. As anticipated, the analyses also indicated that the mean item difficulties differed across groups when adjusted for linguistic characteristics of the items (R^2 change from step 1 to step 2: .089 for Grade 4, .203 for Grade 7, and .235 for Grade 10).

The standardized regression coefficients are presented in Table 2 for each grade level regression model to show the effects of linguistic features on item difficulty. Statistically significant ($p < .05$) coefficients are marked with an asterisk and indicate that that linguistic variable contributed statistically significant unique variance to the prediction of item difficulties. The Type I error rate was maintained at .05 for each set of comparisons through a modified Bonferroni technique. Statistically significant differences between the effects of item features at the different grade levels are shown with subscripts.

The adjusted R^2 values for the main effects of the entire set of linguistic features for each grade level are shown in the last row of Table 2. In terms of the sizes of these effects, small R^2 values start at about .01, medium effects are shown by R^2 values around .09, and large effects are evident in R^2 values of about .25 and up (Cohen, 1988). R^2 values can be interpreted as the amount of variance in the dependent variable, item difficulty, explained by the independent variables, the language characteristics. These R^2 values show a medium-to-large effect of the set of linguistic features on mathematics item difficulty at Grade 4, dropping to a small-to-medium effect at Grade 10.

At Grade 4, five individual language elements showed unique and statistically significant effects on item difficulty: prepositions, ambiguous words, complex verbs (verbs with three or more words), pronouns, and math vocabulary. Note that these relationships are negative; the greater the number of linguistic elements, the

TABLE 2
Effects of Linguistic Features at Each Grade Level

	Standardized Regression Coefficients		
Linguistic Features	*Grade 4 Item Means*	*Grade 7 Item Means*	*Grade 10 Item Means*
Total words	–.047	.152	–.111
Words > six letters	–.018	.113	.124
Sentences	.048	–.010	.066
Prepositions	–.146*$_{b}$	–.064	.073$_{b}$
Relative pronouns	.031	–.048	.034
Ambiguous words	–.194*$_{b}$	–.099	–.066$_{b}$
Homophones	–.001	.032	.054
Passive voice	.144	.004	–.020
Clauses	–.021	.081	–.080
Complex verbs	–.090*$_{b}$	–.036$_{c}$	.147*$_{b,c}$
Infinitives	.090	–.096	–.042
Pronouns	–.148*$_{a,b}$	.067$_{a}$	.016$_{b}$
Math vocabulary	–.184*	–.178*	–.164*
Conditionals	.058	.037	.052
Comparatives	–.009	–.199*	–.093
Cultural references	.049	–.032	.030
Total adjusted R^2	0.134*	0.070*	.041*

Subscript a = statistically significant difference between Grades 4 and 7.
Subscript b = statistically significant difference between Grades 4 and 10.
Subscript c = statistically significant difference between Grades 7 and 10.
* $p < .05$.

lower the mean item difficulty, that is, the more difficult the item. Two features that showed apparently significant positive effects on item means, passive voice and infinitive verb forms, actually had no meaningful zero-order correlation with item means. Their regression coefficients are the result of suppression, an artifact in the analyses resulting from shared variance with other predictors in the regression equation (see Pedhazur, 1997) and thus can be ignored. At Grade 7, the number of language features impacting item mean scores diminished to two: math vocabulary and comparative terms. Ambiguous words and infinitive verb forms were not correlated with item means, and their apparent effect on item difficulty is due to suppression and can be ignored. At Grade 10, math vocabulary alone had a unique negative effect on item difficulty, whereas complex verbs were marginally related to easier items. Comparative forms almost reached the criterion for statistical significance as a negative influence at Grade 10 (p = .051).

Table 3 shows the means, number of items, and standard deviations for the individual linguistic features at each grade level. Length features, such as the mean number of sentences and clauses in each test item, tended to increase with grade level as would be expected. Some syntactic features were also more evident at the

TABLE 3
Frequency of Linguistic Features in Items at Each Grade Level

	Grade								
	4			*7*			*10*		
Linguistic Features	*M*	*N*	*SD*	*M*	*N*	*SD*	*M*	*N*	*SD*
Total words	39.32	208	27.415	39.47	203	23.382	56.97	183	31.534
Words > six letters	4.13	208	3.068	4.53	203	3.275	7.11	183	4.019
Sentences	3.03	208	2.083	3.42	203	1.959	4.34	183	2.387
Prepositions	3.23	208	2.798	5.05	203	3.784	7.76	183	5.261
Relative pronouns	.55	208	.604	.61	203	.719	.75	183	.791
Ambiguous words	1.50	208	1.255	1.21	203	1.048	2.11	183	1.330
Homophones	2.07	208	1.287	.86	203	.939	1.32	183	1.114
Passive voice	.15	208	.452	.21	203	.569	.26	183	.675
Clauses	3.28	208	2.383	4.50	203	2.407	5.85	183	3.190
Complex verbs	.20	208	.489	.24	203	.520	.52	183	1.053
Infinitives	.30	208	.587	.51	203	.886	.73	183	1.143
Pronouns	.90	208	1.448	1.09	203	1.559	.82	183	1.432
Difficult math vocabulary	.71	208	1.042	1.29	203	1.250	1.60	183	1.134
Conditionals	.15	208	.370	.14	203	.346	.20	183	.416
Comparatives	.28	208	.762	.49	203	1.276	.35	183	1.157
Cultural references	.40	208	.688	.43	203	.682	.85	183	1.035

higher grades, like the number of relative pronouns and prepositions. However, other item features were probably a result of content that was selected for assessment at that grade level, such as the number of comparative terms, which had its highest value at Grade 7. Some linguistic elements had fairly even and low representation throughout the grades, including conditional words and passive voice.

DISCUSSION

In a number of ways, this research has overcome some of the weaknesses of earlier studies on the effects of language on mathematics test items. This comprehensive evaluation of mathematics test items encompassed all 594 test items at three grade levels of a major state assessment. Large samples or entire populations of students, in the case of ELLs, were used to compute item difficulties. The mathematics standards used in the 2000 assessment had received outstanding external review regarding their quality and comprehensiveness (Finn & Petrilli, 2000). No test items were changed or modified. All of these analyses were conducted on actual mathematics assessment items selected for the general state assessment after rigorous item development, field testing, and review procedures. These test items represent the types of items similar to those used in other large-scale assessments designed for general populations.

In answer to the first research question, the linguistic features of mathematics test items measured in this study, as a set, have a meaningful impact on student performance with a moderate-to-large effect at Grade 4, a medium effect at Grade 7, and a smaller effect at Grade 10. In response to the second research question, however, this research found no disproportionate impact on potentially vulnerable student groups, ELL and SWD, as a result of these linguistic test item elements, either as a set or individually. When the interaction effects of group with language features were conducted, the overall change in R^2 was nonsignificant, and none of the individual interaction terms for the language elements had a unique effect on item difficulty. The group main effect was all that distinguished the performance of these three student groups from one another, not a differential response to test item language.

When investigating which language features have the greatest impact on student performance, several specific linguistic features exhibited unique, independent effects. The characteristic of difficult mathematics vocabulary shows a consistent effect for all student groups at all grade levels, confirming the effect found by Hanson et al. (1998). This is a content-relevant item feature that would be expected to relate to item difficulty. As such, it is probably better characterized as a mathematics item feature than a strictly language feature. The use of ambiguous or multiple-meaning words has statistically significant effects at Grade 4, also a confirmation of earlier research (Cummins et al., 1988; De Corte et al., 1985). 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. The use of problems requiring comparative terms also has a statistically significant impact on student performance at Grade 7 where these features are more common, and an almost-significant effect at Grade 10, substantiating earlier studies that showed comparison problems to be more difficult for students (Cummins et al., 1988; De Corte et al., 1985; Lean et al., 1990). In fact, the Grade 7 standards identified for assessment include specific reference to inequalities, leading to a greater number of items with comparative wording. This feature is clearly a marker for relevant mathematics content and should be considered a mathematical as well as a linguistic item feature.

An argument can be made that the inclusion of math vocabulary and comparatives as linguistic features confounds the results as these characteristics overlap with content knowledge in affecting the difficulty level of items. To explore the impact of these two features, the data were reanalyzed excluding them as variables in the linguistic set. 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. The reason that the number of words had not been statistically significant in the original analysis is that it is correlated with math vocabulary and comparatives and shares most of their predictable variance, thus not contributing unique variance in the prediction of item difficulties. A larger number of words in the item was very modestly correlated with easier items. The R^2 change at each grade level was smaller than that with all features and was not significant at Grade 10: .110 at Grade 4 ($p < .000$), .024 at Grade 7 ($p = .012$), and .014 at Grade 10 ($p = .082$). These results emphasize the relatively small effect that language *per se* had on item difficulty in this item set, particularly at secondary levels. However, it also shows that no additional language features surface to increase item difficulty when the math-relevant features are removed from the regression model.

The limitation that analyses of the test items from different grade levels could not be combined in this study deserves comment. Test items were written to assess developmentally appropriate standards for the mathematical content of each grade. There are several reasons why test item content might be confounded with other, perhaps irrelevant, grade and age characteristics. Children differ widely in their cognitive and language development at the three grade levels studied. The assumptions made by content standard experts and test item writers about background knowledge and related academic achievement, such as reading ability, are vastly different from elementary through high school ages. Further, content can be quite dissimilar from grade to grade even within one subject area, and mathematics is a huge subject area. Finally, the standards selected for assessment represent a small proportion of academic content at each grade, and content experts selecting the standards for assessment would be more likely to have chosen new and different content for assessment at the higher grades rather than repeating previously assessed standards. There is no reason, therefore, to expect much similarity in test item content at such diverse grade levels.

That said, fourth graders were more influenced by test item language in general, as seen in the larger adjusted R^2 value of the language main effect and the greater number of features that show unique predictive powers. It might be that fourth graders, with their less sophisticated verbal skills, are simply more sensitive to complex language in word problems than older students, a finding that would echo the concerns of Solano-Flores et al. (2003). Pronouns might be expected to cause confusion for less skilled linguists because they introduce a (possibly ambiguous) reference to another sentence element, whereas prepositions mark the existence of an additional phrase in the sentence and hence another concept to be understood. Complex verbs were defined as verbs with at least three words ("had been going," "would have eaten"), which suggests the use of multiple or difficult verb tenses. All of these conclusions have policy implications for test item design for younger students. All of the elementary students, not just ELL and SWD but also general education students, were more affected than older students by language that was irrelevant to the mathematics constructs being measured. Test developers and item writers should pay greater attention to the general language development of the students being tested and use wording that does not introduce additional comprehension hurdles over and above the required content. This suggestion is consistent with universal test design principles offered by the National Center on Educational Outcomes (Thompson, Johnstone, & Thurlow, 2002).

Conversely, the 10th graders' item means were negatively influenced only by difficult mathematics vocabulary, even given the greater frequencies of many of the linguistic characteristics in the Grade 10 test items. It is also likely that language features that had effects only at one grade level were a function of the wording of the mathematics instructional indicators chosen for these assessments and the items contained in this sample. The impact of math vocabulary, which affected items at all three grade levels, represents a more robust result due to the differences in content tested at different grades.

The primary restriction to generalization of these results is the fact that the test items used for these analyses were developed to measure mathematics instructional standards at three grade levels in one state. Though the mathematics content standards themselves are comprehensive, the topics selected for assessment measure only a fraction of the overall content and thus may present a threat to validity in terms of content underrepresentation. The material deemed essential at other grade levels would likely include different content and thus different item characteristics and attributes. Although these results are based on a large number of test items at elementary, middle school, and high school levels, these findings cannot be generalized for test items in other content areas or for different grade levels. Additional analyses using another selection of test item content and other grade levels would be necessary to confirm these results.

The lack of differential impact on vulnerable groups may be a result of the rigorous test and item development process undertaken by this test contractor, which is standard for major large-scale assessments. This process includes diversity on item writing and reviewing teams at each stage of item selection and tryout, bias review panels selected to represent multiple perspectives and groups, and statistical analysis of differential item functioning. Items and tests were analyzed with classical and item response theory methodologies. Though the findings from these specific content standards, test items, and student groups are not generalizable, the outcomes of this test development process may be more widely applicable in that most weak, poorly written, confusing, and unfair items are weeded out during high-quality test development. Therefore, one might hypothesize that other, similarly produced large-scale assessments would produce comparable results for different student groups. This does not mean that different student groups perform equally well, however, which is another assessment topic that demands continued research.

One intriguing avenue of future research was suggested in a study of science test items by Sireci and his colleagues (O'Neil, Sireci, & Huff, 2003–2004), who used expert review panels to evaluate the cognitive characteristics and content of science test items from two test administrations. This method could be used to evaluate content consistency of original and simplified math items, as well as adding an expert rating of cognitive difficulty for each item. In the current study, difficulty was measured only by percentage of students passing the item. Additional measures of difficulty or complexity would be useful.

Another idea would be to create pairs of items with identical computational demands, with one of each pair written as a word problem and the other in a calculation-ready format. Word problems can be difficult for students even when the computation required is below grade level (Larsen et al., 1978; Wheeler & McNutt, 1983), perhaps partially due to the fact that some of the reasoning required to set up the calculation of the problem has already been accomplished. A comparison of the magnitude of increased difficulty of the word format over the calculation format for different student groups would provide insight into the relative effects of language on math item comprehension for each group. However, this test would not allow for comparison of noncomputation problems and concepts, such as interpretation of charts or graphs, estimation, or geometric reasoning.

Future research could also focus on the relationship of mathematics achievement, reading achievement, and language proficiency to performance within student groups, which was not controlled or assessed in this study. For example, the effects of item language attributes on students at different levels of mathematics or reading achievement could be compared, under the hypothesis that language features are more influential and have a greater negative effect for lower achieving students within the general student population. Studies could be designed in which students are grouped on the basis of their English language proficiency and verbal skills, and then the effects of test item features could be evaluated regarding those differences rather than the potentially arbitrary distinctions of disability, language, or cultural difference. Depending on their previous educational experience and their exposure to English, ELL students display a huge range of general language skills and of English proficiency. Additionally, some native English-speaking students receiving special education services may have language delays that impact their performance on math word problems. Once a certain threshold of language proficiency is acquired, language features of test items may no longer pose a barrier to performance, whereas below that threshold differential effects might be found. Defining this threshold would be of interest for large-scale assessment programs such as those mandated by No Child Left Behind.

ACKNOWLEDGMENTS

This study was supported by funding provided by the Kansas State Department of Education and the National Center on Educational Outcomes. Opinions expressed herein are those of the authors and do not necessarily reflect those of the sponsoring agency.

REFERENCES

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33,* 4–14.

Abedi, J., Courtney, M., & Leon, S. (2003). *Research-supported accommodation for English language learners in NAEP* (CSE Tech. Rep. No. 586). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (CSE Tech. Rep. No. 536). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Brown, P. J. (1999). *Findings of the 1999 plain language field test*. Newark: University of Delaware, Delaware Education Research & Development Center.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20,* 405–438.

De Corte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77,* 460–470.

Finn, C. E., & Petrilli, M. J. (Eds.). (2000). *The state of state standards 2000.* Retrieved January 6, 2005, from the Thomas B. Fordham Foundation Web site: http://www.edexcellence.net/doc/Standards2000.pdf

Glasnapp, D., Poggio, J., & Omar, M. (2000). *Technical report, Y2002 Kansas assessments in mathematics, reading, and writing.* Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Hanson, M. R., Hayes, J. R., Schriver, K., LeMahieu, P. G., & Brown, P. J. (1998, April). *A plain language approach to the revision of test items.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Johnson, E., & Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assessment for Effective Intervention, 29,* 35–45.

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. (1999). *Ensuring accuracy in testing for English language learners: A practical guide for assessment development*. Washington, DC: Council of Chief State School Officers.

Larsen, S. C., Parker, R. M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. *Learning Disabilities Quarterly, 1,* 80–85.

Lean, G. A., Clements, M. A., & Del Campo, G. (1990). Linguistic and pedagogical factors affecting children's understanding of arithmetic word problems: A comparative study. *Educational Studies in Mathematics, 21,* 165–191.

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics, 21,* 83–90.

Lord, C., Abedi, J., & Poosuthasee, N. (2000). *Language difficulty and assessment accommodations for English language learners.* Dover: Delaware Department of Education. Retrieved February 1, 2006, from the Delaware Department of Education Web site: http://www.doe.state.de.us/aab/DSTP_research.html

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

O'Neil, T., Sireci, S. G., & Huff, K. L. (2003–2004). Evaluating the consistency of test content across two successive administrations of a state-mandated science assessment. *Educational Assessment, 9,* 129–151.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71,* 53–104.

Rivera, C., & Stansfield, C. (2001, April). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. R., & Poggio, J. P. (2002). *The differential impact of accommodations in statewide assessment.* Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Research summary retrieved January 30, 2005, from the National Center on Educational Outcomes Web site: http://education.umn.edu/NCEO/TopicAreas/Accommodations/Kansas.htm

Solano-Flores, G., Trumbull, E., & Kwon, M. (2003, April). *The metrics of linguistic complexity and the metrics of student performance in the testing of English-language learners.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston: Allyn and Bacon.

Tindal, G., Anderson, L., Helwig, R., Miller, S., & Glasgow, A. (2000). *Accommodating students with learning disabilities on math tests using language simplification.* Eugene, OR: University of Oregon Research, Consultation, and Teaching Program. Retrieved January 30, 2005, from the Behavioral Research and Teaching Web site at the University of Oregon: http://brt.uoregon.edu/files/Accom_lang_simplfy.pdf

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Rep. 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved January 10, 2006, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Wheeler , L. J., & McNutt, G. (1983). The effects of syntax on low-achieving students' abilities to solve mathematics word problems. *Journal of Special Education, 17,* 309–315.

Yang, X., Shaftel, J., Glasnapp, D., & Poggio, J. (2005). Qualitative or quantitative differences? Latent class analysis of mathematical ability for special education students. *Journal of Special Education, 38,* 194–207.

APPENDIX
Linguistic Complexity Checklist

Grade _______ Form _______ Part _______ Item_______

Count the instances of each of these in the problem.

1. _____ total number of words in the item
2. _____ number of different words with 7 letters or more. List: ________________________________
   ____
3. _____ number of sentences
4. _____ number of prepositional phrases (beginning with *from, at, by, in, out, after, among, following,* etc.)
5. _____ number of relative pronouns (*that, who, whom, whose, which*)
6. _____ number of slang, idiomatic, ambiguous, or multiple-meaning words or phrases (*feet, change, set, cool, grab-bag,* etc.) List: ____
7. _____ number of homophones or near homophones (*price/prize, their/there/they're, too/two/to, add/ad, buy/by/bye,* etc.) List: ____
8. _____ number of uses of passive voice (*were sold, was paid, had been computed*, etc.)
9. _____ number of clauses (with subject and predicate: dependent, independent, adverbial, relative, etc.)
10. _____ number of complex verb forms of 3 words or more (*would have been, will have done,* etc.*)*
11. _____ number of infinitive verb phrases (*to drive, to make, to follow,* etc.)
12. _____ number of pronouns (*she, her, hers, he, him, his, it, they, theirs*, etc.)
13. _____ number of unusual or difficult but specific mathematics vocabulary words (*likelihood, probability, perimeter, pentagon, reflection, symmetry, quotient, equation, complementary, coordinate,* etc.) List: ____
14. _____ number of conditional constructions (*if-then, if-would, if-could, if-will,*etc.)
15. _____ number of comparative constructions (*more than, fewer than, less than, greater than,* etc.)
16. _____ number of references to American holidays (*Labor Day, Memorial Day, July 4th, Thanksgiving, Halloween*, etc.) List: ____
17. _____ number of references to specific American cultural events or situations (*picnic, camping, dormitory*, etc.) List: ____

Are there any other features of the item that you think are difficult that are not captured or measured above? What are they?