The Effects of Accommodations on the Assessment of Limited English Proficient (LEP) Students in the National Assessment of Educational Progress (NAEP)

CSE Technical Report 537

Jamal Abedi
CRESST / University of California, Los Angeles

Carol Lord
California State University, Long Beach

Christy Kim Boscardin and Judy Miyoshi
CRESST / University of California, Los Angeles

September 2001

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

LEP Accommodations in NAEP
Jamal Abedi, Project Director, CRESST/University of California, Los Angeles

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number RN95127001, Task#1.2.113.1, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

Acknowledgments

We would like to thank all the researchers—faculty, staff, and students—who participated in and informed this study with their generous efforts. We are especially indebted to Joan Herman for her insightful comments and advice. Frances Butler and Robin Stevens provided clarification of terminology and definitions, and Fred Moss provided excellent editorial support for the report. His skill and efficiency facilitated our efforts in completing the report, and we are grateful.

We appreciate the assistance of:

Bahar Ashori
Katie Hutton
Heather Larson
Bobby Mendez
Rory Constancio
Roy Nakawatase
Alicia Soto
Cheryl Sparti
Liliana Velasquez
Katharine Fry

THE EFFECTS OF ACCOMMODATIONS ON THE ASSESSMENT OF LIMITED ENGLISH PROFICIENT (LEP) STUDENTS IN THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP)

Jamal Abedi
CRESST/University of California, Los Angeles

Carol Lord
California State University, Long Beach

Christy Kim Boscardin and Judy Miyoshi
CRESST/University of California, Los Angeles

EXECUTIVE SUMMARY

Recent federal and state legislation, including Goals 2000 and the Improving America's Schools Act (IASA), calls for inclusion of all students in large-scale assessments such as the National Assessment of Educational Progress (NAEP). This includes students with limited English proficiency (LEP). However, we have clear evidence from recent research that students' language background factors impact their performance on content area assessments. For students with limited English proficiency, the language of test items can be a barrier, preventing them from demonstrating their knowledge of the content area.

Various forms of testing accommodations have been proposed for LEP students. Empirical studies demonstrate that accommodations can increase test scores for both LEP and non-LEP students; furthermore, the provision of accommodations has helped to increase the rate of inclusion for LEP students in the NAEP and other large-scale assessments. There are, however, some major concerns regarding the use of accommodations for LEP students. Among the most important issues are those concerning the validity and feasibility of accommodation strategies.

- **Validity.** The goal of accommodations is to level the playing field for LEP students, not to alter the construct under measurement. Consequently, if an accommodation significantly affects the performance of non-LEP students, the validity of the accommodation could be questioned.
- **Feasibility.** For an accommodation strategy to be useful, it must be implementable in large-scale assessments. Strategies that are expensive, impractical, or logistically complicated are unlikely to be widely accepted.

This study focused on the validity and feasibility of accommodation strategies on a small-scale level. In order to test for validity, both LEP and non-LEP students were assessed under accommodated and non-accommodated conditions, and their performance was compared. Feasibility was a key consideration; we selected accommodation strategies for which implementation would be practical in large-scale assessments. Because previous studies have identified the nontechnical vocabulary of test items as a source of difficulty for LEP students (Abedi, Hofstetter, & Lord, 1998; Abedi, Lord, & Plummer, 1995), we chose two forms of accommodation targeting this issue.

Methodology

This pilot study was conducted between November 1999 and February 2000, in two southern California school districts and at one private school site. The purpose of the pilot study was to test the instruments, shed light on the issues concerning the administration of accommodations, explore the feasibility problems that we might encounter in the main study and, ultimately, provide data to help us modify the main study design. A total of 422 students and eight teachers, from six school sites (14 eighth-grade science classes), participated in this pilot study.

A science test with 20 NAEP Grade 8 secure items was administered in three forms: one with the original items and no accommodation, and two with accommodations focusing on potentially difficult English vocabulary. One form of accommodation consisted of a customized English language dictionary at the end of the test booklet. The other form of accommodation consisted of English glosses[1] and Spanish translations of the glosses in the margins of the test booklet.

The customized dictionary—used in this study for the first time as an accommodation for LEP students—contained only words that were included in the test items. The customized English dictionary was grade appropriate and was compiled by CRESST researchers. Providing full-length English dictionaries to test subjects has two major drawbacks: They are difficult to transport, and they provide too much information on the content material being tested. For these reasons, the entries for the words contained in the test were excerpted (with permission from the publisher) to create customized dictionaries that do not burden administrators and students with the bulk of a published dictionary. The pronunciation guide, font and type size were identical to those used in the original dictionary reference.

For each test booklet form, a follow-up questionnaire was developed to elicit student feedback. The follow-up questionnaire was p70

laced in the test booklet immediately after the science test. The questions were tailored to the type of science test the student completed. For example, students who received an accommodation were asked whether that accommodation helped them to answer the science test items. Students’ responses to these questions would be particularly helpful in designing the main study.

[1] A gloss is an individual definition or paraphrase (plural: glosses). According to Webster, a gloss is “a note of comment or explanation accompanying a text, as in a footnote or margin.” A glossary is a collection of glosses; Webster: “a list of difficult, technical, or foreign terms with definitions or translations . . .”. The glosses included brief definitions, paraphrases, or translations.

Also included in the test booklet was a science background questionnaire, which included items selected from both the 1996 NAEP Grade 8 Bilingual Mathematics test booklet and an earlier CRESST language background study. The questionnaire included items regarding the student's country of origin, ethnicity, language background, language of instruction in science classes, and native language and English language proficiency.

In their responses to the science background questionnaire, most of the LEP students self-reported their ethnicity as Hispanic, followed by White, Asian, American Indian, and Other. Most of the non-LEP students self-reported their ethnicity as White, followed by Hispanic, Asian, Black, American Indian, and Other.

A science teacher questionnaire was also introduced midway through the pilot study. This form was used at three sites to obtain information from the science teachers about each class, including type of science class, language of instruction, science topics covered so far that year, and students' English proficiency.

Test administrators received a science test administration script and were asked to complete a feedback questionnaire after each test administration. Test administrators distributed the six test booklets (three accommodation conditions by two forms) randomly within each classroom. The test directions were read aloud to the students. To address the different treatments, general directions were read aloud to the whole class; then specific directions were read aloud, targeted to each treatment group. Students were given 25 minutes to complete the 20-item science test, 3 minutes to complete the follow-up questionnaire, and 8 minutes to complete the science background questionnaire.

Approval to conduct the study was received from the Office for Protection of Research Subjects (OPRS) at the University of California, Los Angeles (UCLA). Test administrators included CRESST research staff and retired teachers and school administrators, all of whom had prior experience with test administration.

Results

This study examined the effectiveness of accommodations by addressing the difficulty of English vocabulary within test items from a NAEP science assessment. We compared LEP and non-LEP students' scores on 20 science items under three different conditions: the standard NAEP condition (no accommodation), customized dictionary, and glossary. The analyses provided clear results with respect to the performance levels of LEP/non-LEP students, the effectiveness of the accommodations for LEP students, and the validity of the accommodated assessment.

- **Performance gap.** LEP students performed lower than non-LEP students. For LEP students, the mean score was 8.97 (*SD* = 4.40, *n* = 183) and for non-LEP students the mean was 11.66 (*SD* = 3.68, *n* = 236). The difference between the performance of LEP and non-LEP students is relatively large and is statistically significant (*t* = 6.83, *df* = 417, *p* = .000).

- **Effectiveness of accommodations.** LEP students performed substantially higher under the accommodated conditions than under the standard condition. The mean for the LEP students under the customized dictionary condition was 10.18 (*SD* = 5.26, *n* = 55); under the glossary condition, the mean was 8.51 (*SD* = 4.72, *n* = 70); and under the standard condition, the mean was 8.36 (*SD* = 4.40, *n* = 58). As the data suggest, LEP students did particularly well under the customized dictionary condition. The results of an analysis of variance (ANOVA) indicated that the difference between means for LEP students under the three accommodation conditions was significant (*F* = 3.08, *df* = 2, 180, *p* = .048).
- **Validity.** The accommodations had no significant effect on the scores of the non-LEP students. For non-LEP students, the mean score for the customized dictionary condition was 11.37 (*SD* = 3.79, *n* = 82); for the glossary condition, the mean was 11.96 (*SD* = 3.86, *n* = 75); and for the standard condition, the mean was 11.71 (*SD* = 3.40, *n* = 79). The results of an analysis of variance showed no significant difference between the performance of non-LEP students under the three conditions (*F* = .495, *df* = 2, 233, *p* = .610).

These results suggest, first, that the customized dictionary enabled LEP students to perform at a significantly higher level, and second, that the accommodation strategies used in this study did not impact the construct, and the validity of the assessment was not compromised. These results are particularly encouraging, given the ease of administration of the accommodations that were used.

In student responses to the follow-up questionnaire, LEP students reported greater difficulty with the language of the test items. (Follow-up questionnaires were similar, but not identical, for the three accommodation levels of the test.)

- More LEP students than non-LEP students indicated there were words that they did not understand in the science test.
- More LEP students than non-LEP students wanted explanation of some of the difficult words.
- More LEP students than non-LEP students expressed interest in using a dictionary during the test.
- More LEP students than non-LEP students indicated that it would have helped them if the test had explained words in another language.
- More LEP students than non-LEP students expressed a preference for a dictionary during the test.

Analyses based on the background variables showed no significant gender differences. However, a significant difference was found between the performance of students who spoke only English in the home and those who spoke a language other than English in the home, with students who spoke a language other than

English in the home performing significantly lower. This finding is consistent with the literature and with the main findings of this study.

Analyses of self-reported data showed that students who spoke a language other than English in the home indicated that they spoke that language more with their parents and less with their brothers, sisters, and friends. These findings, reflecting a generation gap, are consistent with the existing literature.

The results of analyses of self-reported data on English proficiency were also consistent with the literature and with the earlier findings of this study. As expected, LEP students reported significantly lower proficiency in English than their non-LEP counterparts.

Limitations

Because this was a pilot study that was planned to test the instruments and logistics for the main study, the generalizability of findings from this study is extremely limited. The generalizability of the study is further limited to grade level (Grade 8), content area (science), LEP language background (primarily Spanish), and accommodation type (customized dictionary and glossary).

It should also be noted that an accommodation for one grade level may not necessarily be appropriate, or even considered an accommodation, for another grade level. Students in lower elementary grades may not know how to use a dictionary or may be in the process of learning to use a dictionary, whereas students in higher elementary grade levels and above may be accustomed to regularly using a dictionary. For older students, dictionary use during a testing situation is considered an accommodation, whereas for younger students, dictionary use during a testing situation may not be considered an effective form of accommodation because they may not know how to use a dictionary.

In an effort to find classrooms with an equal number of LEP and non-LEP students, site selection was based on state demographic information at the school site level. However, state demographic information does not necessarily reflect the LEP and non-LEP distribution for individual classes at a school site. Therefore, site selection in the main study should be based on demographic information collected at the classroom level.

A large proportion of the LEP population in southern California is native Spanish speaking. Accordingly, for the glossary accommodation we included English glosses and Spanish translations. In our sample, 88% of the LEP students were Hispanic and 26% of the non-LEP students were Hispanic. LEP students with first languages other than Spanish may have benefited from the English glosses, but the accommodation tells us little about the potential impact of translations in their first languages.

Implications and Recommendations

This study addresses several major issues concerning accommodations for LEP students in NAEP. Although these analyses report on the pilot phase of the study, there are nevertheless several implications for future NAEP assessments.

Since NAEP is a large-scale assessment, feasibility considerations are important. NAEP assessments involve a large number of LEP students, so ease of administration may be a determining factor. Any element that reduces the burden on states, schools, and students will potentially have a positive impact on future NAEP administrations. Educators are developing accommodation strategies that may reduce the gap between LEP and non-LEP scores in large-scale assessments. Not all of these strategies may turn out to be easily administered. One-on-one testing, for example, may be a highly effective form of accommodation, but it may not be feasible in large-scale assessments such as the NAEP.

Providing a customized dictionary is a viable alternative to providing traditional dictionaries. Dictionaries are, in fact, already widely used as instructional aids for LEP students, so the concept is not an unfamiliar one for them. Including a customized dictionary as part of the test booklet can minimize the economic and administrative burden and may help to overcome shortcomings on the validity of accommodations using dictionaries. However, the economic and technical feasibility of providing a customized dictionary as a potential form of accommodation should be evaluated through cost-benefit analyses.

Gathering additional information about the academic performance and the language proficiency levels of students may help to clarify issues associated with inconsistency in the definition of limited English proficiency and the inclusion criteria for standardized assessments. The reading achievement data from the Stanford 9, supplied by the schools, provided valuable information on the language proficiency levels of students, beyond the LEP designations. Given the inconsistency in the LEP designation criteria, collecting additional information about a student's academic and language performance would provide a more comprehensive picture of the student's academic knowledge. More accurate conclusions would be possible from analyses of contextual data, such as a student's performance in other content areas and information on family and language background.

Critical Steps to Follow: Necessity for the Main Study

The results of experimentally controlled accommodation studies may provide assistance to NAEP in its future assessments. This study was designed to address two of the major issues of concern for NAEP, the validity and feasibility issues. Regarding validity, it is important to understand how accommodations impact assessment in NAEP. Any systematic effect of accommodation would impact both the trend and the reporting of NAEP. Regarding feasibility, even a minor modification in the design of accommodations—to make accommodation more implementable and logistically easier—would enhance the design for inclusion of students with limited English proficiency.

As indicated earlier, this pilot study was conducted to help in designing the main study. The generalizability of the findings is limited for the following reasons:

- The number of subjects in this pilot study is small; therefore, there may not be enough statistical power to ascertain and estimate effects.
- Due to the nature of a pilot study, instruments and logistics were often modified throughout this pilot study, based on what we learned from the previous stages of the study.
- Because this was a pilot study, we did not aim to select a truly representative sample of students.
- Because of time and resource limitations, we included students in Grade 8 only. To broaden the level of generalizability, other grade levels and other accommodation strategies should be included.
- We also recommend the addition of another language (for example, Chinese) to have a more representational sample.

The main study will greatly increase the generalizability of the findings.

THE EFFECTS OF ACCOMMODATIONS ON THE ASSESSMENT OF LIMITED ENGLISH PROFICIENT (LEP) STUDENTS IN THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP)

Jamal Abedi
CRESST/University of California, Los Angeles

Carol Lord
California State University, Long Beach

Christy Kim Boscardin and Judy Miyoshi
CRESST/University of California, Los Angeles

We now have clear evidence that students' language background and the language of assessment impact student performance on content area tests (see for example, Abedi, Lord, Hofstetter, & Baker, 1998; Abedi, Lord, & Plummer, 1995; Aiken, 1971, 1972; Cocking & Chipman, 1988; De Corte, Verschaffel, & De Win, 1985; 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). Language is therefore a crucial issue in the assessment of students with limited English proficiency[1] (LEP).

Based on the wealth of evidence concerning the impact of language on content-based assessment, it can be argued that since most state and national assessment tools are constructed and normed for native English speakers, using such assessments for LEP students may not be fair. It would follow that until more valid and fair assessment tools are provided, LEP students should not be included in such assessments.

[1] Limited English Proficient (LEP) is the official term found in federal legislation and is the term used to define students whose first language is not English and whose proficiency in English is currently at a level where they are not able to fully participate in an English-only instructional environment (Olson & Goldstein, 1997).

English language learner (ELL) is a term used in some citations found in this report. English language learner, as defined by LaCelle-Peterson and Rivera (1994), broadly refers to students whose first language is not mainstream English. ELLs include students who may have very little ability with the English language (frequently referred to as LEP) and those who may have a high level of proficiency.

The term LEP will be used in this report because our accommodations are specifically intended for use with this population of ELLs. When the term ELL appears in citations in this report, the authors are usually referring to the LEP population.

However, the authors of this report would like to acknowledge LaCelle-Peterson and Rivera's perspective that ELL is a positive term because it implies a second language. LEP, on the other hand, conveys the perspective that the student has a deficit or a "limiting" condition.

On the other hand, recent federal and state legislation, including Goals 2000 and the Improving America's Schools Act (IASA), calls for inclusion of all students in assessments. This includes LEP students. However, if LEP students are to be included, the issue of the impact of students' language background on their content-based performance must be addressed.

Previous studies have shown that utilizing some forms of accommodation can increase test scores for both LEP and non-LEP students. For example, in an experimentally controlled study, Abedi, Hofstetter, Lord, and Baker (1998) found that a combination of glossary use and extra time increased LEP students' performance by over half a standard deviation. Other forms of accommodation, such as linguistic modification, may narrow the performance gap between LEP and non-LEP students (Abedi et al., 1995; Abedi, Hofstetter, Lord, & Baker, 1998).

Provision of accommodations has helped to increase the rate of inclusion for LEP students (Mazzeo, 1997). Based on the promising results from using accommodations in the 1996 National Assessment of Educational Progress (NAEP) main assessment, accommodations were provided in the 1997 assessment in art and in the 1998 assessments in reading, writing, and civics.

There are, however, some major concerns regarding the use of accommodations for LEP students. Among the most important issues are those concerning the validity and feasibility of accommodation strategies. As indicated earlier, providing accommodations has increased LEP students' performance, but at the same time non-LEP students have also benefited. This may be problematic, since the purpose of using accommodations is to reduce the gap between LEP and non-LEP students, not to alter the construct under measurement. The use of accommodation strategies that affect the construct is questionable. Feasibility is another major issue in the provision of accommodations. Valid accommodation strategies may not be useful if they cannot be easily implemented in large-scale assessments.

This study focuses on the validity and feasibility issues of accommodation strategies. In the study, both LEP and non-LEP students were tested under accommodated and non-accommodated conditions; this provided the basis for testing the validity of accommodation. In addition, we selected two accommodation strategies for which implementation was feasible in large-scale assessments: a customized dictionary and a glossary.

Dictionaries have been suggested as a form of accommodation (Kopriva, 2000). There are, however, caveats concerning the use of dictionaries for accommodation. First, there are validity issues. The accommodation strategy should not impact the construct. Accordingly, the accommodation should not provide content-related information. However, a standard dictionary would provide access to both content and non-content terms. Further, there are various types of dictionaries, differing in purpose, content, form, and scope. Use of different dictionaries might result in different levels of performance.

A second issue is feasibility. Providing the same edition of a dictionary to all participants would be difficult. It would be unrealistic to require all students to bring the same version of a dictionary to an assessment. Furthermore, providing students an opportunity to bring outside materials to the test would pose difficult issues of screening. On the other hand, requiring the administrator to provide dictionaries for all students could pose logistical problems.

To deal with feasibility concerns, we introduced the idea of a customized dictionary, for the first time in this study. The customized dictionary contains only the vocabulary for items that occur in the test. In consultation with library experts and teachers, a widely used standard dictionary was selected. This dictionary was used to create definitions for only words and wordsenses that were in the test, resulting in a customized dictionary.

In addition to the customized dictionary, a glossary was included in the study as a second form of accommodation. The glossary accommodation provided Spanish translations and brief English glosses in the page margins.

These two accommodation strategies were used in the study, along with a standard form of the test as a comparison or control condition. Performance of students under the two accommodation strategies was compared with performance of students under the standard, control condition.

Research Question and Hypotheses

The main research question in this study was whether or not the accommodation strategies that were used in the study reduced the performance gap between LEP and non-LEP students. First, we sought to determine the impact of accommodations on LEP students’ performance.

H_{01}: LEP students tested under accommodation conditions perform the same as LEP students tested with no accommodation.

H_{11}: LEP students tested under accommodation conditions perform better than LEP students tested with no accommodation.

The research question concerning the validity of accommodation is of particular importance in any accommodation study. The following research hypotheses addressed the validity of accommodation.

H_{02}: Non-LEP students tested under accommodation conditions perform the same as non-LEP students tested with no accommodation.

H_{12}: Non-LEP students tested under accommodation conditions perform better than non-LEP students tested with no accommodation.

We also addressed the question of effectiveness of these accommodations as a strategy for increasing test validity for LEP students.

H_{03}: The performance gap between accommodated and non-accommodated performance is the same for LEP and non-LEP students.

H_{13}: Accommodation strategies that are used in this study reduce the gap between accommodated and non-accommodated performance more for LEP students than for non-LEP students.

Literature Review

Based on a nationally representative sample of school districts in 1991, the number of LEP students in Grades K-12 was estimated to be more than 2.3 million (Olson & Goldstein, 1997). In recent efforts to increase participation of language minority students in large-scale assessments, accommodations and adaptations have been proposed as strategies for including these students. About 55% of U.S. states are now providing various accommodations to comply with the mandated inclusion criteria.

Recent studies have examined the impact of language proficiency among both native and non-native English speakers on content-based performance. Differential performances between LEP students and native English speakers in subject areas such as mathematics and science have been attributed to differences in English language proficiency levels. Difficulties in the language of content-based test items have been identified as a significant factor in overall content-based performance.

This literature review provides a brief overview of issues related to the inclusion of LEP students in large-scale assessments, focusing on the following areas:

1. differences in performance between LEP and non-LEP students across content areas;
2. linguistic factors related to science performance;
3. the effects of accommodations.

Content-Based Performance and Limited English Proficient Students

Previous studies have shown that the differences between achievement levels of LEP students and native English speakers are significant (Cocking & Chipman, 1988). Specifically, in mathematics, studies have shown that English proficiency levels are associated with performance on solving word problems (Carpenter, Corbitt, Kepner, Linquist, & Reys, 1980; Mestre, 1988). A study by Butler and Castellon-Wellington (2000) found that native English speakers outperformed both fluent English proficient (non-native English speakers) and limited English proficient students in standardized mathematics assessments. However, Abedi and Leon (1999), in a study using data from several different school districts nationally, demonstrated that the performance gap between LEP and non-LEP decreases as the level of language demand in test items decreases. For example, they showed that the performance gap between LEP and non-LEP students was greatest in reading, decreased substantially in science and became non-existent in math items, particularly those involving mainly computations (see also Abedi, Leon, & Mirocha, 2000).

As Mestre (1998) suggested, language deficiencies may contribute to the misinterpretation of word problems. Cocking and Chipman (1988) concluded that Spanish-dominant students scored higher on the Spanish version of a math placement test than on the same test in English. A 6-year longitudinal study by Moss and Puma (1995) found that LEP students who attended public schools were particularly disadvantaged.

The positive relationship between language proficiency and academic performance has been established by several studies. A study by De Avila, Cervantes, and Duncan (1978) demonstrated that oral language proficiency was a significant predictor of academic performance (De Avila, 1997). De Avila et al. (1978) found that there was a linear relationship between the five levels of a widely used oral language proficiency assessment and performance on a standardized test, the CTBS-U (De Avila, 1997). A replication of this study (De Avila, Duncan, & Navarrete, 1988) found that academic performance was directly associated with literacy skills.

A study conducted by the Minnesota Assessment Project found that more LEP students passed the math tests than the reading tests (Thurlow et al., 1998). Thurlow et al. suggested that the overall poor performance of LEP students may be a result of low reading and comprehension skills, due to unfamiliarity with American English idioms and vocabulary. Previous research has suggested that the types of language or discourse required in an academic setting may be very different from the home practices and experiences of many language minority students (Heath, 1983).

As suggested by many researchers, the level of language proficiency is one of the contributing factors to differences in achievement levels (Abedi et al., 1995; Cocking & Chipman, 1988). To ensure the validity of content area assessments, the effects of language proficiency on performance in content areas such as mathematics and science can be minimized. By reducing the difficulties associated with English language proficiency level, we can establish more valid inferences about LEP students' content area knowledge.

As pointed out in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985), "for a non-native English speaker, and for a speaker of some dialects of English, every test given in English becomes, in part, a language or literacy test" (p. 75). For accurate assessment of such students' content knowledge, accommodations are considered an alternative strategy to ensure validity and reliability of content assessments in mathematics and science. One of the challenges for inclusion of all students in large-scale assessments is that standardized test developers usually assume that the test takers have no language difficulties that would interfere with test performance (Lam & Gordon, 1992; Zehler, 1994).

Linguistic Variables and Science Performance

Previous studies have suggested that linguistic modifications of math word problems are associated with increased math test performance. Certain linguistic features, such as unfamiliar lexical items and passive voice verb constructions, have been implicated as potential contributors to the difficulty of text interpretation (Abedi et al., 1995).

Studies have suggested that cognitive development in science is greatly dependent upon the linguistic development of a student (Anstrom, 1997; Kessler, Quinn, & Fathman, 1992). The acquisition of certain linguistic skills, such as interpreting logical connectors and specialized vocabulary, is considered a prerequisite for demonstrating the advanced reasoning skills used in scientific communication (Anstrom, 1997). The discourse patterns common in scientific texts, such as compare / contrast, cause / effect, and problem / solution, require a high level of linguistic functioning that may be problematic for language minority students (Anstrom, 1997). Scientific language frequently contains complex sentences using passive voice constructions, which may pose greater challenges to language minority students trying to comprehend scientific texts than to students whose first language is English.

Scientific texts often use jargon that may pose challenges for understanding. According to Halliday and Martin (1993), "scientific texts are found to be difficult to read; and this is said to be because they are written in 'scientific language', a 'jargon' which has the effect of making the learner feel excluded and alienated from the subject-matter" (p. 69).

A study by Cassels and Johnstone (1984) concluded that using simpler words brought about an improvement in students' performance on chemistry multiple-choice tests. Replacing a question such as "Which is the least stable sulfide among the following?" with a simplified question such as "Which one of the following sulfides is easiest to break down to its elements?" increased percent correct from 40% to 49%.

According to Abedi, Hofstetter, and Lord (1998), clarifying the language of math items by modifying the linguistic structures and non-technical vocabulary enabled LEP students to achieve higher scores and narrowed the score gap between LEP and non-LEP students.

Language demands in standardized content assessments often exceed the language proficiency levels of LEP students. An evaluation of 11th-grade standardized math and science assessments by Butler and Castellon-Wellington (2000) concluded that "approximately two-thirds to three-quarters of the test items on the mathematics and science subsections, respectively, had general vocabulary rated as uncommon or used in an atypical manner" (p. 98). Butler and Castellon-Wellington also found that the majority of the test items in both standardized mathematics and science assessments contained challenging syntax and vocabulary. As suggested by Gesinger and Carlson (1992), "testing procedures must be sensitive to the needs of LEP students and those from cultural minorities" (p. 2).

Accommodations

The purpose of accommodations is to help remove any irrelevant variances associated with the construct so that students' content knowledge can be accurately measured (McDonnell, McLaughlin, & Morison, 1997). Behind testing accommodations is the theoretical assumption that the elimination of language barriers in testing formats will give students the optimal opportunity to show their true ability in the subject area. Previous studies have shown that students who are being instructed in their native language demonstrate their knowledge in content areas much better in that language or in a combination of the first and second languages (Zehler, 1994).

The availability of testing accommodations can provide an environment conducive to greater participation of LEP students in large-scale testing. August and McArthur (1996) reported that the National Center for Education Statistics (NCES) has found that teachers included more LEP students in NAEP tests when more accommodations were available. An evaluation of the NAEP inclusion criteria found that increases in the percentage of LEP students included will be possible if the list of accommodations and adaptations can be expanded (Mazzeo, Carlson, Voelkl, & Lutkus, 2000). With additional accommodations, other than translated or interpreted versions of the tests, more students may be encouraged to take the tests in English (August, Hakuta, & Pompa, 1994).

In a survey of types of accommodations, Butler and Stevens (1997) categorized approaches as modifications of the test or of the test procedure (see Figure 1).

State policies for accommodations. Shepard, Taylor, and Betebenner (1998) found that accommodations consistently raised the relative position of LEP students on performance-based assessments. In Florida, for example, accommodations for LEP students include flexible scheduling, additional time, clarification of a word or phrase for general directions, and use of dictionaries (Abedi, Boscardin, & Larson, 2000). A study conducted by the North Central Regional Educational Laboratory (NCREL) in 1996, however, found that 7 out of 50 states assessed LEP students with no accommodations, and only half of the states allowed testing accommodations for LEP students. The recommendations of a panel from a symposium sponsored by

Two Categories of Accommodations for English Language Learners

Modifications of the Test

Assessment in the native language
Text changes in vocabulary
Modifications of linguistic complexity
Addition of visual supports
Use of glossaries in English
Use of glossaries in native language
Linguistic modifications of test directions
Additional example items

Modifications of the Test Procedure

Extra assessment time
Breaks during testing
Administration in several sessions
Oral directions in the native language
Small-group administration
Separate room administration
Use of dictionaries
Questions read aloud in English
Answers written directly in test booklet
Directions read aloud or explained

Figure 1. Potential accommodation strategies for English language learners (Butler & Stevens, 1997).

the U.S. Department of Education, Office of Bilingual Education and Minority Languages Affairs (National Clearinghouse for Bilingual Education, 1997), included the use of native language assessments, bilingual versions of assessments, alternative modes of response, and portfolios of student work.

Some of the most widely used forms of accommodation in state assessments are identified as flexible scheduling, extra time, simplified instructions, and dictionary and glossary usage. In New York City, the mathematics assessments are currently translated into five languages: Chinese, Haitian, Creole, Russian, and Spanish (Abedi et al., 2000). Rhode Island offers native language test versions in Grades 4, 8, and 10, which include Spanish, Portuguese, Laotian, and Cambodian (Stansfield, 1998). In Massachusetts, all state assessments are offered in Spanish and use a specialized scoring system involving bilingual and content area teachers (Stansfield, 1998).

Problems with direct translation. Previous studies have indicated that there are several linguistic and cultural problems associated with direct translation of tests into native language (see, for example, Abedi, Hofstetter, & Lord, 1998; Olmedo, 1981). For example, there are numerous dialects within Spanish that may differ across countries and regions of the world. Given the cultural context of a word, a direct translation may not provide the same meaning across dialects and cultures. As pointed out in a report prepared by the Council of Chief State School Officers

(Kopriva, 2000), “confusion can result from rules of syntax or word order that differ in a student’s home language. Yet another common source of student confusion comes from words that mean something different in English than in the student’s home language” (p. 42).

Item analysis revealed that a large percentage of Spanish items used in NAEP math assessments had item statistics that were dissimilar to those of the same items in English (Anderson, Jenkins, & Miller, 1996). Abedi, Hofstetter, and Lord (1998) found that eighth-grade Hispanic students designated as LEP scored higher on NAEP math items in English compared to their peers who received the same items administered in Spanish. However, those students receiving instruction in Spanish performed higher on the math items in Spanish than on either modified or standard English items.

In addition, technical difficulties associated with direct translation of tests have been pointed out by many researchers (Figueroa, 1990). One of the most serious difficulties is trying to establish the reliability and the validity of translated tests. As Olmedo (1981) pointed out, translated items may exhibit psychometric properties substantially different from those of the original English items. Since direct translation is not possible, the slight modifications in the translated version to conform to the rules and patterns of the new language may significantly change the psychometric properties of the item. Consequently, the reliability and validity of translated tests need to be firmly established for limited English proficient students before inferences about their test performance are made.

A study by Valencia and Rankin (1985) reported that the McCarthy Scales of Children’s Abilities (Massoth & Levenson, 1982; McCarthy, 1972) translated into Spanish showed bias against Mexican American Spanish-speaking children in the verbal and numerical memory subtests. Valencia and Rankin concluded that the effect of word length and acoustic similarity on information-processing load might have contributed to the content biases.

According to Liu, Thurlow, Erickson, Spicuzza, and Heinze (1997), direct translation of tests is thought to be beneficial for only two types of LEP students: (a) students who received grade-appropriate instruction or educational experience in their first language or in a bilingual program, and (b) students who are more fluent in their first language than their second, even though they have not been instructed in their first language, and who choose to take a translated version of a test (August et al., 1994, cited in Liu et al., 1997). A study by Thurlow et al. (1998) indicated that students found idiomatic expressions in English difficult to understand and that the Spanish translations were not very helpful.

A report prepared by the Council of Chief State School Officers (CCSSO; Kopriva, 2000) also suggested that “while many LEP students are orally proficient, at least conversationally, in their home language, we should not assume they will be literate in their home language unless they have had steady, consistent, and in-depth instruction in these specific skills” (p. 52). Solano-Flores and Nelson-Barber (2000) pointed out that a simplistic belief that adapting a test (e.g., by translating it into another language or by providing accommodations) is enough to properly serve diverse populations can have the catastrophic effect of “contributing to perpetuating inequalities in the assessment of these groups” (p. 4).

Glossary and dictionary usage. The use of a glossary is a potential form of accommodation for LEP students in large-scale assessments. For the 1995 NAEP mathematics assessments, glossaries in both Spanish and English were used as accommodations for LEP students. A study by Abedi, Hofstetter, and Lord (1998) found that both students with limited English proficiency and English-proficient students benefited from an English glossary along with extended time in mathematics assessments.

One of the positive aspects of using glossaries or dictionaries as accommodations is that these materials are widely used as part of instruction (Kopriva, 2000). Based on an accommodation study evaluating the effect of Spanish translation on performance, Thurlow et al. (1998) concluded, “It seems that the students would have preferred some sort of glossary to explain the vocabulary word” (p. 5). According to Thurlow et al., the students found the Spanish translation did not always “help them understand the [vocabulary] word because they often did not know the word in Spanish either” (p. 5).

Extended time. A meta-analysis conducted by Chiu and Pearson (1999) found that extended time was the most frequently investigated accommodation. Of 30 research studies that they reviewed, almost half (47%) of the accommodations provided extended time or unlimited time. A recent study by Ofiesh (1997) found differential timing effects for learning disabled (LD) and non-learning disabled (NLD) students when the Nelson Denny Reading Test (Riverside Publishing, Itasca, IL; Perkins, 1984) was administered to students in postsecondary schools. Ofiesh found that the target populations benefited from the accommodation while the NLD students were at neither an advantage nor a disadvantage with the extra time. In another study, Montani (1995) found that providing unlimited time increased the scores of both the LD and NLD students in mathematics tests. Abedi, Hofstetter, and Lord (1998) found that the provision of extra time increased performance of non-LEP students only slightly, but extra time plus a glossary had a significant impact on math performance for both LEP and non-LEP students**.**

According to a meta-analysis by Chiu and Pearson (1999), extended or unlimited time accommodations benefited both the target population and the control groups. The study found the comparative advantage for the target population to be only modest. However, some studies (Braun, Ragosta, & Kaplan, 1988; Willingham, Bennett, Braun, Rock, & Powers, 1988) have found that providing extra time appeared to give too much of an advantage to students with learning disabilities. Since the results of providing extra time do not appear to be consistent across studies, it may be that the effect depends in part on other factors such as the nature of the content or item type, or the background of a particular group of students.

Recommendations for Testing

As previous studies have cautioned, in order to derive valid inferences about test results, test developers need to take into consideration the effect of the linguistic and cultural characteristics of the test takers (Gonzales, Castellano, Bauerle, & Duran, 1996). To be valid for LEP students, assessments have to be linguistically and culturally appropriate. Accommodations may provide a systematic way to minimize linguistic and cultural differences. According to a recent report by Shepard et al. (1998), “very few LEP students received accommodations specific to their language needs” (p. 53).

For construct validity purposes, accommodations need to be validated with the intended test takers in mind. According to Gonzales et al. (1996), “it is ethically inappropriate for an evaluator to use a standardized assessment procedure when there is no evidence of construct validity to its practical application for making diagnostic and placement decisions” (p. 452).

Methodology

This investigation was a pilot study to examine the use of accommodations by LEP students on a test comprised of NAEP science questions. The study took place between November 1999 and February 2000 in two southern California school districts and at one private school site. A total of 422 students and 8 teachers, from 6 school sites (14 eighth-grade science classes) participated in the study.

A science test with 20 NAEP items was administered in three forms: one with original items and two with accommodations focusing on potentially difficult English language vocabulary. One form of accommodation included a customized English language dictionary at the end of the test booklet. The other form of accommodation included English and Spanish language glossaries in the margins of the test booklet. In addition, a follow-up questionnaire and a science background questionnaire were administered. Student scores on the unaccommodated tests were compared with scores on the accommodated tests. Participants, instruments, and procedure are described below.

Participants

A total of 422 eighth-grade science students, ages 13-14 years, from 6 school sites, participated in the study. Of the 422 students, 199 were female and 222 were male (information was incomplete for one student).

Teachers provided the English proficiency levels of students from school records. Of the 422 students, 183 students were identified as being limited English proficient (LEP) and 236 were identified as proficient English speakers (non-LEP) (see Table 1). Data were not available for 3 students.

The method used to determine English language proficiency and to monitor the academic progress of students in language programs varies across states and even within school districts. In general, many combinations of information, including registration and enrollment records, home language surveys, interviews, observations, referrals, classroom grades and academic performance, and test results, are used to determine a student's proficiency level and to monitor academic progress (Olson & Goldstein, 1997).

Given the myriad methods and combinations of methods that school districts can use to identify, place, and teach LEP students, it is extremely difficult to make comparisons across districts and institutions. This study, with participants from

Table 1

LEP and Non-LEP Students (*N* = 422)

	LEP	Non-LEP	Total
Site 1	64 (15.2%)	0	64 (15.3%)
Site 2	61 (14.5%)	0	61 (14.6%)
Site 3	0	37 (8.8%)	37 (8.8%)
Site 4	32 (7.6%)	28 (6.6%)	60 (14.6%)
Site 5	6 (1.4%)	139 (32.9%)	145 (34.6%)
Site 6	20 (4.7%)	32 (7.6%)	52 (12.4%)
Total students	183	236	419 (100.0%)

Note. Data not available for 3 (0.7%) students.

school sites in two different school districts and one private school site, used LEP and non-LEP designations from school-site records, information obtained from a science background questionnaire, and state testing results as criteria for analyses and comparison of the LEP and non-LEP groups. However, we realize that some discrepancies across sites may still exist regarding LEP and non-LEP status of students (i.e., an LEP student from one school district may not be considered an LEP student in another school district), and we caution that the results should be interpreted accordingly.

In their responses to the science background questionnaire, most of the LEP students self-reported their ethnicity as Hispanic, followed by White, Asian, American Indian, and Other. Most of the non-LEP students self-reported their ethnicity as White, followed by Hispanic, Asian, Black, American Indian, and Other (see Table 2).

Table 2

LEP Classification and Ethnicity (*N* = 422)

	LEP	Non-LEP
American Indian	2 (.5%)	1 (.2%)
Asian	7 (1.7%)	31 (7.3%)
Black	0	8 (1.9%)
Hispanic	158 (37.4%)	60 (14.2%)
White	10 (2.4%)	97 (22.9%)
Other	2 (.5%)	31 (7.4%)
Total students	179	228

Note. Data not available for 15 (3.6%) students.

Instruments

Students completed a science test, a follow-up questionnaire, and a science background questionnaire. Teachers completed a science teacher questionnaire. Test administrators followed a science test administrator script developed for this study by CRESST, and each administrator was asked to complete a test administrator feedback questionnaire.

Science test. Each student was given a 20-item science test. Multiple-choice items from the 1996 main NAEP eighth-grade science assessment were selected for the test. The items chosen were judged to contain words that students might find difficult or unfamiliar, or words used in a sense or context that students might find difficult or unfamiliar. Judgements were based on nontechnical words only; for example, a word such as “location” would be considered in item selection, but a content-related word such as “tectonic” would not be considered. Three test booklet types were created.

1. One test booklet (Unaccommodated) contained only the original items, as a control or comparison treatment.

2. One test booklet (Dictionary) included the original items and a customized English-language dictionary containing almost all the words in the test, including the content-related words. The dictionary was printed on paper of a contrasting color and was stapled to the end of the test booklet.

3. One test booklet (Glossary) contained the original items and glossary entries for potentially difficult words. Words were explained in the margins on each test page. In the right margin were short definitions or explanations in English; in the left margin of the page were Spanish translations of the definitions or explanations.

For each of the three booklet types, two counterbalanced forms were created. The items in the first half of form A occurred in the second half of form B; items in the second half of form A occurred in the first half of form B. Thus, there were a total of six different forms of the science test.

Since the items were from a secured NAEP test, actual items are not provided in this report. However, Figure 2 shows a comparable item for illustrative purposes. In the control booklet (unaccommodated) the item would have appeared as it does in Figure 2.

The locations of earthquakes in the past ten years are marked on a world map. What can we learn from this map?

A. Earthquakes happen with the same frequency everywhere on the Earth.

B. Earthquakes usually happen along the edges of tectonic plates.

C. Earthquakes most often happen near the middle of continents.

D. Earthquakes do not seem to happen in any regular pattern.

Figure 2. Illustrative comparable test item.

In the Dictionary booklet, the items appeared as in the control booklet (no glosses in the margins), but the dictionary appended to the test booklet contained definitions for almost all words from the items. Nouns, verbs, and adjectives were included in the dictionary, but high-frequency words such as articles, pronouns, and some prepositions were not included. It was assumed that students who did not know these words would not be helped by dictionary definitions of them.

Word definitions were based on those in the *Longman Dictionary of American English* (1997) and included those wordsenses occurring in the test items. For the item represented by Figure 2, the dictionary would contain the words and phrases *location, earthquakes, past, years, marked, world, map, learn, happen, same, frequency, everywhere, Earth, usually, along, edges, tectonic plates, near, middle, continents, seem,* and *regular pattern*. A typical Dictionary entry might be:

location: a particular place or position

Because the dictionary included almost all words from the items, it included definitions of content-related vocabulary, such as “tectonic plates” and “continents.” The choice to include all words was made so that the results of this study could be more meaningfully compared to the results of other studies, in which students were provided actual dictionaries as an accommodation.

In the Glossary booklet, the same items appeared as in the control booklet, but the left margin of each page contained Spanish translations of vocabulary words or phrases judged to be potentially difficult. A typical Spanish gloss might be:

location: lugar

Glosses were drafted for each test item by a bilingual Spanish / English research assistant with experience in middle school classrooms. The glosses were reviewed and edited by a bilingual teacher / translator who was a native of Chile with teaching experience in California junior colleges.

The right margin of each page contained the same potentially difficult words from the item, each followed by a brief gloss in English, based on the appropriate wordsense from the *Longman Dictionary of American English* (1997); for example:

location: place or position

Follow-up questionnaire. For each test booklet type, a follow-up questionnaire was developed to elicit student feedback. The follow-up questionnaire was placed in the test booklet immediately after the science test. Questions were tailored to the type of science test the student completed. The different forms contained from six to nine questions; for example:

Unaccommodated science test:

Would it help if the test explained words in another language?

Science test with customized dictionary:

Did the dictionary help you understand the questions?

Science test with glossary:

Did you read the explanations in the margins in English (on the right side of the page)?

For the three forms of the follow-up questionnaire, see Appendix A.

Science background questionnaire. Included in the test booklet was a science background questionnaire (see Appendix B) with 35 questions selected from both the 1996 NAEP Grade 8 Bilingual Mathematics booklet and an earlier language background study (Abedi et al., 1995).

The questionnaire included inquiries about the student's country of origin, ethnicity, language background, language of instruction in science classes, and English proficiency; for example:

What country do you come from?

How long have you lived in the United States?

Do you speak a language besides English?

Have you ever studied science in a language other than English?

How long have you studied science in English?

Does your family often get a newspaper written in English?

Do you read English well?

Demographic form. Teachers were asked to complete a demographic form for each class that participated in the study. The form was used to gather information about student gender, ethnicity, free lunch program eligibility status, LEP or non-LEP status, SAT-9 scores, and language spoken at home (see Appendix C).

Science teacher questionnaire. A questionnaire was introduced midway through the pilot study and used at sites 4 through 6 to obtain information from each science teacher about each class, including type of science class, language of instruction, topics covered so far this year, and teacher judgment of students' English proficiency (see Appendix D).

Script for science test administrators. A script was prepared for the test administrators to ensure consistent testing procedures across classrooms and across school sites (see Appendix E).

Test administrator feedback form. Each test administrator was asked to provide feedback and comments on each administration. This information was gathered mainly to address or improve test administration procedures, thus resulting in modification of the test administrator script (see Appendix F).

Procedure

Human subjects approval. Approval to conduct the study was received from the Office for Protection of Research Subjects (OPRS) at the University of California, Los Angeles (UCLA). Student consent forms were not used for this study in order to match the testing procedures used for NAEP testing. The OPRS's Human Subject Protection Committee at UCLA approved this request.

Test administrators. Test administrators included CRESST research staff and a group of retired teachers and school administrators who also had prior experience with test administration.

Site selection. The initial goal for site selection was to use eighth-grade science classrooms with an equal distribution of LEP and non-LEP students. A demographic form was developed by CRESST and sent to teachers to elicit language background information about the students in their classrooms (see Appendix C for the demographic form).

Based on feedback from the teachers, it became clear that it would be extremely difficult to locate sites with an equal balance of LEP and non-LEP students in the same classroom. From the more than 30 sites contacted, 6 were confirmed for participation. A letter to the principal described the study (see Appendix G). Both the school site and the teacher participant received an honorarium of $125.

Testing procedures. Test administrators distributed the six test booklet forms randomly within each classroom. The test directions were read aloud to the students. Students were informed that their score on the test would not be a part of their grade for the class. To address the different treatments, the directions were read aloud to the whole class; then each treatment group was read specific directions that were targeted to that group. For example, if a student received a Glossary-A or Glossary-B test booklet, the directions were as follows:

> If the bottom line on your test booklet says “Glossary-A” or “Glossary-B,” please raise your hand. These directions are for you. In the margins of the pages in your test booklet, certain words are explained. If the meaning of a word is not clear, you may look at the explanation in the margin. On the right side of the page, you will find explanations in English [assistant test administrator hold up a “Glossary” test booklet, open to page 3, and point to the English glosses]. On the left side of the page, you will find explanations in Spanish [assistant test administrator point to the Spanish glosses].

All test booklets contained a sample question. The test administrator asked students to read the sample question silently and to circle the correct answer. The sample question, not related to science, was used to make sure that students understood the correct response format (i.e., circling as opposed to darkening or “X-ing” in the correct response). (For the complete test administrator script, see Appendix E.)

Students were given 25 minutes to complete the 20-item science test, 3 minutes to complete the follow-up questionnaire, and 8 minutes to complete the 25-item science background questionnaire.

Each teacher was asked to complete the science teacher questionnaire (see Appendix D), and the test administrators completed the test administrator feedback questionnaire (see Appendix F).

Analysis

Student science test scores were compared to investigate (a) the validity of the accommodations and (b) the possible differential impact of accommodations on groups of students with different language backgrounds.

Results

This section provides findings on the impact of the accommodations and a discussion of findings in regard to the follow-up and background questionnaires.

Accommodation Results

The main research question in this study was whether or not accommodations addressing the difficulty of English vocabulary in test items reduce the performance gap between limited English proficient students and proficient speakers of English in content-based areas such as science. A sample of 422 students were tested under accommodated and non-accommodated conditions. To examine the validity of accommodated assessments, proficient speakers of English (non-LEP students), who do not normally receive any form of accommodation, were also included in this study. The non-LEP students were tested under both accommodated and unaccommodated conditions.

Twenty science test items were selected from the 1996 NAEP secured science main assessment items. Two counterbalanced booklets were formed using the same items but in different order (form A and form B; see description in the Instruments section above). The two forms were randomly assigned to students under different accommodation conditions. Fifty five percent received form A and 45% received form B. Students' performance under the two forms was compared for any significant form effect. No significant difference was found between the scores for the two forms (t = -1.38, df = 420, p = .169); therefore, scores from both forms were treated equally.

We now turn to the findings concerning the performance gap between LEP and non-LEP students. We compared the performance of LEP and non-LEP students under the three accommodation conditions (dictionary, glossary, and standard condition). Table 3 presents the mean, standard deviation and number of students for each group of LEP / non-LEP students by accommodation condition.

There was a large performance gap between LEP and non-LEP students. Consistent with the literature, LEP students performed substantially lower than non-LEP students. For LEP students, the mean score was 8.97 (*SD* = 4.40, *n* = 183) and for non-LEP students the mean was 11.67 (*SD* = 3.68, *n* = 236), a difference of about two thirds of a standard deviation.

We tested the level of significance of the differences between the means reported in Table 3. A two-factor ANOVA model was applied. Factor A was students' LEP status (LEP versus non-LEP) and Factor B was assessment conditions (dictionary versus glossary versus standard condition). Factor A main effect (difference between performance of LEP and non-LEP students) was significant (*F* = 46.40, *df* = 1, 413, *p* = .000), suggesting that LEP students in general performed lower

Table 3

Means, Standard Deviations, and Numbers of Students by LEP Status and Accommodation Conditions

LEP status / accommodation condition[a]	Original	Dictionary	Glossary	Total
LEP				
M	8.36	10.18	8.51	8.97
SD	4.40	5.26	4.72	4.40
N	58	55	70	183
Non-LEP				
M	11.71	11.37	11.96	11.67
SD	3.39	3.79	3.86	3.68
N	79	82	75	236
Total				
M	10.29	10.86	10.34	10.50
SD	3.48	4.46	4.61	4.22
N	137	138	147	419

[a] Data were not available for 3 students.

than non-LEP students, a finding that was discussed earlier. Factor B main effect (performance under different testing conditions) was not significant for the overall group (*F* = 0.66, *df* = 2, 413, *p* = .515). The interaction between A (LEP status) and B (testing condition), however, was significant (*F* = 3.43, *df* = 2, 413, *p* = .033). This significant interaction suggests that LEP and non-LEP students performed differently under different testing conditions.

However, the main hypothesis in this study dealt with the effectiveness of accommodation in reducing the performance gap between LEP and non-LEP students. To test this hypothesis, we compared the performance of LEP students under the three testing conditions.

To test the hypothesis of effectiveness of accommodation in reducing the performance gap between LEP and non-LEP students, we compared LEP students' scores on science items under the three accommodation conditions: customized dictionary, glossary, and standard NAEP conditions. LEP students performed higher under the accommodated conditions than under the standard condition. The mean for LEP students under the customized dictionary condition was 10.18 (*SD* = 5.26, *n* = 55); under the glossary condition, the mean was 8.51 (*SD* = 4.72, *n* = 70); and under the standard condition, the mean was 8.36 (*SD* = 4.40, *n* = 58). As the data suggest, LEP students did particularly well under the customized dictionary condition. The results of an analysis of variance (ANOVA) indicated that the difference between means for LEP students under the three accommodation conditions was significant (*F* = 3.08, *df* = 2, 180, *p* = .048). The results of multiple comparison tests suggested that the performance of LEP students under the dictionary condition was significantly higher than the performance of LEP students under the standard condition. However, when the performance of LEP students under the glossary condition was compared with the performance of LEP students under the standard condition, the difference was not significant.

Abedi, Hofstetter, Lord, and Baker (1998) demonstrated that the translation of English-language assessment tools into students' native language may not help if the language of instruction is English. To test the hypothesis of effectiveness of a Spanish glossary in reducing the gap for LEP students with Hispanic language background, we compared the mean science score of Hispanic students across the three accommodation conditions (dictionary, glossary, and original). The mean science score for Hispanic LEP students tested with the original booklet was 8.21 (*SD* = 4.27, *n* = 53). The mean for Hispanic LEP students utilizing the dictionary accommodation was 10.28 (*SD* = 5.25, *n* = 46). Under the glossary condition, the Hispanic LEP student mean was 8.03, (*SD* = 4.41, *n* = 59).

The results of an analysis of variance indicated that the difference between the mean scores under the three accommodation conditions was significant (*F* = 4.40, *df* = 2, 155, *p* = .01). This difference was mainly between the dictionary condition and the others; the mean performance for the glossary condition was almost identical to the mean for the standard condition. These results confirmed the earlier findings by Abedi, Hofstetter, Lord, and Baker (1998) that translating an instrument or providing a glossary in students' native language may not help if the language of instruction is English. However, an English dictionary may be more effective in reducing the science performance gap between LEP and non-LEP students, since it may help students with the language factors in assessment.

The validity of the accommodations in this study was tested by comparing the performance of non-LEP students across the accommodation conditions. Accommodations should not affect the performance of non-LEP students. That is, there should not be any significant differences between the performance of non-LEP students tested under the accommodated condition(s) and that of non-LEP students tested under the standard condition, with no accommodation. The results of analyses suggested that the accommodations had no significant effect on the scores of the non-LEP students. For non-LEP students, the mean science score for the dictionary accommodation was 11.37 (*SD* = 3.79, *n* = 82); for the glossary condition, the mean was 11.96 (*SD* = 3.86, *n* = 75); and for the standard condition, the mean was 11.71 (*SD* = 3.39, *n* = 79). The results of an analysis of variance showed no significant difference between the performance of non-LEP students under the three accommodation conditions (*F* = .495, *df* = 2, 233, *p* = .610).

The non-significant results indicate that the accommodation strategies used in this study did not affect the outcome of measurement. Thus, concerns over the validity of accommodations may not be warranted.

Classroom effects. To examine the effects of the multilevel structure of data (students nested in classrooms), a 2-level hierarchical model was used in the analysis. The sources of educational influence on students occur in the context of classrooms, which give rise to multilevel data (Burstein, 1993). Using hierarchical linear models, the effects of different accommodations for LEP and non-LEP students were examined in detail. The 2-level model included the student-level variables in level 1, represented by Figure 3. Figure 4 represents the differences across classrooms examined in level 2.

The preliminary results of the analysis are presented in Table 4. The differences in the science performance mean across classes are statistically significant (p = .000). However, as discussed in the Methodology section, to control for teacher and class effects, test booklets were randomly assigned to students within a classroom. The significance of classroom effect in this case may be a result of the small *n* in this pilot study. Randomization may not be effective when *n* size is small. Given the significance of the variance, the classroom differences are an important factor to consider in the model.

After adding the estimation of classroom differences in the model, LEP status and the reading achievement score on the SAT-9 were determined as strong predictors of science performance. The results indicated that the LEP students on average performed about 3 points higher than the non-LEP students, after controlling for differences in reading performance.

$Yij = \beta oj + \beta 1j(\text{LEP}) + \beta 2j\ (\text{Reading Score}) + \beta 3j\ (\text{Dictionary}) + \beta 4j\ (\text{Glossary})$

$+\ \beta 5j\ (\text{LEP*Dictionary}) + \beta 6j\ (\text{LEP*Glossary}) + rij \quad r(N, \sigma^2)$

where

Yij – individual outcome

βoj – the class mean

$\beta 1j$ – the effect of LEP compared to non-LEP students

$\beta 2j$ – the effect of reading score on SAT-9 (covariate)

$\beta 3j$ – the effect of Dictionary compared to Standard test booklet

$\beta 4j$ – the effect of Glossary compared to Standard test booklet

$\beta 5j$ – the effect of Dictionary accommodation for LEP students

$\beta 6j$ – the effect of Glossary accommodation for LEP students

rij – the error associated with the level 1 model

Figure 3. Level-1 model.

$\beta oj = \gamma 00 + \mu 0j$

$\beta 1j = \gamma 10$

$\beta 2j = \gamma 20$

$\beta 3j = \gamma 30$

$\beta 4j = \gamma 40$

$\beta 5j = \gamma 50$

$\beta 6j = \gamma 60$

where

$\gamma 00$ – the overall mean across classes

$\gamma 10$ – the mean for LEP students

$\gamma 20$ – the mean of reading scores

$\gamma 30$ – the mean for non-LEP with dictionary accommodation

$\gamma 40$ – the mean for non-LEP with glossary accommodation

$\gamma 50$ – the mean for LEP students with dictionary accommodation

$\gamma 60$ – the mean for LEP students with glossary accommodation

$\mu 0j$ – the error associated with βoj (the variability of classrooms)

Figure 4. Level-2 model.

Table 4

Examination of Science Test Performance Using a Hierarchical Linear Model

Fixed effects	Coefficient	*SE*	*T* ratio	
Mean across classes	6.699	1.151	5.821	
Mean reading score, non-LEP students	3.295	0.734	4.487	
Mean with dictionary accommodation, non-LEP students	0.049	0.007	6.506	
Mean with glossary accommodation, non-LEP students	-0.132	0.510	-0.259	
Mean for LEP students with dictionary accommodation	-0.082	0.517	-0.159	
Mean for LEP students with glossary accommodation	1.149	0.799	1.438	
Mean for LEP students on reading	-0.94	0.776	-1.21	
Random effects	Variance component	*df*	χ^2	*p* value
Mean across classes	10.332	9	368.342	0.000
Level-1 error	9.914			

The dictionary and the glossary accommodations had no significant effect on the performance of non-LEP students. However, the results suggested that the use of a dictionary may help LEP students. Even though the *p*-value does not hold any statistical significance, there is some evidence for a positive accommodation effect for LEP students. This finding is consistent with our results derived using analysis of variance. This preliminary analysis suggests that the use of a customized dictionary as an accommodation may contribute to validity for LEP students in large-scale assessments.

Follow-Up Questionnaires

As indicated earlier, we used three different test booklets:

1. a booklet with original versions of the test items, with no dictionary or glossary;

2. a booklet with a customized dictionary attached;

3. a booklet with glossary of terms (Spanish translations and brief English glosses in the page margins).

For each of these booklets, a follow-up questionnaire was developed to gather feedback from students on the language of the test items and the level of utilization and usefulness of the accommodations they received (none, dictionary, and glossary). Different booklets had different sets of follow-up questions. For example, the questionnaire in both the non-accommodated and the dictionary-accommodated test booklets consisted of 1 open-ended question and 5 close-ended questions, whereas the questionnaire in the glossary-accommodated test booklet consisted of 1 open-ended question and 8 close-ended questions. See Appendix A for the follow-up questionnaires. Numbers were assigned to Likert scale options as follows: 1 = *no/never*; 2 = *yes, some/sometimes/maybe;* and 3 = *yes, many/a lot/yes, definitely*.

Follow-up questions, original booklet. To examine the pattern of responses across the LEP categories (comparing responses of LEP and non-LEP students), frequencies of responses to the follow-up questions were obtained for the two groups. Table 5 presents the frequency of responses for each of the five Likert-type questions for LEP and non-LEP students using the original (non-accommodated) test booklet. The first question asks, "In the science test, were there words that you did not understand?" Response options for this question range from *no* to *yes, some/maybe* to *yes, many/a lot*. Values of 1 to 3 were assigned to the three response options respectively.

Table 5

Frequency Distribution of the Responses to the Follow-Up Questions for the Original (Non-Accommodated) Booklet

	No		Yes, some / maybe		Yes, many / a lot	
Question	Non-LEP	LEP	Non-LEP	LEP	Non-LEP	LEP
1. In the science test, were there words that you did not understand?	24 30.4%	8 13.8%	51 64.6%	45 77.6%	3 3.8%	3 5.2%
2. Would it help if the test explained some of the more difficult words?	19 24.1%	5 8.6%	49 62.0%	38 65.5%	11 13.9%	13 22.4%
3. Would you like to be able to use a dictionary during a test like this?	19 24.1%	6 10.3%	55 69.6%	32 55.2%	4 5.1%	17 29.3%
4. If you had a dictionary to use during the test, how much would you use it?	15 19.0%	5 8.6%	37 46.8%	41 70.7%	26 32.9%	10 17.2%
5. Would it help if the test explained words in another language?	68 86.1%	31 53.4%	10 12.7%	19 32.8%	0 0.0%	6 10.3%

Of 134 total respondents, 24 (30.4%) non-LEP students responded “No” to question 1, indicating that there were no words in the science test that they did not understand. However, only 8 (13.8%) LEP students responded “No” to this question. The large gap between LEP and non-LEP students on this question suggests that LEP students perceived the vocabulary of science test items as more difficult than the non-LEP group did. A larger percentage of LEP students (77.6%) than non-LEP students (64.6%) also indicated that they had some difficulty understanding the science questions. Also, as expected, a smaller percentage of non-LEP students indicated that they found many words in the science test that they did not understand. Of the non-LEP students, 3.8% selected this option as compared with 5.2% for LEP students.

Follow-up question 2 asks whether it would help if the test explained some of the more difficult words. A higher percentage of LEP students indicated that it would. Of the total 134 respondents, 24 indicated that explanation of difficult words would not be helpful. Of these 24, 19 respondents were non-LEP students (21.1% of the non-LEP group), and only 5 were LEP students (8.6% of the LEP group). However, there was an opposite trend of response in the highest category, *yes, many*. More LEP students indicated that it would help if the test explained some of the more difficult words (22.4% for the LEP group as compared with 13.9% for the non-LEP group).

The same trend can be seen for follow-up questions 3 and 4, which ask about use of a dictionary. More LEP students indicated that they would like to be able to use a dictionary and they would use it if they had one. Similarly, more LEP students indicated that it would help if the test explained words in another language.

To compare the response patterns of LEP and non-LEP students on these follow-up questions, we created an average rating for each question by assigning numbers (rank) to the responses (1 = *no/never*, 2 = *yes, some/maybe*, and 3 = *yes, many/yes, a lot*).

Table 6 presents means and standard deviations for the ranks by LEP and non-LEP groups for the original (non-accommodated) booklet. Mean ranks for LEP students are higher for all questions, except question 4, suggesting that LEP students in general would prefer more assistance. The mean rating for question 1 ("In the science test, were there words that you did not understand?") for non-LEP students was 1.73 (*SD* = .53) as compared with a mean of 1.91 (*SD* = .44) for LEP students. For question 2 ("Would it help if the test explained some of the more difficult words?"), the mean for non-LEP students was 1.90 (*SD* = .61) as compared with a mean of 2.14 (*SD* = .55) for LEP students.

Table 6

Means and Standard Deviations of Ranks for Responses to Follow-Up Questions, Original Booklet

	Non-LEP students			LEP students		
Questions	Mean	*SD*	*N*	Mean	*SD*	*N*
1. In the science test, were there words that you did not understand?	1.73	.53	78	1.91	.44	56
2. Would it help if the test explained some of the more difficult words?	1.90	.61	79	2.14	.55	56
3. Would you like to be able to use a dictionary during a test like this?	1.81	.51	78	2.20	.62	55
4. If you had a dictionary to use during the test, how much would you use it?	2.14	.72	78	2.09	.51	56
5. Would it help if the test explained words in another language?	1.13	.34	78	1.55	.68	56

Note. Responses: 1 = *no*, 2 = *yes, some/maybe*, 3 = *yes, many/a lot*.

We compared the response patterns of LEP and non-LEP students on all five follow-up questions in the original booklet using multivariate analysis of variance (MANOVA). In this MANOVA model, the Likert-type scores of the five follow-up questions were used as the dependent variable, and students' LEP status (LEP / non-LEP) was used as the independent variable. Table 7 summarizes the results of this multivariate analysis. The multivariate test was significant (Wilks' $\lambda = .75$, $F = 8.22$, $p < .01$) indicating that LEP and non-LEP students responded differently to the set of follow-up questions. The univariate *F*-test, however, suggested that of the five questions, four elicited different responses from the two groups, but question 4 had the same response pattern across the two groups (LEP / non-LEP). The responses to question 4 indicated that many of the non-LEP students, as well as the LEP students, would use a dictionary.

Follow-up questions, dictionary booklet. The purpose of follow-up questions in the dictionary booklet was to find out whether students used the customized dictionary. However, questions similar to those in the original booklet were also asked of students who received the dictionary booklet. Table 8 presents the summary results of analyses on the dictionary booklet follow-up questions.

The trend of frequency distributions in Table 8 for the dictionary booklet is similar to that reported in Table 5 for the original-version booklet. LEP students indicated that there were more words in the science test that they did not understand, in comparison to the non-LEP students. LEP students, more than their non-LEP counterparts, thought that it would help if the test explained words in another language, and that it would help if the test used easier words. However, LEP and non-LEP students gave similar responses when they were asked if they

Table 7

Multivariate ANOVA Results for Response Patterns for Follow-Up Questions, Original Booklet

	SS		*df*		*MS*			
Variable	Hypo.	Error	Hypo.	Error	Hypo.	Error	*F*	*P*
Question 1	1.10	31.74	1	129	1.10	.25	4.46	.037
Question 2	2.01	43.99	1	129	2.01	.34	5.88	.017
Question 3	4.39	39.58	1	129	4.39	.31	14.32	.000
Question 4	0.053	53.23	1	129	0.053	.41	0.128	.721
Question 5	5.96	34.21	1	129	5.96	.27	22.46	.000

Note. *SS* = Sum of Squares. *MS* = Mean Squares.

Table 8

Frequency Distribution of the Responses to Follow-Up Questions, Dictionary Booklet

	No		Yes, some / maybe		Yes, many / a lot	
Questions	Non-LEP	LEP	Non-LEP	LEP	Non-LEP	LEP
1. In the science test, were there words that you did not understand?	27 32.9%	11 20.0%	52 63.4%	40 72.7%	1 1.2%	3 5.5%
2. Did you look up words in the dictionary?	34 41.5%	23 41.8%	43 52.4%	30 54.5%	3 3.7%	1 1.8%
3. Did the dictionary help you understand the questions?	32 39.0%	22 40.0%	23 28.0%	14 25.5%	24 29.3%	17 30.9%
4. Would it help if the test explained words in another language?	64 78.0%	23 41.8%	12 14.6%	26 47.3%	2 2.4%	5 9.1%
5. Would it help if the test used easier words?	24 29.3%	2 3.6%	39 47.6%	19 34.5%	16 19.5%	18 32.7%

looked up words in the dictionary. Both groups also found the dictionary helpful in understanding the questions.

Table 9 reports means, standard deviations, and numbers of students responding to the dictionary questions. Mean Likert-scale scores for questions 2 and 3 (concerning using the dictionary and whether or not the dictionary was helpful) were almost identical for LEP and non-LEP students, but for questions 1, 4, and 5, the means were very different. The results of multivariate analysis of variance (MANOVA) comparing LEP and non-LEP students on the five dictionary follow-up questions confirm our earlier statement that LEP and non-LEP students responded differently on questions 1, 4, and 5.

Follow-up questions, glossary booklet. The glossary follow-up questionnaire contained 8 Likert-type items and 1 open-ended question. In addition to the questions that were asked in the original booklet and dictionary booklet questionnaires (such as “Were there words that you did not understand?” and “Would it help if the test used easier words?”), there were questions specifically related to the use of the glossary. Table 10 presents frequencies and percentages of students’ responses to the glossary follow-up questions.

Table 9

Means and Standard Deviations of Ranks for the Responses to the Follow-Up Questions, Dictionary Booklet

	Non-LEP students			LEP students		
Questions	Mean	*SD*	*N*	Mean	*SD*	*N*
1. In the science test, were there words that you did not understand?	1.68	.50	80	1.85	.49	54
2. Did you look up words in the dictionary?	1.61	.56	80	1.59	.53	54
3. Did the dictionary help you understand the questions?	1.90	.84	79	1.90	.86	53
4. Would it help if the test explained words in another language?	1.21	.47	78	1.67	.64	54
5. Would it help if the test used easier words?	1.90	.71	79	2.41	.59	39

Note. Responses: 1 = *no*, 2 = *yes, some/maybe*, 3 = *yes, many/a lot*.

Table 10

Frequency Distribution of the Responses to the Follow-Up Questions for the Glossary Booklet

	No		Yes, some / maybe		Yes, many / a lot	
Questions	Non-LEP	LEP	Non-LEP	LEP	Non-LEP	LEP
1. In the science test, were there words that you did not understand?	35 47.7%	10 14.3%	36 48.0%	51 72.9%	3 4.0%	7 10.0%
2. Did you read the explanation in the margins in English (on the right side of the page)?	18 24.0%	9 12.9%	52 69.0	37 52.0%	4 5.3%	20 28.6%
3. Did the English explanations help you understand the questions?	18 24.0%	5 7.1%	40 53.3%	31 44.3%	16 21.3%	32 45.7%
4. Did you read the explanation in the margins in Spanish (on the left side of the page)?	67 89.3%	43 61.4%	6 8.0%	18 25.7%	1 1.3%	5 7.1%
5. Did the Spanish explanations help you understand the questions?	70 93.3%	34 48.6%	4 5.3%	23 32.9%	0 0.0%	1 1.4%
6. Would you like to be able to use a dictionary during a test like this?	22 29.3%	8 11.4%	35 46.7%	27 38.6%	15 20.0%	32 45.7%
7. If you had a dictionary to use during the test, how much would you use it?	19 25.3%	1 1.4%	46 61.3%	47 67.1%	7 9.3%	19 27.1%
8. Would it help if the test used easier words?	20 26.7%	3 4.3%	36 48.0%	18 25.7%	17 22.7%	18 25.7%
9. What else would make it easier for you to understand the questions on the test?						

The trend of responses in Table 10 is similar to the trend reported in Table 5 and Table 8 for the original and dictionary booklets. LEP students, more than their non-LEP counterparts, indicated that there were words that they did not understand. The LEP group also indicated, more than the non-LEP group, that it would help if the test used easier words.

Responses given by LEP students were different from those of non-LEP students. LEP students, more than non-LEP students, indicated that they read the explanations in the margin (the glossary). More LEP students responded that the English and Spanish explanations helped them understand the questions. In response to the question "Would you like to be able to use a dictionary during a test like this?" 29.3% of non-LEP students said "No," they would not like to use a dictionary as compared with 11.4% of LEP students who said that they would not. On the other hand, 20% of non-LEP students said "Yes," they would like to use a dictionary, as compared with 45.7% of LEP students.

Table 11 presents means, standard deviations and numbers of students responding to the glossary follow-up questions. Similar to the means reported in Table 6 and Table 9, the trend of higher means for LEP students is evident from the data in Table 11.

Table 12 reports the results of multivariate analysis of variance for the eight questions in the glossary follow-up questionnaire, comparing mean Likert scores of LEP and non-LEP students. As the data in Table 12 suggest, in all eight questions the differences in mean between LEP and non-LEP were significant.

Different follow-up questionnaires were used for the three testing groups: the original, the dictionary, and the glossary groups. Some of the questions were identical across the three groups, and other questions were similar. The similarity of the follow-up questions across the three testing groups may warrant the following general conclusions. However, the follow-up questions were not significantly related to the science test scores.

In general, the responses provided by LEP students imply that they had more difficulty with the language of test items than the non-LEP students had. For example:

Table 11

Means and Standard Deviations of Ranks for the Follow-Up Questions, Glossary Booklet

	Non-LEP students			LEP students		
Questions	Mean	*SD*	*N*	Mean	*SD*	*N*
1. In the science test, were there words that you did not understand?	1.57	.58	74	1.96	.50	68
2. Did you read the explanation in the margins in English (on the right side of the page)?	1.81	.51	74	2.17	.65	66
3. Did the English explanations help you understand the questions?	1.97	.68	74	2.40	.63	68
4. Did you read the explanation in the margins in Spanish (on the left side of the page)?	1.11	.36	74	1.42	.63	66
5. Did the Spanish explanations help you understand the questions?	1.05	.23	74	1.66	.74	68
6. Would you like to be able to use a dictionary during a test like this?	1.90	.72	72	2.36	.69	67
7. If you had a dictionary to use during the test, how much would you use it?	1.83	.58	72	2.27	.48	67
8. Would it help if the test used easier words?	1.96	.72	73	2.38	.63	39

Note. Responses: 1 = *no*, 2 = *yes, some/maybe*, 3 = *yes, many/a lot*.

Table 12

Multivariate ANOVA Results for Follow-Up Questions, Glossary Booklet

	SS		*df*		*MS*			
Variable	Hypo.	Error	Hypo.	Error	Hypo.	Error	*F*	*P*
Question 1	4.70	39.91	1	104	4.70	.29	16.00	.000
Question 2	3.73	45.09	1	104	3.73	.33	11.26	.001
Question 3	5.55	58.94	1	104	5.55	.43	12.80	.000
Question 4	3.38	34.74	1	104	3.38	.26	13.22	.000
Question 5	11.81	38.52	1	104	11.81	.28	41.69	.000
Question 6	3.03	54.04	1	104	3.03	.39	7.62	.007
Question 7	5.39	22.00	1	104	5.39	.16	33.32	.000
Question 8	4.57	23.66	1	104	4.57	.17	26.26	.000

Note. SS = Sum of Squares. *MS* = Mean Squares.

1. More LEP than non-LEP students indicated that in the science test, there were words that they did not understand.

2. More LEP than non-LEP students wanted explanation of some of the difficult words.

3. More LEP than non-LEP students expressed interest in using a dictionary during the test.

4. More LEP than non-LEP students indicated that it would help them if the test explained words in another language.

5. More LEP than non-LEP students expressed a preference for a dictionary during the test.

Background Questionnaire

As indicated in the Methodology section, along with the science test and the follow-up questionnaire, a background questionnaire was included in the test booklet. The background questionnaire consists of 35 questions. These questions can be categorized as follows:

1. *Demographic questions*: Questions 1–5 are demographic questions about country of origin, length of time in the U.S., gender, zip code, and ethnicity.

2. *A language other than English*: Questions 6–14 ask students whether they use a language other than English at home and with relatives, and if they do, how proficient they are with that language.

3. *Studied a subject in other languages*: Questions 15–18 ask students whether they studied science or any other subjects in a language other than English.

4. *Self-reported English proficiency*: Questions 19–22 ask students to self-report their level of English proficiency (understanding, speaking, reading, and writing).

5. *Home environment*: Questions 23–27 ask about home environment; for example, whether there are newspapers, books, and encyclopedias in English in the home, and number of hours of television viewing.

6. *School and interest*: Questions 28–29 ask about school changes and plans for future schooling, and questions 30-31 ask about students' interest in science.

7. *Self-reported grades*: Questions 32–34 ask students to self-report their grades in school.

Results of analyses of background questions. Some of the background questions may not be directly related to the main hypotheses of this study discussed earlier; however, they provide useful information. We report the results of analyses of the background questionnaire using the categories described above.

Demographic questions. The findings of previous studies suggest that length of time in the U.S. is one of the strong predictors of school achievement for LEP students. To examine replicability of this finding in our study, we computed the correlation between students' performance in science and the length of time that students have been in the U.S. This correlation was .0865 ($p > .05$), which is not statistically significant.

Whether there was a gender effect on scores was another interesting research question that we could address using the background questions. Performance of male and female students in science was compared. The mean science score for the male students was 10.61 (*SD* = 4.25, *n* = 222) and for females, 10.40 (*SD* = 4.18, *n* = 199). A *t*-test result of .50 (*df* = 419, *p* = .617) indicates that the difference between the mean scores of male and female students is not statistically significant.

A language other than English. Students were asked whether they speak a language besides English at home. We compared the performance of students who responded "Yes" to this question with the performance of those who responded "No." The mean science score for those responding "Yes" was 9.99 (*SD* = 4.20, *n* = 307). The mean for those responding "No" was 12.54 (*SD* = 3.50, *n* = 94), a difference of about two thirds of a standard deviation. This difference between the performance of those students who spoke a language other than English at home and those who spoke only English at home is statistically significant (*t* = 5.34, *df* = 399, *p* = 0.00).

Questions 7 to 10 asked students how much they speak that language with others (parents, brothers and sisters, friends at school, and outside). Since these four questions are about the use of the other language, we created a composite variable of the questions that asked "How much do you speak that language with . . .". The questions have three response categories, *always or most of the time, sometimes,* and *never or hardly ever."* We assigned 1 to *always or most of the time,* 2 to *sometimes,* and 3 to *never or hardly ever.* Thus, values for the composite variable range from 4 (always or most of the time speaks the language with others) to 12 (never or hardly ever).

Table 13 shows means and standard deviations for the four questions on the use of a language other than English.

Table 13

Means, Standard Deviations, and Numbers of Respondents for the Four Questions (7-10) About the Use of a Language Other Than English

Variable	Mean	*SD*	*N*
Q7, Parents	1.55	.66	317
Q8, Brothers / sisters	2.05	.73	313
Q9, Friends at school	2.30	.78	315
Q10, Outside	2.27	.77	315
Composite	8.04	2.34	320

Since *always or most of the time* was coded as 1 and *never* as 3, the larger the mean for these four questions, the less the language is spoken with others. As Table 13 shows, the mean for question 7 (*M* = 1.55, *SD* = .66) is smaller that the means for the other questions. This question asks students how much they speak that language with their parents. The small mean for this question (as compared with the means for the other questions) suggests that students spoke that language more with their parents than with brothers and sisters or friends.

These four questions (Q7 to Q10) were answered mainly by the non-native English speaking students; therefore, comparison across LEP groups (LEP versus non-LEP) was not meaningful. However, we examined the relationship between this composite variable (use of a language other than English) and students' performance in science. A PM correlation of .238 significant beyond .01 nominal level suggests that there is a relationship between speaking a language other than English and performance in science. The composite variable is a proxy of students' LEP status; therefore, this finding is consistent with our earlier finding that LEP students performed at a significantly lower level in science than non-LEP students.

Questions 11 to 14 asked students to self-report their proficiency level in the language other than English that they use. The format (response options) of these questions was similar to the format of questions on the use of another language, discussed earlier. A value of 1 was assigned to *very well,* 2 to *fairly well,* and 3 to *not very well.*

Table 14 presents means, standard deviations, and numbers of respondents for questions 11 to 14. As data in Table 14 suggest, students had more difficulty

Table 14

Means, Standard Deviations, and Numbers of Respondents for the Four Questions (11-14) About Level of Proficiency in the Language Other Than English

Variable	Mean	*SD*	*N*
Q11, Speak	1.59	.63	311
Q12, Understand	1.43	.61	309
Q13, Read	1.97	.80	311
Q14, Write	2.00	.78	310
Composite	6.91	2.34	314

with writing (*M* = 2.00, *SD* = .78) and reading (*M* = 1.97, *SD* = .80) and less difficulty with understanding (*M* = 1.43, *SD* = .61) and speaking (*M* = 1.59, *SD* = .63).

A composite variable consisting of all self-reported first language proficiency was created. The mean for this variable (see Table 14) is 6.91 (*SD* = 2.34), which is higher than the midpoint of 6.00 (maximum score is 12; 3 points for each question by 4 questions). This higher-than-midpoint mean suggests that students believed they had difficulty in the language that they spoke mainly with their parents and sometimes with their other family members and friends. A PM correlation coefficient of .189, significant beyond the .01 nominal level, suggests that a relationship exists between proficiency in the first language and students' performance in science. This relationship, although not very strong (only 3.6% of the variance of joint distribution), is in the opposite direction. That is, the more proficient the student claimed to be in his/her first language, the lower the level of science performance he/she demonstrated.

Self-reported English proficiency. Questions 19 to 22 asked students to self-report their level of English language proficiency. The format (response options) of these questions was similar to the format of the questions on self-reported proficiency on the first language, discussed earlier. A value of 1 was assigned to *very well*, 2 to *fairly well*, and 3 to *not very well.*

Table 15 reports means, standard deviations, and numbers of respondents for questions 19 to 22. As Table 15 shows, students self-reported relatively high levels of English proficiency, higher than the levels of proficiency for the first language (by

Table 15

Means, Standard Deviations, and Numbers of Respondents for the Four Questions (19-22) About Level of English Language Proficiency

Variable	Mean	*SD*	*N*
Q19, Understand	1.20	.44	405
Q20, Speak	1.23	.45	412
Q21, Read	1.31	.50	408
Q22, Write	1.38	.54	408
Composite	5.07	1.56	412

those who spoke a language other than English). However, compared with the means for self-reported proficiency in understanding (*M* = 1.20, *SD* = .44) and speaking English (*M* = 1.23, *SD* = .45), the means for reading (*M* = 1.31, *SD* = .50) and writing (*M* = 1.38, *SD* = .54) were higher, suggesting more difficulty in these two areas of language.

A PM correlation coefficient of -.255 (6.5% of the variance), significant beyond the .01 nominal level, suggests that a relationship exists between students' level of language proficiency and their scores on the science test. Unlike the correlation reported earlier for the self-reported proficiency in a language other than English, the relationship is in the expected direction. That is, the higher the level of language proficiency, the higher a student's performance in science.

Discussion

Research Hypothesis and Findings

The main hypothesis of this study concerned the effectiveness of accommodations. That is, how effective were the accommodation strategies that were used in the study? As reported in the Results section, overall, the provision of accommodations did not impact students' performance. For all students, mean scores of 10.29 for the original version of the test, 10.86 for the dictionary version, and 10.34 for the glossary version suggest that accommodations did not have any sizable impact on students' performance in general. The provision of accommodations did not impact the performance of non-LEP students. Mean scores of 11.71 for the unaccommodated (original) test, 11.37 for the test-plus-dictionary version, and 11.96 for the test-plus-glossary version indicate that accommodations did not have a sizable impact on their performance. However, when the performance of students under accommodated and non-accommodated assessments was compared across the students' LEP status, interesting trends were apparent.

A comparison of LEP students' performance on the accommodated tests with their performance on the unaccommodated test revealed that the accommodations actually contributed to improved performance of LEP students. LEP students who were provided the customized dictionary performed significantly better than those LEP students assessed under the standard NAEP condition. Providing the definitions of words in a glossary format also helped LEP students, but the effect did not reach a level of statistical significance.

Addendum

Both accommodations focused on potentially difficult vocabulary. However, only the dictionary accommodation resulted in significantly higher scores for LEP students. An interesting question is why the glossary accommodation did not show similar results. There are a number of possible reasons, which we are currently exploring:

- Did students find it easier to use the dictionary than the glossary? Did they use the dictionary more?

- In the glossary version of the test booklet, inclusion of Spanish translations and English glosses made the pages rather busy visually; did this divert the students' attention from the science questions?

- Did the glossary leave out important words? The dictionary included more words per item than the glossary version, and the words for the glossary that were selected by researchers may not have been the words that the students actually looked up in the dictionary.

- Was the dictionary more informative than the glossary? The dictionary definitions were longer than the corresponding glosses; students may have found them more helpful.

A dictionary is, in a sense, a mini-encyclopedia. Since the dictionary included all content words, an important question is whether the dictionary provided content-area information that helped students answer the science questions. We are reviewing items and definitions to determine this. However, the fact that the dictionary definitions did not help non-LEP students is strong evidence that the accommodation did not provide content information.

The second hypothesis, a major concern in any accommodation study, questioned the validity of accommodation. The results of this study clearly indicate that a customized dictionary helped LEP students. The question remaining is whether the accommodation

a. reduced the performance gap between LEP and non-LEP students;

b. increased the performance gap between LEP and non-LEP students;

c. increased the performance of all students.

To address this validity concern, we compared the accommodated and unaccommodated performance of non-LEP students. If a given accommodation strategy affected the construct under measurement, then the accommodated non-LEP students should have performed significantly better than the non-accommodated non-LEP students. The results of our analyses indicated that the accommodations did not affect the performance of non-LEP students. The means of non-LEP student groups across the three accommodation conditions (original, dictionary, and glossary) were not significantly different.

The results of this study suggest that, among the two accommodation strategies that were used, the customized dictionary was effective in reducing the gap between LEP and non-LEP scores. The accommodation did not affect the validity of the assessment. The results also show that, once variability in reading performance was taken into account, the LEP students outperformed the non-LEP students in science. This is consistent with previous findings, which show a strong correlation between language proficiency and academic performance.

Follow-Up Questions

As discussed in the Methodology and Results sections of this report, students were asked to respond to a set of follow-up questions and a set of background questions. The purpose of the follow-up questions was to see whether students who received accommodations found them useful and how much students actually used the accommodations (for example, how often they referred to the dictionary and how much they used the glossary). The analyses of the follow-up questions show that more LEP students than non-LEP students reported that they actually utilized the accommodations and that the dictionary and glossary were useful.

Background Questions

Student background information includes factors such as community, school, home, and individual characteristics that impact students in academic settings (Butler & Stevens, 1997). It is well documented that some components found in the science background questionnaire of this study play an important role in academic performance (De Avila et al., 1978; Heath, 1983; Thurlow et al., 1998). Gonzales et al. (1996) emphasized the importance of factoring in linguistic and cultural characteristics for assessments in order for them to be valid. This study analyzed the relationship between those background characteristics and the use of accommodations in an evaluation.

The background questionnaire used for this study included 35 questions, categorized as follows, for analyses:

1. demographic questions
2. a language other than English
3. studied a subject in other languages
4. self-reported English proficiency
5. home environment
6. school and interest
7. self-reported grade points

Students' responses to the background questions provided data for testing additional hypotheses concerning the impact of students' background variables, including their language background variables, in relation to their performance. Our analyses showed no significant gender differences. However, a large significant difference was found between the performance of students who spoke only English in the home and those who spoke a language other than English in the home. Students who spoke a language other than English performed significantly lower than those who did not. This finding is consistent with the literature and the findings that are reported earlier in this paper; because students who spoke a language other than English were mainly LEP students, their performance was lower than the monolingual English-speaking students (non-LEP). Of the total number of LEP students participating in this study, 96% spoke a language other than English.

Self-reported data on the level of first and second language proficiency also provided useful information. Students who spoke a language other than English in the home indicated that they spoke that language more with their parents and less with their brothers, sisters, and friends. These findings, consistent with the existing literature, reflect a generation gap and suggest that older family members may not have sufficient English language facility to communicate comfortably with their children in English. The children, therefore, find it necessary to use their native language when communicating with their parents and grandparents.

The results of analyses on the self-reported data about English proficiency were also consistent with the literature and with the earlier findings of this study. LEP students reported significantly lower proficiency in English than their non-LEP counterparts.

Limitations

This study focuses on a particular population and utilizes specific testing materials. Therefore, the generalizability of the study is limited. Its analyses are limited by the following parameters:

1. Grade level: Grade 8

2. Sample size: a pilot study

3. Content area: science

4. LEP language background: primarily Spanish

5. Accommodation type: dictionary and glossary

It should be noted that an accommodation for one grade level may not necessarily be appropriate, or even considered an accommodation, for another grade level. Students in lower elementary grade levels may not know how to use a dictionary, or may be in the process of learning to use a dictionary, whereas students in higher elementary grade levels and beyond may be using a dictionary to learn. For this latter group, dictionary use during a testing situation is considered an accommodation. For example, for students in Grade 3 and beyond, the use of a dictionary has already been taught.

In an effort to find classrooms with an equal number of LEP and non-LEP students, site selection was based on state demographic information at the school site level. However, state demographic information does not necessarily reflect the LEP and non-LEP distribution for all classes at a school site. Therefore, future site selection should be based on demographic information collected at the classroom level.

A large proportion of the LEP student population in southern California is native Spanish speaking. Accordingly, for the glossary accommodation we included English glosses and Spanish translations. In our sample, 88% of the LEP students were Hispanic, and 26% of the non-LEP students were Hispanic. LEP students with first languages other than Spanish may have benefited from the English glosses, but the accommodation tells us nothing about the potential impact of translations in their first languages.

Implications and Recommendations

This study addresses several major issues concerning accommodations for LEP students. Although these analyses report on the pilot phase of the study, there are nevertheless several implications for future NAEP assessments.

Since the NAEP is a large-scale assessment, feasibility considerations are important. NAEP assessments involve a large number of LEP students, and ease of administration is a factor. Any element that reduces the burden on states, schools, and students will have a potential positive impact on future NAEP administrations.

Educators are developing accommodation strategies that may reduce the gap between LEP and non-LEP scores in large-scale evaluations. Not all of these strategies may turn out to be easily administered. One-on-one testing, for example, may be a highly effective form of accommodation, but it may not be feasible in large-scale assessments such as the NAEP.

In this study we included only accommodation strategies that we considered easy to implement. A major innovation of this study was the use of a customized dictionary, as an accommodation, in the assessment of students with limited English proficiency. As the study demonstrated, providing a customized dictionary is a viable alternative to providing traditional dictionaries.

Dictionaries are, in fact, already widely used as instructional aids for LEP students, so the concept is not an unfamiliar one for the students. Providing students with actual dictionaries in a testing situation requires extra logistical arrangements and additional cost. In contrast, the customized dictionary's limited number of pages allowed it to be attached directly to the test booklet, minimizing the economic and administrative burden. However, the economic and technical feasibility of providing a customized dictionary as a potential form of accommodation must be evaluated through cost-benefit analysis before a decision can be made concerning its advisability.

Another area of consideration is the inclusion of additional background queries in future studies. Collecting additional information about the academic performance and the language proficiency level of students may help to clarify issues associated with inconsistency in the definition of LEP and the inclusion criteria for standardized assessments. The inclusion of reading achievement data from SAT-9, supplied by the schools, provided valuable information on the language proficiency levels of students beyond the LEP designations.

Given the inconsistency in the LEP designation criteria, gathering additional information about a student's academic and language performance would provide a more comprehensive picture of the student's academic knowledge. More accurate conclusions would be possible from analyses of contextual data, such as student's performance in other content areas and information on family and language background.

References

Abedi, J., Boscardin, C. K., & Larson, H. (2000). *AERA Special Interest Group. Summaries of research on inclusion of students with disabilities and limited English proficient students in large-scale assessments.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., & Lord, C. (1998). *Impact of selected background variables on students' NAEP math performance: NAEP TRP Task 3D: Language background study*. Los Angeles: University of California, Center for the Study of Evaluation / National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., Lord, C., & Baker, E. (1998). *NAEP math performance and test accommodations: Interactions with student language background* (Draft report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Leon, S. (1999, December). *Impact of students' language background on content-based performance: Analyses of extant data* (Draft report). Los Angeles: University of California, Center for the Study of Evaluation / National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2000). *Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data.* Los Angeles: University of California, 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: University of California, Center for the Study of Evaluation / National Center for Research on Evaluation, Standards, and Student Testing.

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

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Education Research, 42,* 359-385.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anderson, N. E., Jenkins, F. F., & Miller, K. E. (1996). *NAEP Inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics.* Princeton, NJ: Educational Testing Service.

Anstrom, K. (1997). *Academic achievement for secondary language minority students: Standards, measures and promising practices.* Washington, DC: National Clearinghouse for Bilingual Education.

August, D., Hakuta, K., & Pompa, D. (1994). *For all students: Limited English proficient students and Goals 2000.* Washington, DC: National Clearinghouse for Bilingual Education.

August, D., & McArthur, E. (1996). *U.S. Department of Education, National Center for Education Statistics. Proceedings of the Conference on Inclusion Guidelines and Accommodations for Limited English Proficient Students in National Assessment of Educational Progress* (NCES 96-861). Washington, DC: National Center for Education Statistics.

Braun, H., Ragosta, M., & Kaplan, B. (1988). Predictive validity. In W. W. Willingham, M. Ragosta, R. E. Bennett, H. Braun, D. A. Rock, & D. E. Powers (Eds.), *Testing handicapped people* (pp. 83-97). Boston, MA: Allyn and Bacon.

Burstein, L. (1993, July). *TRP investigations of validity of NAEP measures.* Paper presented at the TRP / NCES meeting, Washington, DC.

Butler, F. A., & Castellon-Wellington, M. (2000). *Students' concurrent performance on tests of English language proficiency and academic achievement* (Draft Deliverable to OBEMLA). Los Angeles: University of California, Center for the Study of Evaluation / National Center for Research on Evaluation, Standards, and Student Testing.

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education, 61*, 613-615.

Chiu, C. W. T., & Pearson, D., (1999, June). *Synthesizing the effects of test accommodations for special education and limited English proficient students.* Paper presented at the National Conference on Large Scale Assessment, Snowbird, UT.

Cocking, R. R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement in 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.

De Avila, E. (1997, November). *Setting expected gains for non and limited English proficient students* (NCBE Resource Collection, Series No. 8). Washington, DC: George Washington University, National Clearinghouse for Bilingual Education.

De Avila, E., Cervantes, R., & Duncan, S. (1978). Bilingual exit criteria. CABE *Research Journal, 1*(2).

De Avila, E. A., Duncan, S. E., & Navarrete, C. J. (1988). *Finding out/Descubrimiento* (Teacher's Resource Guide). Northvale, NJ: Santillana.

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.

Figueroa, R. A. (1990). Best practices in the assessment of bilingual children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (pp. 93-106). Washington, DC: National Association of School Psychologists.

Gesinger, K. F., & Carlson, J. F. (1992). Assessing language-minority students. *Assessment, Research and Evaluation, 3*(2), 1-4.

Gonzales, V., Castellano, J. A., Bauerle, P., & Duran, R. (1996). Attitudes and behaviors toward testing-the-limits when assessing LEP students: Results of a NABE-sponsored national survey. *The Bilingual Research Journal, 20,* 433-463.

Halliday, M. A. K., & Martin, J. R. (1993.) *Writing science: Literacy and discursive power*. Pittsburgh, PA: University of Pittsburgh Press.

Heath, S. B. (1983). *Ways with words language, life, and work in communities and classrooms*. Cambridge: Cambridge University Press.

Jerman, M., & Rees, R. (1972). Predicting the relative difficulty of verbal arithmetic problems. *Educational Studies in Mathematics, 4,* 306-323.

Kessler, C., Quinn, M. E., & Fathman, A. K. (1992). Science and cooperative learning for LEP students. In C. Kessler (Ed.), *Cooperative language learning: A teacher's resource book* (pp. 65-84). Englewood Cliffs, NJ: Prentice Hall Regents.

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

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

LaCelle-Peterson, M. W., & 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.*

Lam, T. C., & Gordon, W. I. (1992). State policies for standardized achievement testing of limited English proficient students. *Educational Measurement: Issues and Practice, 11*(4),18-20.

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

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

Liu, K., Thurlow, M., Erickson, R., Spicuzza, R., & Heinze, K. (1997). *A review of the literature on students with limited English proficiency and assessment* (Minnesota Report 11). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Longman Dictionary of American English (2nd ed.). (1997). White Plains, NY: Longman.

Massoth, N. A., & Levenson, R., Jr. (1982). The McCarthy Scales of Children's Abilities as a predictor of reading readiness and reading achievement. *Psychology in the Schools, 19,* 293-296.

Mazzeo, J. (1997, March). *Toward a more inclusive NAEP*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES 2000-473). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities.* San Antonio, TX: The Psychological Corporation.

McDonnell, L., McLaughlin, M. J., & Morison, P. (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington DC: National Academy Press.

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. 200-220). Hillsdale, NJ: Lawrence Erlbaum Associates.

Montani, T. O. (1995). *Calculation skills of third-grade children with mathematics and reading difficulties*. Unpublished doctoral dissertation, Rutgers, the State University of New Jersey, New Brunswick.

Moss, M., & Puma, M. J. (1995). *Prospects: The congressionally mandated study of educational growth and opportunity. First year report on language minority and* *limited English proficient students* (Prepared for the U.S. Dept. of Education, Office of the Under Secretary, Office of Educational Research and Improvement, Educational Resources Information Center). Cambridge, MA: Abt Associates.

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

National Clearinghouse for Bilingual Education. (1997). *High stakes assessment: A research agenda for English language learners. Symposium summary.* Washington, DC: George Washington University, National Clearinghouse for Bilingual Education.

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

North Central Regional Educational Laboratory. (1996). *Part 1: Assessment of students with disabilities and LEP students. The status report of the assessment programs in the U.S. state student assessment program database.* Oakbrook, IL: North Central Regional Educational Laboratory and the Council of Chief State School Officers.

O'Connor, M., & Michaels, S. (1993). Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology and Education Quarterly, 24,* 318-335.

Ofiesh, N. S. (1997). *Using processing speed tests to predict the benefit of extended test time for university students with learning disabilities.* Unpublished doctoral dissertation, The Pennsylvania State University.

Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078-1085.

Olson, J. F., & Goldstein, A. A. (1997). *The inclusion of students with disabilities and limited English proficiency students in large-scale assessments: A summary of recent progress* (NCES 97-482). Washington DC: U.S. Department of Education, National Center for Education Statistics.

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

Perkins, D. (1984). Assessment on the use of the Nelson Denny Reading Test. *Forum for Reading, 15,* 64-69.

Rothman, R. W., & Cohen, J. (1989). The language of math needs to be taught. *Academic Therapy, 25,* 133-42.

Shepard, L., Taylor, G., & Betebenner, D. (1998*). Inclusion of limited-English-proficient students in Rhode Island's Grade 4 mathematics performance assessment* (CSE Tech.

Rep. No. 486). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Solano-Flores, W., & Nelson-Barber, S. (2000, April). *Cultural validity of assessments and assessment development procedures*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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.

Stansfield, C. (1998). *English language learners and state assessments*. Paper presented at the annual meeting of the Massachusetts Association of Bilingual Educators, Leominster, MA.

Thurlow, M. L., Liu, K. K., Quest, C., Thompson, S. J., Albus, D., & Anderson, M. (1998, September). *Findings from research on accommodated statewide assessments for English language learners.* Paper presentation at the 1998 annual CRESST conference, University of California, Los Angeles.

Valencia, R. R., & Rankin, R. J. (1985). Evidence of content bias on the McCarthy Scales with Mexican American children: Implications for test translation and nonbiased assessment. *Journal of Educational Psychology, 77,* 197-207.

Willingham, W. , Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A. & Powers, D. E. (Eds.). (1988). *Testing handicapped people.* Boston, MA: Allyn and Bacon.

Zehler, A. (1994). *An examination of assessment of limited English proficient students.* Arlington, VA: Development Associates, Special Issues Analysis Center.

APPENDIX A

Follow-up Questionnaires for Three Groups:

Control (No Accommodation)

Dictionary

Glossary

Follow-Up Questionnaire
Science Test–No Accommodation

1. In the science test, were there words that you did not understand?

   No ___ Yes, some ___ Yes, many ___

2. Would you like to be able to use a dictionary during a test like this?

   No ___ Maybe ___ Yes, definitely ___

3. Did the dictionary help you understand the questions?

   Never ___ Sometimes ___ A lot ___

4. If you had a dictionary to use during the test, how much would you use it?

   No ___ Maybe ___ Yes, definitely ___

5. Would it help if the test explained words in another language?

   No ___ Maybe ___ Yes, definitely ___ What language? ___________________________

6. What else would make it easier for you to understand the questions on the test?

   ____

   ____

   ____

   ____

Follow-Up Questionnaire
Science Test With Dictionary

1. In the science test, were there words that you did not understand?

   No ___ Yes, some ___ Yes, many ___

2. Did you use the dictionary attached at the end of your test booklet to look up words?

   No ___ Yes, some ___ Yes, a lot ___

3. Did the dictionary help you understand the questions?

   No ___ Yes, some ___ Yes, a lot ___

4. Would it help if the test explained words in another language?

   No ___ Maybe ___ Yes, definitely ___ What language? _________________________

5. Would it help if the test used easier words?

   No ___ Maybe ___ Yes, definitely ___

6. What else would make it easier for you to understand the questions on the test?

   ____

   ____

   ____

   ____

Follow-Up Questionnaire
Science Test With Glossary

1. In the science test, were there words that you did not understand?

   No ___ Yes, some ___ Yes, many ___

2. Did you read the explanations in the margins in English (on the right side of the page)?

   No ___ Yes, some ___ Yes, a lot ___

3. Did the English explanations help you understand the questions?

   No ___ Yes, some ___ Yes, a lot ___

4. Did you read the explanations in the margins in Spanish (on the left side of the page)?

   No ___ Yes, some ___ Yes, a lot ___

5. Did the Spanish explanations help you understand the questions?

   No ___ Yes, some ___ Yes, a lot ___

6. Would you like to be able to use a dictionary during a test like this?

   No ___ Maybe ___ Yes, definitely ___

7. If you had a dictionary to use during the test, how much would you use it?

   Never ___ Sometimes ___ A lot ___

8. Would it help if the test used easier words?

   No ___ Maybe ___ Yes, definitely ___

9. What else would make it easier for you to understand the questions on the test?

   ____

   ____

   ____

   ____

APPENDIX B

Science Background Questionnaire

Science Background Questionnaire

1. What country do you come from?________________________________

2. How long have you lived in the United States? _____________ years

3. Are you a male or a female? Male ___ Female ___

4. What is your zipcode? _______________

5. Which best describes you (check one)?

   ___ White (not Hispanic)
   ___ Black (not Hispanic)
   ___ Hispanic
   ___ Asian or Pacific Islander
   ___ American Indian or Alaskan Native
   ___ Other __________________________

6. Do you speak a language besides English (check one)? Yes ___ No ___

   If **yes**, what is that language? _________________________

   If **no**, skip down to question #15.

7. How much do you speak that language with your parents?

Always or most of the time	Sometimes	Never or hardly ever
___	___	___

8. How much do you speak that language with your brothers and sisters?

Always or most of the time	Sometimes	Never or hardly ever
___	___	___

9. How much do you speak that language with your friends at school?

Always or most of the time	Sometimes	Never or hardly ever
___	___	___

10. How much do you speak that language with your friends **outside** school?

Always or most of the time	Sometimes	Never or hardly ever
___	___	___

11. Do you **speak** that language well ?

Very well	Fairly well	Not very well
___	___	___

12. Do you **understand** that language well ?

Very well   Fairly well   Not very well

___   ___   ___

13. Do you **read** that language well ?

Very well   Fairly well   Not very well

___   ___   ___

14. Do you **write** that language well ?

Very well   Fairly well   Not very well

___   ___   ___

15. Have you ever studied science in a language other than English?

___ Yes   ___ No (if **no**, skip to #17)

16. If so, how long were you taught science in a language other than English (choose one)?

___ Less than one year
___ More than one year
___ All my life

17. Have you studied any subjects at school in a language other than English?

___ No
___ Yes (what subjects?) ________________________

18. How long have you studied science in English?

___ All my life
___ Less than one year
___ More than one year

19. Do you **understand spoken English** well?

Very well   Fairly well   Not very well

___   ___   ___

20. Do you **speak English** well?

Very well   Fairly well   Not very well

___   ___   ___

21. Do you **read English** well?

Very well   Fairly well   Not very well

___   ___   ___

22. Do you write English well?

Very well	Fairly well	Not very well
___	___	___

23. Does your family get a newspaper which is written in English regularly?

Yes	No	I don’t know
___	___	___

24. Is there an encyclopedia which is written in English in your home?

Yes	No	I don’t know
___	___	___

25. Are there more than 25 books in English in your home?

Yes	No	I don’t know
___	___	___

26. Does your family get any English language magazines?

Yes	No	I don’t know
___	___	___

27. How much television do you watch in a day?

___ None
___ 1 hour or less
___ 2 hours
___ 3 hours
___ 4 hours
___ 5 hours
___ 6 hours or more

28. In the last two years, how many times have you changed schools because you moved?

___ None
___ 1
___ 2
___ 3 or more

29. How far do you think you will go in school?

___ I will not finish high school.
___ I will graduate from high school.
___ I will have some education after high school.
___ I will graduate from college.
___ I will go to graduate school.
___ I don’t know.

30. I like science.

Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
___	___	___	___	___

31. I am good at science.

Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
___	___	___	___	___

32. What are your grades in science since sixth grade?

___ Mostly As
___ Mostly Bs
___ Mostly Cs
___ Mostly Ds
___ Mostly below D
___ Classes not graded

33. What are your grades in English since sixth grade?

___ Mostly As
___ Mostly Bs
___ Mostly Cs
___ Mostly Ds
___ Mostly below D
___ Classes not graded

34. What are your grades as a whole since sixth grade?

___ Mostly As
___ Mostly Bs
___ Mostly Cs
___ Mostly Ds
___ Mostly below D
___ Classes not graded

APPENDIX C

Demographic Form

University of California, Los Angeles
Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, and Student Testing
301 GSE&IS
Los Angeles, CA 90095-1522

LEP Test Accommodation Study
Demographic Form

School Name: ________________________ Test Date: _______________ Class Subject: _________

Teacher Name: _______________________ Class Grade: _____________

	Student name	Gender	Ethnicity	Does student participate in school free lunch program	Is student LEP or non-LEP	SAT-9 reading score	SAT-9 math score	Language arts rate 1-5	Language spoken at home
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									

LEP Test Accommodation Study
Demographic Form

(continued)

	Student name	Gender	Ethnicity	Does student participate in school free lunch program	Is student LEP or non-LEP	SAT-9 reading score	SAT-9 math score	Language arts rate 1-5	Language spoken at home
24									
25									
26									
27									
28									
29									
30									
31									
32									
33									
34									
35									

Teacher: After completing this form, please return it to the test administrators on the day of the test. You may also fax it to XXX at XXX within seven days after the test date. If you have any questions, please call XXX at XXX.

APPENDIX D

Science Teacher Questionnaire

Science Teacher Questionnaire

School Name: ______________________________ Teacher Name: __________________________

Date: ______________________ Class Period: ______ Class Time: ________________________

Type of science class:
(check one)

______ Integrated Science
______ General Science
______ Life Science
______ Earth Science
______ Other ____

Language of instruction:
(check one)

______ English Only
______ Spanish Only
______ English Sheltered
______ SDAIE
______ Other ____

Topics covered so far
this year:
(check all that apply)

______ contour maps
______ energy transformations
______ energy sources
______ evolution
______ biomes
______ soil erosion
______ the human body
______ phases of matter
______ physics of motion
______ climate
______ properties of water
______ air pressure
______ interpreting graphs

1. How many months have you been teaching this classroom of students? _________ months.

2. How many students are in your class (present at time of testing)? ________ students.

3. Approximately how many of the students in your class are:

   a. Limited English Proficient (LEP)—non-native English speakers _____________
   b. Fluent English Proficient (FEP)—originally LEP, transitioned to FEP _____________
   c. Initially Fluent in English (IFE)—native English speakers _____________

4. In terms of *ethnic background*, approximately how many of your students are:

   a. Latino / Hispanic ________
   b. Caucasian ________
   c. African American ________
   d. Asian / Pacific Islander ________
   e. Other __________________________
   f. Other __________________________

5. In terms of *native language,* approximately how many of your students speak:

   a. English __________
   b. Spanish __________
   c. Vietnamese __________
   d. Other _________________________
   e. Other _________________________
   f. Other _________________________

6. In terms of *English language use,* about how many of your students speak:

   a. English only ________
   b. Spanish only ________
   c. English dominant, Spanish first language ________
   d. Spanish dominant, Spanish first language ________
   e. English dominant, other first language ________
   f. Other ________________________________ ________
   g. Other ________________________________ ________

7. In terms of *general science achievement,* how many of your students would you rate as having:

   a. low-level science understanding ________
   b. medium-level science understanding ________
   c. high-level science understanding ________

8. In terms of *reading* English proficiency, how many of your students are:

   a. Completely fluent in reading the English language ________
   b. Somewhat fluent in reading the English language ________
   c. Not at all fluent in reading the English language ________

9. In terms of *writing* English proficiency, how many of these students are:

   a. Completely fluent in writing the English language ________
   b. Somewhat fluent in writing the English language ________
   c. Not at all fluent in writing the English language ________

10. In terms of *oral* English proficiency, how many of these students are:

    a. Completely fluent in speaking the English language ________
    b. Somewhat fluent in speaking the English language ________
    c. Not at all fluent in speaking the English language ________

11. If you have any comments about the study, the testing experience, or your students or classroom, please include them below.

Thank you very much for your time and assistance!

APPENDIX E

Test Administrator Script

ADMINISTRATION SCRIPT

LEP STUDY

February 2000

ADMINISTRATION SCRIPT

(TOTAL TESTING TIME: 46 MINUTES)

INSTRUCTIONS to the administrator are printed in **BOLD CAPITAL LETTERS** and should not be read to the students. All words in plain print are to be read to the students.

Good morning. My name is ________ and this is my colleague __________.

At UCLA we are looking at science tests. We want to make sure that the questions on science tests are clear and not confusing. By taking this science test today, you can help us in designing better science tests for future students.

Your score on this test will not be part of your grade for this class. However, it is important that you do your best work so that the results are accurate. This will help teachers write better science tests in the future.

We thank you and your teacher, Ms./Mr. ______________, for participating.

We'll be giving to each of you a test booklet and a UCLA pencil; the pencil is yours to keep after the test. Please don't open your test booklets until I tell you to. There should be no talking during the test. It is important that you do your own work and not share answers.

PASS OUT TEST BOOKLETS

On the cover of the test booklet, please write your name clearly, the date, your teacher's name, and the class period. Don't write on the line at the bottom that says ID.

Now, please open your test booklet to page 1. Please follow along in your test booklet as I read the directions aloud.

DIRECTIONS

"Directions: Read each question carefully and answer it as well as you can.
You will have 25 minutes to answer 20 questions.
Mark your answers in your booklet. Circle only one letter for each question.
If you change your answer, erase your first answer completely.
We will now do a sample question together.
Read the sample question. Draw a circle around the best answer.
You should have drawn a circle around D, because there are 120 minutes in 2 hours."

Now look at the cover of your test booklet. Look at the bottom line. If the bottom line on your test booklet says "Test-A" or "Test-B," raise your hand.

CHECK

Good. Your test booklet has no additional directions. However, some test booklets have additional directions.

If the bottom line on your test booklet says “Dictionary-A” or “Dictionary-B,” raise your hand.

CHECK

Note that there are dictionary pages at the end of your test booklet. The dictionary pages are yellow.

ASSISTANT TEST ADMINISTRATOR: HOLD UP A “DICTIONARY” TEST BOOKLET AND TURN TO THE FIRST YELLOW PAGE.

Please find them now, beginning with Page D-1. On page D-1, look at the first word under “A.” That is the word “above.”

CHECK TO MAKE SURE STUDENTS FOUND DICTIONARY PAGE D-1.

Please follow along as I read the definition: “above: in or to a higher position than something else.” In the science test, if the meaning of a word is not clear, you may look up the word in these dictionary pages at any time during the test.

If the bottom line on your test booklet says “Glossary-A” or “Glossary-B,” raise your hand.

CHECK

In the margins of the pages in your test booklet, certain words are explained. If the meaning of a word is not clear, you may look at the explanation in the margin. On the right side of the page, you will find explanations in English.

ASSISTANT TEST ADMINISTRATOR: HOLD UP A “GLOSSARY” TEST BOOKLET, OPEN TO PAGE 3, AND POINT TO ENGLISH GLOSSES.

On the left side of the page, you will find explanations in Spanish.

ASSISTANT TEST ADMINISTRATOR: HOLD UP A “GLOSSARY” TEST BOOKLET, OPEN TO PAGE 3, AND POINT TO SPANISH GLOSSES.

CHECK FOR STUDENT UNDERSTANDING.

You will have 25 minutes to answer 20 science questions. The last science question is on page 19 of your test booklet. When you come to the stop sign on page 19, stop.

SHOW STOP SIGN.

If you finish early, you may go back and check your work.

ASSISTANT TEST ADMINISTRATOR: NOTE TIME AND WRITE START AND STOP TIME ON BOARD:

START:
STOP:

Now turn to page 3 and begin.

ALLOW 25 MINUTES.

AFTER 25 MINUTES.

STOP. Now please turn to page A-1, just after page 19. At the top of this page it says, “Follow-up Questionnaire.” We would like your opinion on the questions in this test. Please answer the questions on page A-1 now.

ALLOW 3 MINUTES OR UNTIL ALL STUDENTS HAVE FINISHED.

Now please turn to the next page, page B-1. At the top of this page it says, “Science Background Questionnaire.” This section asks for some information about you. Please answer the questions on pages B-1 to B-5 now.

ALLOW ABOUT 8 MINUTES OR UNTIL ALL STUDENTS HAVE FINISHED.

We will now collect your test booklets; you may keep the pencil. Thank you very much for being a part of this testing program. We hope that the results and your comments will help teachers to write tests that are fairer and easier to understand.

APPENDIX F

Test Administrator Feedback Form

Test Administrator Feedback Form

TEST ADMINISTRATOR: Please take a moment to give us your feedback and comments.

Date of test:
Teacher:
Class period:
Name(s) of Administrator(s):

1. Were all 6 forms of the test distributed randomly?

2. Did students appear to understand that some of the tests contained dictionary pages at the back, and some had glossary entries in the page margins? Did students with those test forms appear to use the dictionary? The glossary?

3. Was 25 minutes enough time for students to finish the science test?

4. Were the students confused at any point?

5. Did students comment about the difficulty of the science test?

6. Did you observe any negative impact due to simultaneous administering of different accommodations (i.e., dictionary and glossary)?

7. Additional comments?

APPENDIX G

Letter to the Principal

Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, and Student Testing
UCLA Graduate School of Education & Information Studies
405 Hilgard Avenue, 301 GSEIS Building
Los Angeles, CA 90095-1522
(310) 206-1532
Fax (310) 825-3883

Date

XXX
XXX
XXX
XXX

Dear Principal XXX,

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA is currently conducting a study on the validity, feasibility, and differential impact of accommodations for 8th-grade LEP students in science classes.

In this study, we selected a set of science test questions from the 1996 NAEP assessment for administration to 8th-grade students who represent various language backgrounds. We have selected four test treatments, including the control treatment. In addition, a language background questionnaire and a student accommodation follow-up questionnaire complete the assessment procedure, which will take one class period.

We will need one to four Grade 8 classes containing BOTH English speaking and English Language Learner (ELL) students who are currently enrolled in science. We need to know the number of English speaking and ELL students in each science class to ensure that all classes meet our study design. We would like to get out to school sites in January 2000.

We will pay each teacher $125.00 and each school site $125.00 for participating in the study.

If you have any questions or concerns, please call XXX at XXX or me, XXX, at XXX. We will be contacting the science department teachers to follow up on your school site's interest in participating in this study. Thank you for your consideration.

Sincerely,

XXX
XXX

APPENDIX H

Table A1

Multivariate ANOVA Results for Follow-Up Questions, Dictionary Booklet

	SS		*MS*			
Variable	Hypo.	Error	Hypo.	Error	*F*	*P*
Question 1	1.38	27.28	1.38	.24	5.71	.019
Question 2	.002	34.92	.002	.31	.006	.941
Question 3	.037	80.26	.038	.71	.052	.819
Question 4	6.43	34.49	6.43	.31	21.07	.000
Question 5	6.67	51.63	6.67	.46	14.60	.000

Note. *SS* = Sum of Squares. *MS* = Mean Squares.