APPLIED MEASUREMENT IN EDUCATION, *14*(3), 219–234

The Language Factor in Mathematics Tests

Jamal Abedi
University of California, Los Angeles/National Center for Research on Evaluation, Standards, and Student Testing

Carol Lord
California State University, Long Beach

In this study we investigated the importance of language in student test performance on mathematics word problems. Students were given released items from the National Assessment of Educational Progress mathematics assessment, along with parallel items that were modified to reduce their linguistic complexity. In interviews, students typically preferred the revised items over the original counterparts. Paper-and-pencil tests containing original and revised items were administered to 1,174 8th grade students. Students who were English language learners (ELLs) scored lower on the math test than proficient speakers of English. There were also differences in math performance with respect to socioeconomic status (SES) but not gender. Linguistic modification of test items resulted in significant differences in math performance; scores on the linguistically modified version were slightly higher. Some student groups benefited more from the linguistic modification of items—in particular, students in low-level and average math classes, but also ELLs and low SES students.

Research has drawn attention to the importance of language in student performance on assessments in content-based areas such as mathematics (see, for example, Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, & Plummer, 1995; Aiken, 1971; Aiken, 1972; Cocking & Chipman, 1988; De Corte, Verschaffel, & DeWin, 1985; Garcia, 1991; Jerman & Rees, 1972; Kintsch & Greeno, 1985; Larsen, Parker, & Trenholme, 1978; Lepik, 1990; Mestre, 1988; Munro, 1979; Noonan, 1990; Orr, 1987; Rothman & Cohen, 1989; Spanos, Rhodes, Dale, & Crandall, 1988). Nationally, children perform 10% to 30% worse on arithmetic word problems than on comparable problems presented in numeric format (Carpenter,

Requests for reprints should be sent to Jamal Abedi, UCLA–CSE/CRESST, 300 Charles E. Young Drive North, GSE & IS Building, Third Floor, Los Angeles, CA 90095–1522.

Corbitt, Kepner, Linquist, & Reys, 1980). The discrepancy between performance on verbal and numeric format problems strongly suggests that factors other than mathematical skill contribute to success in solving word problems (August & Hakuta, 1997; Cummins, Kintsch, Reusser, & Weimer, 1988; LaCelle-Peterson & Rivera, 1994; Zehler, Hopstock, Fleischman, & Greniuk, 1994).

English language learner (ELL) students score lower than students who are proficient in English on standardized tests of mathematics achievement in elementary school as well as on the Scholastic Aptitude Test and the quantitative and analytical sections of the Graduate Record Examination. Although there is no evidence to suggest that the basic abilities of ELL students are different from non-ELL students, the achievement differences between ELL and non-ELL students are pronounced (Cocking & Chipman, 1988; Mestre, 1988).

In this study we compared the performance of ELLs and proficient speakers of English on math word problems from the National Assessment of Educational Progress (NAEP) tests and investigated whether modifying the linguistic structures in the test items affected student test performance.

Two separate field studies were conducted. In the first investigation, the Student Perceptions Study, 36 eighth-grade students were interviewed. These students were given original NAEP math items and parallel revised items (with simplified language) in a structured interview format to investigate the students' perceptions and preferences.

In the second study, the Accuracy Test Study, 1,174 eighth-grade students took paper-and-pencil math tests including 10 original NAEP math items, 10 items with linguistic modifications, and five noncomplex control items. Students' scores on the original and linguistically modified items were compared.

MODIFICATION OF MATH ITEMS

The corpus of math items used for this investigation was the 69 released items from the 1992 NAEP main math assessment. From the set of linguistic features appearing in these items, several features were identified as potentially problematic for ELL students. Judgments were based on expert knowledge and on findings of previous empirical studies (including, among others, Adams, 1990; Bever, 1970; Biber, 1988; Botel & Granowsky, 1974; Bormuth, 1966; Celce-Murcia & Larsen-Freeman, 1983; Gathercole & Baddeley, 1993; Chall, Jacobs, & Baldwin, 1990; Forster & Olbrei, 1973; Hunt, 1965, 1977; Jones, 1982; Just & Carpenter, 1980; Kane, 1968; Klare, 1974; Lemke, 1986; MacDonald, 1993; MacGinitie & Tretiak, 1971; Paul, Nibbelink, & Hoover, 1986; Pauley & Syder, 1983; Perera, 1980; Slobin, 1968; Wang, 1970).

For those items with language that might be difficult for students, simpler versions were drafted, keeping the math task the same but modifying nonmath vocab-

ulary and linguistic structures; math terminology was not changed. (Math experts checked original and modified versions to ensure that the math content was parallel.) Problematic features were removed or recast. For a given math item, more than one feature might be revised. Linguistic features that were modified included the following (see Abedi et al., 1995, for further discussion):

- Familiarity or frequency of nonmath vocabulary—unfamiliar or infrequent words were changed (a certain reference file > Mack's company).
- Voice of verb phrase—passive verb forms were changed to active (if a marble is taken from the bag > if you take a marble from the bag).
- Length of nominals—long nominals were shortened (the pattern of the puppy's weight gain > the pattern above).
- Conditional clauses—conditionals were replaced with separate sentences, or the order of conditional and main clause was changed (if two batteries in the sample were found to be dead > he found three broken skateboards in the sample).
- Relative clauses—removed or recast (the total number of newspapers that Lee delivers in 5 days > how many newspapers does he deliver in 5 days).
- Question phrases—complex question phrases were changed to simple question words (which is the best approximation of the number > approximately how many).
- Abstract or impersonal presentations—made more concrete (… 2,675 radios sold > … 2,675 radios that Mrs. Jones sold).

Some changes involved more than one feature; in the fourth example in the aforementioned list, the revised version no longer contains the conditional clause or the passive voice verb, and more frequent or familiar vocabulary has been used.

STUDENT PERCEPTIONS STUDY

In the first study, the following questions were investigated in interviews with eighth-grade students. For mathematics items with parallel mathematics content, do students respond differently to items that contain different linguistic structures? Do students find linguistically simpler items easier to comprehend? Do they show a preference for items with simpler language?

Procedure

A total of 36 students at four school sites in the Los Angeles area were interviewed. The students represented a cross section of ethnic and language backgrounds; their

native languages, in addition to English, included Spanish, Cambodian, and Vietnamese. Their current grades in math class ranged from A to D.

Each recorded interview lasted 10 to 15 min. After a brief introductory conversation, the student was asked to read a pair of math items—the original item and the corresponding revised item—and was asked the following questions: "If you were really in a hurry on a test and you had to pick one of these problems to do, which one would you do? Read it aloud to me. Now read the other one aloud to me. Are there words in either of them that might be confusing for some students or hard for them to understand? What is it about the one you chose that seems easier?" Each student responded to four pairs of items.

Results

In the first set of interviews, 19 students from two schools participated. As the data in Table 1 indicate, a majority of these students picked the revised version for items 1 and 2. On average, the original items were selected 37% of the time (P_{orig} or = .37), and the revised items were selected 63% of the time (P_{rev} = .63). A z statistic comparing P_{orig} with P_{rev} was 2.18, which is significant at the .05 nominal level. Students significantly preferred the revised items over the original items.

In the second set of interviews, a different set of four pairs of items was presented to 17 students. As Table 2 shows, most of the students chose the revised version for all pairs; 16.9% of the students preferred the original items (P_{orig} = .17), and 83.1% preferred the revised items (P_{rev} = .83). A z statistic comparing P_{orig} with P_{rev} was 5.47, significant beyond the .01 nominal level; in general, students preferred the revised items.

Many students voiced a global judgment that the language in the revised item was easier to comprehend; comments included, "Well, it makes more sense," and "It seems simpler; you get a clear idea of what they want you to do." Some students made specific reference to time pressure as a factor in taking tests; some commented on the length of the items. Responses included, "It's easier to read, and it gets to the point, so you won't have to waste time," and "'Cause it's, like, a little bit less writing." Some students commented on the difficulty of vocabulary items.

TABLE 1
Student Perceptions Study: First Set

Item No.	*Original Item Chosen*	*Revised Item Chosen*
1	3	16
2	4	15
3	10	9
4	11	8

z = 2.18. $p < .05$.

TABLE 2
Student Perceptions Study: Second Set

Item No.	*Original Item Chosen*	*Revised Item Chosen*
5	3	14
6	4.5[a]	12.5
7	2	15
8	2	15

[a]One student was ambivalent about his choice.
$z = 5.47$. $p < .01$.

They indicated that the vocabulary in the revised items was more familiar to them, as in the following comments: "This one uses words like 'sector' and 'approximation,' and this one uses words that I can relate to," and "Because it's shorter and doesn't have, like, complicated words."

In some instances, students chose the original item. One student said the original item was more interesting. Another said the original item was more challenging. A student said that the two items were very much the same, so he picked the one he had read first (in this case, the original item).

In addition to explicit student comments about the items, further insight about difficulty in comprehending vocabulary and syntax was gained from having students read both versions of each item aloud. When a student is reading, pauses for unfamiliar words or constructions are likely to disrupt the flow of comprehension (Adams, 1990). Some students stumbled on words such as "certain," "reference," "entire," and "closet." In reading aloud an original item containing a passive verb construction, one student substituted an active verb form (the item contained the verb phrase "would be expected," but the student read it aloud as "would you expect to find"), replacing a less familiar construction with a more familiar one. The student read the revised version as it was written.

In general, the student responses showed clear differences between the original and the revised item in each pair. Student preferences for the revised items gave support to the notion that the math items could be linguistically simplified in meaningful ways for the test taker. The interview results supported the plan to test a larger group of students to determine whether the observed differences in student responses to the language of the math items would be reflected as actual differences in math test scores.

ACCURACY TEST STUDY

The purpose of the second field study, the Accuracy Test Study, was to examine the impact of revision of selected linguistic features in NAEP math test items on the

number of test items answered correctly by students. The level of linguistic complexity was experimentally manipulated (items were linguistically simplified) and revised versions of 20 test items were created. Original and revised items in a paper-and-pencil format were presented to students.

Participants

For this study, 1,174 eighth-grade students from 39 classes in 11 schools from the greater Los Angeles area were selected to provide a range of language, socioeconomic, and ethnic backgrounds.

Information was obtained from school personnel on students' English proficiency classification, language background, grade level, type of math class, grades in math class, gender, ethnicity, and socioeconomic status (SES). In Los Angeles schools, a Home Language Survey (HLS) is administered to determine if a language other than English is spoken in the home. Based on the HLS response, English language assessment tests are administered, leading to a classification of the student's English proficiency. These classifications were obtained, where available, for students in the study. The results indicated that approximately 31% of the students were assigned to ELL categories ranging from Initially Fluent in English (4.8%) to Redesignated Fluent (8.7%) to Limited English Proficient (9.2%), and to other categories of English as a Second Language (ESL) (8.3%). Most students were eighth graders (95%); 5% were in Grade 7 or 9. Types of math classes included honors algebra, algebra, high mathematics, average mathematics, low mathematics, and ESL mathematics (including bilingual and sheltered English classrooms). The student group was 54% boys and 46% girls.

Data on student ethnicity classifications were obtained from the schools: 35% were Latino, 26% were White, 19% were African American, 16% were Asian American, and 4% were other or declined to state. Estimating from the limited data available, roughly 36% of the students were categorized as low SES on the basis of participation in free school lunch programs or in Aid to Families with Dependent Children programs. In addition to English, students spoke Spanish, Korean, Chinese, Farsi, and Filipino (as reported most frequently on the Language Background Questionnaire, LBQ).

Instruments

Each test booklet contained a math test and a two-page LBQ, which included items from the NAEP background questionnaires and the National Education Longitudinal Study (88; Ingels, Scott, Lindmark, Frankel, & Myers, 1992) background questionnaire, as well as new items generated for this study.

For the math test, 20 items were selected from the 69 released eighth-grade NAEP items. These items were those judged most likely to impede the student's performance on a test because of language that could be misunderstood, could con-

fuse the student, or could present difficulties that might interfere with the student's focus on the math content of the item. A simplified version of each of the items was written. The language was simplified, but the quantities, numerals, and visuals were retained from the original, so that the math content of the revised items paralleled that of the original items.

To ensure that the mathematical content of both versions of each item was equivalent, two experts in mathematics education independently reviewed each pair of items. They were asked to determine whether the two items differed in mathematical content or were equivalent with respect to the mathematical concepts being assessed. One math expert found no differences between the original and revised items in mathematical content; the other math expert pointed out three instances in which the situation in the revised item might be construed as slightly different. Changes were made in those three items to ensure that the math content in each pair was parallel.

Two different forms of the mathematics test were created. Booklet A contained 10 original items; the revised versions of these items were placed in Booklet B. Ten additional original items were placed in Booklet B, and the revised versions of these were placed in Booklet A. Thus, each form contained 10 original and 10 revised items. In addition, from the 69 original NAEP items, 5 items were selected in which the language was judged to have the least potential for misunderstanding or confusion. Thus, each test booklet contained a total of 25 math items.

In an effort to make Booklets A and B as similar as possible, original test items were assigned to Booklets A and B according to four criteria: type and number of linguistic complexities, presence or absence of a diagram or other visual aid, mathematical classification of the item content according to NAEP categories, and difficulty of the item. The measure of item difficulty used was the item difficulty index (*p* value) of each item from an earlier NAEP administration for eighth-grade students (1992 main assessment in math).

Procedure

Tests were administered by a team of 10 retired teachers and principals experienced in test administration. Test administrators attended a half-day training session, and testing sites were monitored by members of the project staff. Within each classroom, test booklets were distributed with Booklet A and Booklet B alternating; 51% received Booklet A, and 49% received Booklet B. Students were given approximately 1 hr to complete the test.

Results

The main research questions in this study were

- Are there significant differences in the math performance of English language learners and proficient speakers of English?
- Does modifying the linguistic structures in math test items affect students' test performance?

An additional research question was

- Do student background variables such as gender and family SES impact students' math test performance?

Student math scores. Table 3 presents descriptive statistics (mean; *SD;* and number of participants, *n*) for the total sample and for subgroups of students. As Table 3 shows, the mean score on all 25 items for the entire group was 14.01, with *SD* of 6.78. Proficient English speakers showed a substantially higher mean score (M = 15.14, SD = 6.64) than ELLs (M = 11.56, SD = 6.46). Large performance differences were found between students at different categories of SES and type of math class. Students in the low SES group had a lower

TABLE 3
Math Test Scores on 25 Items

Student Groups	*N*	*M*	*SD*
Total	1174	14.01	6.78
ELL classification			
English learners (ELL)	372	11.56	6.46
Proficient English speakers (non-ELL)	802	15.14	6.64
SES (Free lunch or AFDC)			
Low	449	12.47	6.55
High	725	14.96	6.76
Gender			
Male	647	14.20	6.12
Female	527	13.92	6.22
Math test booklet			
Form A	601	13.90	6.93
Form B	573	14.12	6.64
Type of math class			
ESL math	167	5.21	3.90
Low math	53	9.74	4.67
Average math	405	13.14	4.98
High math	249	16.28	5.85
Algebra	178	17.31	5.95
Honors algebra	122	21.33	3.64

Note. ELL = English language learner; SES = socioeconomic status; AFDC = Aid to Families with Dependent Children; ESL = English as a Second Language.

mean score (M = 12.47, SD = 6.55) than students in the high SES category (M = 14.96, SD = 6.76).

As one might expect, the largest difference in math performance was found between students in different math classes. Students in higher level math classes received higher scores; the means of the two booklets ranged from 5.21 (SD = 3.90) for the ESL math classes to 21.33 (SD = 3.64) for the honors algebra classes—a difference of over three standard deviations. Student gender did not have much impact on math performance; mean scores for boys and girls were similar. For boys, the mean was 14.20 (SD = 6.12), and for girls the mean was 13.92 (SD = 6.22).

To examine the relation of students' SES and their ELL classification, joint distributions of SES and ELL categories were obtained. Table 4 shows frequencies and percentages of ELL and non-ELL students at different categories of SES.

As shown in Table 4, the percentage of ELLs in the high SES category is 40.6%, compared with 71.6% of non-ELLs in the high SES category. The chi-square statistic for this table was 103.26 with 1 degree of freedom, which is significant above the .01 nominal level, suggesting that a significant confounding occurs between students' ELL status and their family SES. Cramér's V was .324 for this table, showing a moderate relation between ELL status and SES.

To test the hypothesis of a difference or lack of a difference in performance between ELLs and proficient English speakers at different SES levels, a two-factor analysis of variance (ANOVA) model was applied to the data. The independent variables were students' ELL status and family SES. The dependent variable in this model was student scores on all 25 test items (10 original, 10 revised, and 5 control items).

The mean math score for English language learners (M = 11.56, SD = 6.46) was significantly lower than the mean score for proficient English speakers (M = 15.14, SD = 6.64), and the null hypothesis of no difference between the performance of ELLs and proficient English speakers (factor A main effect) was rejected (F = 52.25, df = 1, 1170, p = 0.00). Factor A (ELL status) explained 4.1% of the variance of the dependent variable (η^2 = .041).

TABLE 4
Students' ELL Classification and Family SES

ELL Classification and SES	*Low SES*		*High SES*	
	Frequency	*% Total*	*Frequency*	*% Total*
ELL	221	59.5	151	40.6
Non-ELL	228	28.4	574	71.6

Note. ELL = English language learner; SES = socioeconomic status.
χ^2 = 103.26. df = 1. p = 0.000. Cramér's V = .297.

Students' family SES (Factor B) also had significant impact on their math performance. Students in the low SES category ($M = 12.47$, $SD = 6.55$) performed significantly lower than those in the high SES group ($M = 14.96$, $SD = 6.76$). Thus, the null hypothesis of equal performance of high or low SES students was rejected ($F = 15.61$, $df = 1, 1170$, $p = 0.00$; $\eta^2 = .012$).

The results of ANOVA also suggest a significant interaction between students' ELL status and their family SES ($F = 17.41$, $df = 1, 1170$, $p = 0.00$; $\eta^2 = .014$). This confirms that ELL and SES were confounded.

Comparing students' performance on linguistically modified versus original items. This study focused on two major questions. First, would the math test performance of ELLs and proficient English speakers be different? The results showed that the proficient English speakers achieved significantly higher math scores. The second question, then, is whether modifying the language of the items affects student performance. That is, to what extent is language modification effective in reducing the performance gap between ELLs and proficient English speakers? To address this second question, we compared the performance of students on the original and revised items.

One group of students answered original items, and another group answered revised items. To control for class, teacher, and school effects, we randomly assigned the two booklets (A and B) within a class. To ensure that the overall math performance of one group was not statistically different from that of the other group, we compared the performance of students who were given Booklet A with those who were given Booklet B. For this comparison, a two-factor ANOVA model was used with Booklet (Form A and B) and ELL classification (ELL and non-ELL) as the two independent variables, and the total score on all 25 math items as the dependent variable. We included ELL classification to examine a possible interaction between ELL status and test form.

Mean math score for Form A was 13.90 ($SD = 6.93$) and for Form B was 14.12 ($SD = 6.64$), with a difference of less than a quarter of a score point (see Table 3). This small difference between mean scores of students taking the two different forms was not significant ($F = 0.98$, $df = 1, 1170$, $p = .32$, $\eta^2 = .00$). The difference between ELLs and non-ELLs was significant ($F = 75.95$, $df = 1, 1170$, $p = .00$, $\eta^2 = .06$), but the interaction between ELL classification and Booklet was not significant ($F = 0.29$, $df = 1, 1170$, $p = .59$, $\eta^2 = .00$). These results suggested that the two groups of students who answered items from two different booklets were from the same population, and consequently the original and modified scores could be combined across the two booklets in comparing student scores.

Because the two groups (taking the different booklets) were randomly equivalent, two separate between-group analyses were conducted. The 10 original items on Booklet A had revised counterparts in Booklet B. The first analysis compared the performance of students who answered the 10 original items in Booklet A with

those students who answered the revised counterparts of those items in Booklet B. The mean revised items (M = 5.45, SD = 2.78) was slightly higher than the mean original items (M = 5.21, SD = 2.97). This difference, however, did not reach the significance level (t = 1.43, df = 1172, p = .153).

In the second analysis, we compared the math performance of students who took the original items in Booklet B with those who took the modified versions of those items in Booklet A. Students who took the modified version of the test performed slightly better (M = 5.79, SD = 2.88, n = 601) than those who took the original version of the test (M = 5.73, SD = 2.79, n = 573). However, although the trend of higher performance of students taking the revised items is consistent across the two booklets, the difference did not reach the significance level (t = .10, df = 1172, p = .760). As discussed later, when the original and revised items were compared for all students (across the two booklets), the difference became statistically significant due in part to the larger sample size.

Because the analyses revealed that the contents of the two math booklets were parallel, we created two composite scores for each subject: (a) the total original score (the sum of correct responses on the 10 original items), and (b) the total revised score (the sum of correct responses on the 10 revised items). Using these two sets of 10 items each, we found that mean student scores were greater for the revised items than for the original items in both cases—that is, the students did better on the revised versions. Mean score on the original items for the entire sample was 5.46 (SD = 2.89) and on the revised items was 5.62 (SD = 2.83). The difference between the means for original and revised items was .16, a relatively small difference but statistically significant (t = 2.95, df = 1173, p = .003). Thus, revising the items resulted in higher math scores overall.

The impact of linguistic modification on different subgroups. To investigate the possibility of differential performance of students on the original and revised items due to ELL classification, type of math class, and family SES, three different factorial ANOVA models were run. In the first model, item type (original and revised scores) was used as a within factor, and students' ELL classification was used as a between factor. The within factor main effect (original and revised scores) was significant (F = 6.41, df = 1, 1172, p = .012, η^2 = .006), indicating that students' mean score on the revised items (M = 5.62, SD = 2.83) was higher than their mean scores on the original items (M = 5.46, SD = 2.89). The between factor main effect (ELLs versus non-ELLs) was also significant (F = 81.89, df = 1, 1172, p = .000, η^2 = .069), indicating that non-ELLs (M = 6.10, SD = 2.75) performed better than ELLs (M = 4.58, SD = 2.73). However, the interaction between type of items and ELL classification was not significant (F = .35, df = 1, 1172, p = .556, η^2 = .00).

In the second model, we compared mean scores on original and revised items across SES categories. The within factor was item type (original or revised), as in the previous model, and the between factor was SES. The between factor main ef-

fect was significant (F = 38.36, df = 1, 1172, p = .000, η^2 = .033), indicating that higher SES students (M = 5.85, SD = 2.91) performed better than lower SES students (M = 4.85, SD = 2.75). As in the previous model, the interaction was not significant (F = .02, df = 1, 1172, p = .973, η^2 = .00).

The third model used item type as within and math classes as between factor. The main effect for the between factor was significant (F = 31.62, df = 1, 1172, p = .000), suggesting that students in different levels of math classes performed differently. As Table 3 shows, students enrolled in the lower level math classes performed significantly lower than students in the higher level classes.

These results indicate that simplifying the language of math test items helped students improve their performance. To further investigate whether linguistic revisions helped certain students more than others, a gain score in students' performance due to language modification of test items was computed. This gain score was defined as the total score on the modified test items minus the total score on the original items. We converted this gain score to a percentage of improvement by dividing the gain score by the mean score on original items. Table 5 presents the average gain score and percentage of improvement for different student groups.

As the data in Table 5 show, the magnitude of the gain score and the percentage of improvement of students' performance due to language modification of test

TABLE 5
Improvement of Performance on Modified Over Original Items on 10 Items

Student Groups	*N*	*Mean Gain*	*SD*	*% Improvement*
Total	1174	.156	1.81	2.9%
ELL classification				
English learners (ELL)	372	.165	1.86	3.7
Proficient English speakers (non-ELL)	802	.144	1.74	2.4
SES (Free lunch or AFDC)				
Low	449	.158	1.84	3.3
High	725	.154	1.79	2.6
Gender				
Male	647	.155	1.80	2.8
Female	527	.156	1.77	2.9
Type of math class				
Low math	53	.358	2.12	6.7
Average math	405	.351	2.02	6.6
ESL math	167	.090	1.40	0.9
High math	249	.028	1.92	0.4
Algebra	178	.051	1.63	0.7
Honors algebra	122	–.074	1.36	–0.8

Note. ELL = English language learner; SES = socioeconomic status; AFDC = Aid to Families with Dependent Children; ESL = English as a Second Language.

items differ across different student groups. The overall gain score for the entire sample is .156 (*SD* = 1.81), which translates to 2.9% improvement due to language modification. The percentage of improvement is slightly higher for ELLs (3.7%) than for non-ELLs (2.4%). Similarly, on the SES categories, the percentage of improvement for low SES students is slightly higher (3.3%) than the percentage for high SES students (2.6%). The percentage of improvement for boys (2.8%) is almost identical with the percentage for girls (2.9%).

The largest discrepancy in percentage of improvement was observed among students in different levels of math classes. Students in the lower level math classes benefited more from the language modification of test items. Students in low level math classes scored 6.7% higher on revised items, and those in average level math classes showed 6.6% improvement on the revised items. The trend did not continue for higher levels of math classes, however; in fact, for the honors algebra class the language simplifications had a small negative effect (–0.8%). Students in ESL math classes showed 0.9% improvement in their math performance on the revised items.

DISCUSSION

The results of this study clearly show the impact of students' language background on their performance on math word problems. First, the study found that English language learners scored significantly lower than proficient speakers of English. This is a cause for concern. Second, it appears that modifying the linguistic structures in math word problems can affect student performance. In interviews, students indicated preferences for items that were simpler linguistically. On paper-and-pencil tests, over a thousand students scored higher, on average, on linguistically modified items; the overall mean score difference was small but statistically significant.

In general, the language modifications had greater impact for low-performing students. In terms of the percentage of improvement of scores on modified over original items

- English language learners benefited more than proficient speakers of English.
- Low SES students benefited more than others.
- Students in low level and average math classes benefited more than those in high level math and algebra classes.

The differences observed here are consistent with previous research studies showing relations between reading ability and arithmetic problem-solving ability (Aiken, 1971, 1972; Larsen et al., 1978; Noonan, 1990). Because language ability is, in general, a predictor of math performance, it is possible that the lan-

guage 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 greater relative improvement in their scores.

The findings here are also consistent with the view that inexperienced problem solvers, lacking highly developed semantic schemata for problem solving, rely more on the text (De Corte et al., 1985); if this is indeed the case, we would expect that the complexity of the text would be a more significant factor for inexperienced and inexpert problem solvers. Our results support this view.

Although the portion of this study that dealt with the identification of complex language was largely exploratory in nature, it provided useful clues in the search for linguistic features that can negatively affect performance for certain groups of students. Data from this study were consistent with previous research suggesting that unfamiliar or infrequent vocabulary and passive voice constructions may affect comprehension for certain groups of students and that average and low-achieving students may be at a relatively greater disadvantage in answering mathematics items with complex language. These studies should be replicated and refined. It is also possible that future studies, with larger numbers of other targeted linguistic features such as those described in this study, will reveal similar effects. Meanwhile, it remains prudent to continue searching for interactions among linguistic, socioeconomic, and other background variables to shed light on the growing issue of the role of language in content area assessment.

Ultimately, this study shows that the interaction between language and mathematics achievement is real. This interaction must be a critical consideration in future mathematics assessment research and practice.

ACKNOWLEDGMENTS

The work reported here was supported in part under the National Center for Education Statistics, Contract RS90159001 as administered by the U.S. Department of Education, Office of Educational Research and Improvement. The findings and opinions expressed here do not reflect the position or policies of the National Center for Education Statistics, the Office of Educational Research and Improvement, or the U.S. Department of Education.

We acknowledge the valuable contribution of Joseph Plummer and Patricia Snyder to this study and the helpful comments of Frances Butler and Joan Herman.

REFERENCES

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance.* Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C. & Plummer, J. (1995). *Language Background as a Variable in NAEP Mathematics Performance: NAEP TRP Task 3D: Language Background Study.* Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Aiken, L. R. (1971). Verbal factors and mathematics learning: A review of research. *Journal for Research in Mathematics Education, 2*(4), 304–13.

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Education Research, 42,* 359–85.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda.* Washington, DC: National Academy Press.

Bever, T. (1970). "The cognitive basis for linguistic structure." In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–353). New York: Wiley.

Biber, D. (1988). *Variation across speech and writing.* New York: Cambridge University Press.

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly, 1,* 79–132.

Botel, M., & Granowsky, A. (1974). A formula for measuring syntactic complexity: A directional effort. *Elementary English, 1,* 513–516.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Jr., Linquist, M. M., & Reys, R. E. (1980). Solving verbal problems: Results and implications from national assessment. *Arithmetic Teacher, 28*(1), 8–12.

Celce-Murcia, M., & Larsen-Freeman, D. (1983). *The grammar book: An ESL/EFL teacher's book.* Rowley, MA: Newbury House.

Chall, J. S., Jacobs, V. S., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind.* Cambridge, MA: Harvard University Press.

Cocking, R. R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement of language minority children. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 17–46). 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., & DeWin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77,* 460–470.

Forster, K. I., & Olbrei, I. (1973). Semantic heuristics and syntactic trial. *Cognition, 2,* 319–347.

Garcia, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly, 26,* 371–391.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (Research Rep. No. 3). Urbana, IL: National Council of Teachers of English.

Hunt, K. W. (1977). Early blooming and late blooming syntactic structures. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging.* Urbana, IL: National Council of Teachers of English.

Ingels, S. J., Scott, L. A., Lindmark, J. T., Frankel, M. R., & Myers, S. L. (1992). *First follow up; student component data file user's manual* (Vol. 1). Washington, DC: National Center for Education Statistics.

Jerman, M., & Rees, R. (1972). Predicting the relative difficulty of verbal arithmetic problems. *Educational Studies in Mathematics, 13,* 269–287.

Jones, P. L. (1982). Learning mathematics in a second language: A problem with more and less. *Educational Studies in Mathematics, 13,* 269–87.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixation to comprehension. *Psychological Review, 87,* 329–354.

Kane, R. (1968). The readability of mathematical English. *Journal of Research in Science Teaching, 5,* 296–298.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92,* 109–129.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly, 10,* 62–102.

LaCelle-Peterson, M., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64,* 55–75.

Larsen, S. C., Parker, R. M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. *Educational Studies in Mathematics, 21,* 83–90.

Lemke, J. L. (1986). *Using language in classrooms.* Victoria, Australia: Deakin University Press.

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

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language. 32,* 692–715.

MacGinitie, W. H., & Tretiak, R. (1971). Sentence depth measures as predictors of reading difficulty. *Reading Research Quarterly, 6,* 364–377.

Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 201–220). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Munro, J. (1979). Language abilities and math performance. *Reading Teacher, 32*(8), 900–915.

Noonan, J. (1990). Readability problems presented by mathematics text. *Early Child Development and Care, 54,* 57–81.

Orr, E. W. (1987). *Twice as less: Black English and the performance of Black students in mathematics and science.* New York: Norton.

Paul, D. J., Nibbelink, W. H., & Hoover, H. D. (1986). The effects of adjusting readability on the difficulty of mathematics story problems. *Journal for Research in Mathematics Education, 17*(3), 163–171.

Pauley, A., & Syder, F. H. (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics, 7,* 551–579.

Perera, K. (1980). The assessment of linguistic difficulty in reading material. *Educational Review,, 32,* 151–161.

Rothman, R. W., & Cohen, J. (1989). The language of math needs to be taught. *Academic Therapy, 25,* 133–142.

Slobin, D. I. (1968). Recall of full and truncated passive sentences in connected discourse. *Journal of Verbal Learning and Verbal Behavior, 7,* 876–881.

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

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

Zehler, A. M., Hopstock, P. J., Fleischman, H. L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students.* Arlington, VA: Development Associates, Special Issues Analysis Center.