Language Accommodations for English Language Learners in Large-Scale Assessments: Bilingual Dictionaries and Linguistic Modification

CSE Report 666

Jamal Abedi, Mary Courtney, James Mirocha, Seth Leon, and Jennifer Goldberg

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of California, Los Angeles

December 2005

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

Project 2.4 Accommodations for English Language Learners
Jamal Abedi, Project Director

The work reported herein was supported under the under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

Acknowledgments

Many people generously contributed to the development of this study. We would like to thank all the researchers—faculty, staff, and students—who participated in and informed the study. We are especially grateful to Joan Herman for her insightful comments and advice.

Frances Butler and Alison Bailey advised on the choice of the study's language assessment and test items.

Jennifer Vincent, project assistant, and Liliana Velasquez, provided essential logistical support to every aspect of the study. Kathryn Morrison led the scoring of the study's open-ended test items with insight and care.

Carolyn Huie Hofstetter of University of California, Berkeley, Patrick Lee of Oakland Unified School District, and Cheryl Sparti, formerly of Los Angeles Unified School District, generously arranged California testing sites. Marian Crislip and Selvin Chin-Chance were the Hawaii hosts for the study.

Special thanks to Jenny Kao for her excellent work in preparing this manuscript, including updating and revising various portions of this report.

We appreciate the dedication and hard work of:

Cliff Alexander
Yueh-Wen Chang
Katherine Chun
Rory Constancio
Yvette Cuenco
Amy Gallatin
John Iwanaga
Jennifer Kaplan
Heather Larson
Ani Moughamian
Roy Nakawatase
Priscilla Parks
Rebecca Parks
Alicia Soto
Susan York

A Special Acknowledgement

We are grateful to these colleagues for their advice and contributions to the discussion of the study design.

Richard Durán, UC Santa Barbara

Julia Lara, Council of Chief State School Officers

John Mazzeo, Educational Testing Service

John Olson, Council of Chief State School Officers

Charlene Rivera, George Washington University

Lorrie Shepard, Univ. of Colorado, Boulder

Catherine Snow, Harvard University

Charles Stansfield, Second Language Testing, Inc.

Table of Contents

LANGUAGE ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNERS IN LARGE-SCALE ASSESSMENTS: BILINGUAL DICTIONARIES AND LINGUISTIC MODIFICATION

Jamal Abedi, Mary Courtney, James Mirocha, Seth Leon, and Jennifer Goldberg

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/University of California, Los Angeles

Executive Summary

Recent attention to issues concerning the instruction and assessment of English language learner (ELL) students has placed them among the top national priorities in education. Policy has noticeably shifted from exclusion to inclusion of ELL students in the assessment and accountability system. However, recent research on and practice in the instruction and assessment of ELL students has raised a new set of concerns. One of the most important of these concerns is that language factors may affect students' ability to demonstrate a true picture of what they know and can do in content areas such as math and science.

To fairly assess content knowledge of ELL students, educational researchers and practitioners recommend the use of accommodation. The purpose of accommodation is to help "level the playing field" with regard to English language comprehension. However, sometimes an accommodation does more than is intended. An accommodation may change the ability to assess the construct under measurement by giving an unfair advantage to those receiving the accommodation, thereby negatively affecting the validity of assessment. Further, some forms of accommodation may cause an additional burden to schools, teachers, and large-scale local and national assessment providers.

This study focused on four issues concerning the use of accommodation for ELL students: validity, effectiveness, differential impact, and feasibility. The major theme of this study is to investigate the following questions:

1. Do accommodation strategies help reduce the performance gap between ELL and non-ELL students? (Effectiveness)

2. Do accommodation strategies impact the performance of non-ELL students on content-based assessments? (Validity)

3. Do student background variables impact performance on the accommodated assessments? (Differential impact)

4. Are accommodations easy to implement or use? (Feasibility)

Methodology

A total of 611 Grade 4 and Grade 8 students and 24 teachers at 11 school sites participated in this study during the spring of 1999. English proficiency designation of students was determined based on school records. Language groups chosen were Spanish, Chinese, Filipino, and Korean. Out of 611 students, 317 students (52%) were identified as being ELL or limited English proficient (LEP).[1] The other 294 students either were native English speakers or had become proficient enough in English to be redesignated. These 294 students were combined into our non-ELL category. Of the 317 ELL students, 241 (76%) belonged to one of the four target language groups.

A science test for each grade level with original multiple-choice and open-ended released NAEP (National Assessment of Educational Progress) items was administered under four conditions. One condition had no accommodation. The remaining three conditions included an accommodation, either an English dictionary, a bilingual dictionary, or a linguistic modification[2] of the test, each of which addressed the challenge of understanding the English lexicon and, possibly, its syntax. Science was chosen because the construct being tested is not language, yet the questions tend to have a high language load.

In addition to the science test, an English reading proficiency test was administered for each grade level. The English reading proficiency test consisted of two 25-minute intact released blocks of the 1994 NAEP (Grade 4 or Grade 8) reading assessment and was used to measure the reading ability of both ELL and non-ELL students. The test contained open-ended and multiple-choice items based on the reading passages. This reading measure was given to determine how students at different levels of reading proficiency may have benefited differently from any accommodation received in this study, regardless of ELL status.

The study also included a student background questionnaire, an accommodation follow-up questionnaire, and teacher and school questionnaires. The background questionnaire was used to determine whether student background affects test performance. The questionnaire queried students' language background, country of origin, and length of time in the U.S., and also asked students to self-assess their proficiency in both English and their native language.

The follow-up questionnaire was used to determine whether either dictionary (English or bilingual) helped students during the science test and how students felt the language in the test could have been made easier to understand.

---

[1] In this report, the descriptor *English language learner* or *ELL* signifies a student whose English proficiency is considered "limited." The designation *limited English proficient* or *LEP* is also used to describe the target students in this study.

[2] The linguistic complexity of the science items—but not the science content—was modified to reduce complexity.

The teacher questionnaire included questions regarding educational background and experience. The school questionnaire contained questions regarding school population, and science and English as a Second Language (ESL) resources.

To control for class, teacher, and school effects, test materials and accommodations were distributed randomly among students. Each test booklet was pre-assigned with the student name and accommodation type.

All open-ended test items were scored by at least two teachers who were trained by the project staff. NAEP guidelines and scoring rubrics were followed for double-scoring the open-ended science and language items. Middle school science teachers and Grade 4 teachers scored the open-ended science items. Middle school language arts teachers and Grade 4 teachers scored the open-ended reading items.

Modifications were made during the course of the study. The most important was developing and adding the linguistically modified versions of the science tests to the available set of accommodations. However, at that stage in the study, the number of Grade 8 study participants became limited, so that no non-ELL Grade 8 students were available to use the linguistically modified science test accommodation for that grade level.

Results

Following are alternative hypotheses based on the four research questions:

H_{11}: Some forms of accommodation are more effective than others in reducing the science performance gap between ELL and non-ELL students.

H_{21}: Accommodations impact the performance of non-ELL students on science tests.

H_{31}: Student background variables impact performance on the accommodated assessments.

To investigate these three hypotheses, Grade 4 and Grade 8 ELL and non-ELL students were tested under one of three accommodation conditions or under a standard condition in which no accommodation was provided. The three forms of accommodation were an English dictionary, a bilingual dictionary, and a linguistically modified (i.e., linguistically simplified) version of the science test items. Conditions were randomly assigned to ELL and non-ELL students within each classroom. Eight comparison groups were possible: 4 conditions by 2 levels of ELL status. Since there was no practical reason to give a bilingual dictionary to a non-ELL student, this group was excluded in the design. Hence, seven groups of students were available for comparison: ELL students under four conditions and non-ELL students under three conditions.

Analyses of Open-Ended Questions

Interrater reliability indices (percent of exact and within-one-point agreement, PM correlation, intraclass correlation, kappa, and alpha coefficients) were computed for open-ended science items. In general, interrater reliability coefficients were high and suggest that the open-ended scoring was objective for both Grades 4 and 8.

To test internal consistency for the reading and science tests, the alpha coefficient was computed for both tests. Internal consistency coefficients for the Grade 4 and Grade 8 reading tests were very high and suggest that the reading tests were uni-dimensional and measured only one factor (reading comprehension). For both science tests, however, the internal consistency coefficient was low, suggesting that the science tests were multidimensional.

Testing the Hypotheses

To test the effectiveness hypothesis, we compared the performance of ELL students who were provided an accommodation with the performance of ELL students tested under the standard condition. A significantly higher performance under any accommodation in this study would suggest that the accommodation was effective.

To test the validity hypothesis, we compared the performance of non-ELL students who were provided an accommodation with the performance of non-ELL students tested under the standard condition. Any significant difference in the performance of non-ELL students would suggest an impact of accommodation on the construct, thus creating concerns over the validity of accommodation.

To examine the differential impact hypothesis, multiple regression analysis was performed under two of the conditions (English dictionary, and standard condition). The science test score was used as the criterion, and multiple background variables were entered as predictor variables. The percent of variance (R^2) explained by each model was compared to determine whether or not background variables impacted these accommodation conditions differently.

Results for Grade 4 Students

Grade 4 ELL students had lower science test scores (M = 11.17, SD = 3.67, n = 205) than Grade 4 non-ELL students (M = 12.73, SD = 3.35, n =196). There were differences between ELL and non-ELL performance under different forms of accommodation. Comparing the performance of ELL students under accommodation and under the standard condition, ELL students scored better under an accommodation. For example, the mean science score for ELL students was 11.97 (SD = 3.47, n = 59) under the English dictionary accommodation, as compared with a mean of 10.04 (SD = 3.66, n = 62) for the standard NAEP condition. However, these differences may be attributable to inherent reading proficiency differences rather than to the accommodation.

For non-ELL students, accommodation did not seem to make a difference. Non-ELL students' mean score was 12.94 (SD = 3.54, n = 88) under the English dictionary accommodation, 12.22 (SD = 3.37, n = 23) under the linguistically modified test version, and 12.64 (SD = 3.16, n = 85) under the standard condition.

To test the effectiveness of accommodation, we performed a one-factor analysis of covariance (ANCOVA). The ANCOVA model compared the means of ELL students under different forms of accommodation (English dictionary, bilingual dictionary, linguistically modified test version, and standard condition). Once again, to control for possible differences in reading proficiency among students with different types of accommodation, the reading score was used as a covariate. A nonsignificant F-ratio of 2.40 (df = 3, 194; p = .07) suggested that the accommodation strategies used in this study did not have significant impact on ELL students' performance. However, since the probability of a Type I error for this model (.07) was close to the .05 significance level, we performed multiple comparison analyses. Of the three comparisons made, two were significant. The mean score for ELL students who were provided with an English dictionary was significantly higher than the mean score for ELL students in the standard condition. Also, ELL students who received a bilingual dictionary performed significantly higher than their peers under the standard condition.

To test the validity of accommodation, the performance of non-ELL students under accommodation was compared with the performance of non-ELL students under the standard condition. To control for students' level of English proficiency, the reading score was used as a covariate. A nonsignificant F-ratio of .774 (df = 2, 181; p = .46) indicated that accommodation strategies did not change the performance of non-ELL students, meaning the accommodation strategies did not affect the validity of the Grade 4 assessment.

Results for Grade 8 Students

As mentioned earlier, the bilingual dictionary accommodation was not used with non-ELL students. In addition, in Grade 8 there were no non-ELL students available to test the new linguistically modified accommodation. Thus, there were six comparison groups in Grade 8.

On average, Grade 8 non-ELL students (M = 12.73, SD = 4.21, n = 69) outperformed Grade 8 ELL students (M = 10.94, SD = 3.61, n = 72) by about 2 points. Among the ELL sample, students under the linguistically modified condition scored the highest (M = 13.27, SD = 3.04, n = 11), followed by students under the English dictionary condition (M = 11.52, SD = 3.53, n = 23), and the standard condition (M = 10.32, SD = 3.99, n = 22). Students under the bilingual dictionary condition scored the lowest (M = 9.38, SD = 2.69, n = 16). Among the non-ELL sample, students under the English dictionary accommodation (M = 12.64, SD = 4.19, n = 36) scored about the same as students under the standard condition (M = 12.83, SD = 4.19, n = 36).

To test the effectiveness hypothesis, we compared the performance of ELL students under accommodation and under the standard condition, using a one-factor ANCOVA model. In this model, as in Grade 4, the reading score was used as a covariate. A significant F-ratio of 2.88 (df = 3, 67; p = .04) suggested that the accommodation impacted the performance of ELL students on the science test. Multiple comparisons revealed that the linguistically modified test accommodation group outperformed the standard condition group. The adjusted mean science score for ELL students under the linguistically modified condition was 13.00 (SE = .95, n = 11) as compared with a mean of 10.49 (SE = .67, n = 22) under the standard condition. Neither the English dictionary group (M = 11.37, SE = .66, n = 23) nor the bilingual dictionary group (M = 9.55, SE = .79, n = 16) scored significantly differently from the standard condition group.

To test the validity hypothesis with the Grade 8 data, we compared the performance of non-ELL students under the English dictionary condition and under the standard condition. A nonsignificant F-ratio of .020 (df = 1, 66; p = .89) indicated that the accommodation did not affect the validity of the science measure.

Discussion

The goal of this study was to examine the effectiveness, validity, and feasibility of selected language accommodations for large-scale science assessments. In addition, student background variables were studied to judge the impact of such variables on student test performance.

The results suggested that some of the accommodation strategies employed were effective in increasing the performance of ELL students and reducing the performance gap between ELL and non-ELL students. The results also suggested that the effectiveness of accommodation may vary across grade levels. Some forms of accommodation strategies were shown to be effective for Grade 4 students but not for Grade 8 students. For example, the English dictionary was among the effective accommodations for students in Grade 4. In Grade 8, however, linguistic modification of the science test items seemed to be more effective than any dictionary usage. These results seem reasonable since content assessments for students in higher grades may be more linguistically complex, not just because of vocabulary, but also because of discourse.

Results also showed that the accommodation strategies used in this study did not impact the performance of the general student population. This finding is encouraging because it suggests that the validity of the assessments was not compromised by the use of accommodation.

The results of this study suggest that many background variables were significantly related to performance on the science assessments. These variables include time lived in the U.S., initial grade attended in the U.S., having attended

school outside the U.S., primary home language of Korean or Spanish, and ability to understand spoken English at school.

Response patterns from ELL students found on the follow-up questionnaire varied from one question to another. As expected, ELL students had more difficulty understanding words on the science assessments than did non-ELL students. The students reported significant differences in the amount of both English and bilingual dictionary usage and in the helpfulness of dictionaries. However, ELL students felt that explanation in another language would benefit them more than non-ELL students did. Among ELL students, those who received the English dictionary or the linguistically modified test version rated the helpfulness of explanation in another language lower than did ELL students under the standard condition.

For some students, the dictionary was an unfamiliar tool. Some students used the dictionary as a spelling tool when writing answers to open-ended items, rather than as a key to understanding unfamiliar vocabulary. It was not unusual to see students leave the dictionaries unopened during the test. Some bilingual dictionaries were especially uninformative. Bilingual dictionaries (which provide translation, not definition) do not translate every word that might be found in an English dictionary. Examples from the Grade 4 test are *clump* and *cycle*. A standard English dictionary was found to be overly informative. The science content information contained in a dictionary definition potentially provided an unfair advantage. A sample term from the Grade 8 test is:

half-life *n* : the time required for half of the atoms of a radioactive substance to change composition.

Here the word *time* provides the answer to the science question.

In addition, according to observation notes made by the test administrators, the Korean and Chinese students in Grades 4 and 8 opened their English and bilingual dictionaries more often than Spanish-speaking Grade 4 students and clearly more often than Spanish-speaking Grade 8 students. Such divergent use of the dictionaries between the students suggests that other factors influencing dictionary usage may be at play. This perhaps could relate to attitude toward or familiarity with dictionaries, or other indiscernible reasons that introduce complications to using published dictionaries as an accommodation.

Students did not always know how to use the bilingual dictionary, either because of unfamiliarity with the bilingual format or because of limited native language literacy. Furthermore, locating and distributing quality dictionaries was logistically difficult. Providing dictionaries on a large scale would also seem nearly unfeasible.

Our next study will utilize observations made in the present study and seek to remedy the problems we encountered.

LANGUAGE ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNERS IN LARGE-SCALE ASSESSMENTS: BILINGUAL DICTIONARIES AND LINGUISTIC MODIFICATION

Jamal Abedi, Mary Courtney, James Mirocha, Seth Leon, and Jennifer Goldberg

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/University of California, Los Angeles

Introduction

Recent policy mandates, such as the No Child Left Behind Act of 2001 (2002), have placed issues concerning the assessment of English language learners (ELLs)[1] among the top national priorities in education, especially since ELL students continue to grow rapidly in numbers. Since assessment results often directly shape curriculum and instruction, it is only fair that accurate assessments be made. However, for students of linguistically diverse backgrounds, content-based assessments may inadvertently function as language tests, thereby affecting the validity of the test results. Because construct-irrelevant factors are introduced in the assessments, results may not accurately reflect what was intended to be measured.

Researchers have found a strong relationship between the level of language proficiency and performance on content-based assessments (Abedi & Leon, 1999; Bailey, 2000; Butler & Castellon-Wellington, 2000) in that performance is confounded with language proficiency. To help reduce language factor effects in the assessment of ELL students, accommodations have been suggested and used in both local and national large-scale assessments (see, for example, Mazzeo, Carlson, Voelkl, & Lutkus, 2000; Olson & Goldstein, 1997; Rivera, Vincent, Hafner, & LaCelle-

[1] Both terms, English language learner (ELL) and limited English proficient (LEP), are used in this report. ELL, as defined by LaCelle-Peterson and Rivera (1994), broadly refers to students whose first language is not mainstream English. ELL students include those who may have very little ability with the English language (frequently referred to as LEP) compared to those who have a high level of proficiency. 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).

The authors of this report would like to acknowledge LaCelle-Peterson and Rivera's perspective that ELL is viewed as a positive term because it implies that the student, in addition to having mastered a first language is now in the process of mastering another language. LEP, however, conveys that the student has a deficit or a "limiting" condition. Because accommodations are specifically intended for use with the LEP population of ELL students, instances of the term ELL generally refer to this LEP population.

Peterson, 1997). Accommodations are provided for ELL students in an attempt to "level the playing field" and measure their content knowledge as fairly and accurately as possible.

Accommodations were provided for ELL students in some of the recent National Assessment of Education Progress (NAEP) test administrations: the 1995 field test, the 1996 main math and science assessments, the 1997 main assessment in art, and the 1998 main assessments in reading, writing, and civics. The 1996 NAEP assessment provided the first series of studies evaluating various testing accommodations and their effectiveness, using over-sampling of ELL students in Grades 4, 8 and 12 (Goldstein, 1997; Mazzeo, 1997).

Assessment research, however, has suggested that there may not be a simple solution for obtaining fair and accurate assessments (Rivera, Stansfield, Scialdone, & Sharkey, 2000). The main limitation of the NAEP accommodation data was the lack of comparison or control groups, which is also true for other national large-scale accommodation data. An accommodation may give an unfair advantage to those receiving it if it affects the construct of what is being measured. Thus language-related accommodation must be evaluated for effectiveness and validity, to ensure that it removes the language barrier for ELL students without altering the construct of the assessment. To do this, both ELL and non-ELL student groups should be tested with and without accommodation.

Another concern with the use of accommodation is its feasibility. Some forms of accommodation are difficult to use and expensive, especially in large-scale assessments. For instance, one-on-one testing would seem ideal, but it would be very expensive and logistically impossible to implement on a large scale.

In this report, we describe the research we conducted in 1999 to begin to study the use of language accommodations in science assessment.

Research Questions

The focus of this study concerned the following issues in the use of accommodation for ELL students: effectiveness, validity, differential impact, and feasibility. These research questions guided our study:

- Do accommodation strategies help reduce the performance gap between ELL and non-ELL students? (Effectiveness)

- Do accommodation strategies impact the performance of non-ELL students on content-based assessments? (Validity)
- Do student background variables impact performance on the accommodated assessments? (Differential impact)
- Are accommodations easy to implement and use? (Feasibility)

To investigate these questions, we tested both ELL and non-ELL students in Grades 4 and 8 under one of three accommodation conditions, or under a standard condition in which no accommodation was provided. The three forms of accommodation were English dictionary, bilingual dictionary, and a linguistically modified (i.e., linguistically simplified) version of the test items. (More discussion on the accommodation types follows in the literature review portion of this report.) Thus eight comparison groups were possible: four levels of accommodation by two levels of ELL status. However, since there was no practical reason to give non-ELL students bilingual dictionaries, this group was excluded, leaving seven comparison groups for investigation.

We included a variety of accommodations to compare their effectiveness. Accommodation strategies were selected based on frequency of usage, nationwide recognition, feasibility, and first-language literacy factors. Each of the accommodation strategies used in this study (English dictionary, bilingual dictionary, and linguistic modification of items) can function as a language aid for ELL students on large-scale assessments. Students from different language and cultural backgrounds were included to examine any possible cross-cultural or cross-language factors that may impact the outcome of accommodated assessment. We tested both ELL and non-ELL students to observe the effects of accommodation on the general student population.

Finally, we included a measure of English reading proficiency because we believe both ELL and non-ELL groups are not homogeneous groups within themselves. Findings from past studies suggested that ELL and non-ELL students vary substantially in their English language capabilities, and the effectiveness of accommodation can depend on students' English language background. Scores on the English reading proficiency tests were used as a covariate. This approach controls for any unsubstantiated initial difference that might occur despite randomizing treatment (accommodation conditions) within the classrooms. Using

reading scores as a covariate provides statistical control of these preexisting group differences.

Literature Review

Federal legislation in the last decade, including the No Child Left Behind Act of 2001, Goals 2000, and the Improving America's Schools Act of 1994, illustrates the continuing interest in reaching higher standards for all students, including ELL students. The rapidly growing ELL population makes this even more challenging. According to the 2000-2001 *Survey of the States' Limited English Proficient Students and Available Educational Programs and Services* (Kindler, 2002), more than 4.5 million public school students were identified as limited English proficient. California enrolled the largest number of public school LEP students, representing one third of the total national LEP enrollment. Since the 1997-1998 school year, there has been a greater than 27% increase in LEP enrollments. As English language learners continue to increase in numbers, both researchers and policymakers strive for accurate assessments of ELL students.

In the past, many ELL students were exempted from exams because large-scale assessments were not effectively assessing those students' content knowledge, and better alternatives were not available. However, exempting students from assessments does not provide a measurement for progress and may deny such students from educational opportunities that are shaped by assessment results. In search of accurate assessments for ELL students, individual classrooms and schools turned to alternatives such as portfolios, interviews, and oral testing. Though perhaps effective on a small scale, such methods are not cost effective and are too time-consuming for large-scale assessments.

The Improving America's School Act of 1994 (IASA) states that "limited English proficient students . . . shall be assessed to the extent practical in the language and form most likely to yield accurate and reliable information on what students know and can do to determine such students' mastery of skills and participants other than English." Furthermore, the No Child Left Behind Act of 2001 calls for stronger accountability and the provision of accommodation for those who need it. Such legislation has inspired us to investigate the usage of accommodation, in the hope of creating more fair and accurate assessments.

In the following pages we will provide a brief overview of issues related to the inclusion of ELL students in large-scale assessments, as well as a discussion of the

accommodations on trial in this study. The following topics are included in the review: validity issues for assessing ELL students; performance differences between ELL and non-ELL students; defining accommodation; state policies for ELL students; evaluating the use of accommodation; bilingual dictionaries and glossaries as accommodation; linguistic modification of test items as accommodation.

Validity Issues for Assessing ELL Students

Using assessments designed for non-ELL students with ELL students often fails to provide valid inferences about ELL content knowledge. Background differences, including English language proficiency, can interfere with a student's ability to demonstrate content knowledge. Consequently, assessment procedures may not yield valid results for ELL students (Gandara & Merino, 1993; LaCelle-Peterson & Rivera, 1994). The recently revised *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) reminds us that "for all test takers, any test that employs language is, in part, a measure of their language skills" (p. 91), which is of particular concern for ELL students, in that "construct-irrelevant components" may be introduced into the test, so that "test results may not reflect accurately the qualities and competencies intended to be measured" (p. 91).

Ideally, assessment instruments will yield beneficial and accurate information about students. In order to provide the most meaningful data, LaCelle-Peterson and Rivera (1994) suggested several questions to be addressed when evaluating assessments:

Technical/validity questions:

- Is the test valid for the school populations being assessed—ELL students?
- Have available translations been validated and normed?
- Has the role of language been taken into account in the scoring criteria?
- Do the scoring criteria for content area assessments focus on the knowledge, skills, and abilities being tested, and not on the quality of the language in which the response is expressed? Are ELL students inappropriately being penalized for lacking English language skills?
- Are raters who score students' work trained to recognize and score ELL responses?

Equity considerations:

- Are ELL students adequately prepared and instructed to know the content being assessed?
- Have ELL students been given adequate preparation to respond to the items or tasks of the assessment?
- Has the content of the test been examined for evidence of cultural, gender or other biases?
- Is the assessment appropriate for the purpose(s) intended?
- Has appropriate accommodation that would give ELL students the same opportunity available to monolingual students been provided?

Performance Differences Between ELL Students and Non-ELL Students

Research has found that students' language background is confounded with their performance in content-based areas (Abedi, Courtney, & Leon, 2001; Abedi & Leon, 1999; Abedi, Leon, & Mirocha, 2001; Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000). The linguistic complexity of test items may confound scores on achievement tests. ELL students may be unfamiliar with linguistic structures of questions, may not recognize vocabulary terms, or may misinterpret an item literally (Durán, 1989; Garcia, 1991).

Aside from linguistic complexity, cultural variables may influence student performance on assessments. Such variables include student disinclination to ask questions, attitudes toward competition, attitudes toward individualism versus collectivism, gender roles, attitudes toward the use of time, attitudes toward the demonstration of knowledge, use of body movements and gestures, and use of eye contact (Liu, Thurlow, Erickson, Spicuzza, & Heinze, 1997). Abedi, Lord, and Hofstetter (1998) found that student background variables such as language background, length of stay in the United States, overall grades, and the number of school changes were valuable predictors of ELL student performance in math and reading.

According to Mazzeo et al. (2000), in the NAEP 1996 mathematics sample, the majority (see Table 1) of LEP students had received science instruction at their grade level. In the sample, there were still large groups of students receiving instruction below grade level. Nineteen percent of Grade 8 LEP students were receiving science instruction 2 or more years below grade level. It would be difficult for Grade 8

Table 1

Percentage of Grade-Level Distribution of LEP Students With English Language Instruction in Science by Grade

What grade level of instruction in the English language is this student currently receiving in science?	Grade 4 (%)	Grade 8 (%)
Above grade level	0	0
At grade level	83	76
One year below grade level	11	5
Two or more years below grade level	6	19

Note. Table adapted from Mazzeo et al., 2000, p. 76. Source: National Center for Education Statistics, NAEP, 1996 Mathematics Assessment.

students to take Grade 8 level assessments when they are only receiving instruction at the Grade 6 level.

NAEP results from 1990 through 1996 showed lower average scores for LEP students included in the NAEP assessment than for the English-proficient students in several content areas, including science (see Table 2; Mazzeo et al., 2000).

Defining Accommodation

Accommodations have been suggested and used for ELL students in assessments. Also referred to as modifications or adaptations, they are changes made to either the test or the testing procedure that help to provide a more accurate measure of content knowledge. Such changes may be prepared ahead of time or

Table 2

Percentage Distribution of Estimated Grade Level of Performance in Science for LEP Students by Grade

At what grade level is this student currently performing in the English language in science?	Grade 4 (%)	Grade 8 (%)
Above grade level	1	1
At grade level	46	31
One year below grade level	27	13
Two or more years below grade level	20	34
I don't know	6	22

Note. Table adapted from Mazzeo et al., 2000, p. 78. Source: National Center for Education Statistics, NAEP, 1996 Mathematics Assessment.

provided to the student during the test. The following are examples of accommodations involving changes in test format:

- using a translated version of the assessment in the student's home language;
- using a bilingual version of the test;
- modifying complex linguistic features of test items; and
- incorporating glossaries into the test.

Examples of accommodations involving changes to the test procedure (from Rivera et al., 2000) include

- allowing extra time to take the test;
- using small-group testing or multiple testing sessions;
- having a familiar test administrator;
- providing dictionaries or glossaries; and
- giving simplified directions or repeating directions aloud.

To assess academic achievement, data are collected in ways that demonstrate a student's knowledge, skills, and abilities. In order to effectively assess ELL students, comprehensive systems that attempt to assess all that students are learning must be used. The use of accommodation may improve the accuracy of test scores by eliminating irrelevant obstacles for ELLs. Therefore, scores earned on tests with appropriate accommodation are more likely to maintain the validity of the test and minimize error in the measurement of the student's abilities.

State Policies for ELL Students

States vary on policies regarding the identification of ELL students and the role of accommodation in assessments for ELL students. During the 1998-1999 school year, 40 states had accommodation policies and 37 of them allowed accommodations (Rivera et al., 2000), bringing accommodation use to 74% nationwide.

We next describe the state policies of two of the two states from which we pooled participants for this study: California and Texas. (For further detail or for information on other states, see Rivera et al., 2000.)

In California, students are identified as ELL students based on a home language survey, an English oral/aural proficiency test, and grade-appropriate literacy testing. Test exemptions are not allowed in California. There is no specific California state policy regarding accommodation on assessments for ELL students (California Department of Education, 2000; Rivera et al., 2000).

In Texas (Texas Education Agency, n.d.), ELL students are identified based on a home language survey; an oral language proficiency test; informal assessment through a teacher/parent interview, student interview or teacher survey; standardized achievement test scores; and classroom grades. Testing accommodation is permitted unless it would make a particular test invalid as a measure for school accountability. The permissible accommodations include translation of directions on all components in a student's native language and translating of some components of the test in a student's native language. School district officials are the decision makers about accommodations for ELL students.

In general, state policies on the process of identifying ELL students contain some similarities, including collecting information about home language and from assessments. Not all states have specific accommodation policies, although all states seem to be addressing concerns for including all students in large-scale assessments. More research, however, is needed to determine the most effective way to accommodate ELL students. Our study investigated the use of English and bilingual dictionaries and linguistically modified items as accommodation.

Evaluating the Use of Accommodation

Although appropriate test accommodation helps "level the playing field," it is important that accommodations do not provide unintentional advantages to students receiving them (Rivera & Stansfield, 1998). For instance, some students may be able to correctly respond to a question only because the item was answered within a dictionary's definition. Also, providing extra time only for ELL students may give them an unfair advantage if other students have lower scores simply due to lack of time to complete test items.

There has been much discussion over using translated versions of a test as an accommodation. Saville-Troike (1991) suggested that where appropriate, all ELL students should have a right to assessment in their native language as well as in English. However, Abedi et al. (1998) found that translating test items from English to other languages may not accommodate students who are taught in English, and

may have more detriments than benefits. Students typically come from linguistically diverse backgrounds, and their level of fluency and literacy in the home language varies. Even in providing translations to those literate in the home language, results must first be determined as comparable to testing in English (National Clearinghouse for Bilingual Education [NCBE], 1997). The 1995 NAEP field test results indicated that translated versions of some items may not have been parallel in measurement properties to the English versions (Olson & Goldstein, 1997). For these reasons, some feel that turning to other types of accommodation may be more appropriate.

Rivera, Vincent, Hafner, and LaCelle-Peterson (1997) noted that 52% of states reported allowing test modifications for LEP students on at least one statewide assessment. Extra time was the most frequent test modification reported by states. The North Central Regional Educational Laboratory (NCREL) (Liu et al., 1997; NCREL, 1996a, 1996b) also found that half of states reported allowing accommodation for LEP students, mainly including a separate setting, a flexible testing schedule, small-group administration, extra time, and simplified directions. Some states, such as Arizona, Hawaii, New Mexico, and New York, used other languages on the test or an alternative test (Liu et al., 1997).

The 1996 NAEP science tests were designed with three samples of schools, using inclusion criteria in the third sample and having a variety of assessment accommodations available. Permitted accommodations for ELL students included one-on-one testing, small-group testing, extended time, oral reading of directions, and a Spanish/English glossary of scientific terms. Students using the glossary were usually given extra time (O'Sullivan, Reese, & Mazzeo, 1997). Abedi et al. (2000) found that linguistically modified testing, extra time, and glossary plus extra time helped ELL students. Evidence indicates that the provision of accommodation results in higher rates of participation for ELL students (Mazzeo et al., 2000; O'Sullivan et al., 1997). However, the availability of accommodation is another challenge to measurement. Bilingual versions of the 1996 NAEP science assessment were not developed due to resource constraints and comparability concerns.

Effective and economical accommodation on national standardized tests will allow schools, districts, and states to be compared more reliably. The next two sections focus specifically on two types of accommodations: bilingual dictionaries/glossaries and linguistic modification.

Bilingual Dictionaries and Glossaries as Accommodation

In order to assist students who have a limited English vocabulary, dictionaries or glossaries have been used as an accommodation. Some states, such as Ohio and Massachusetts, have approved lists of bilingual dictionaries (Rivera & Stansfield, 1998). These dictionaries actually function as glossaries by merely translating words rather than defining them. The states wanted to ensure that larger, expanded dictionaries, which give an unfair advantage to students, are not used.

Customized test glossaries may be a better alternative to dictionaries. They include only the words used on the test and they define the words only in the context in which they appear on the test. These glossaries can also be used more efficiently than an English or bilingual dictionary. They are more practical for national assessments when they accompany the tests and do not have to be separately provided. Students at schools that are unable to provide bilingual dictionaries still have the opportunity to use an accommodation when a customized test glossary is provided.

Both bilingual and monolingual glossaries can be used. The bilingual glossary is a cross-lingual list of words that appear on the test. This kind of glossary translates words that are used to build the context of the item. It does not serve as a reference on the subject being tested by the item or the test. Similarly, a monolingual glossary provides synonyms for words without explaining the material being tested. This prevents leading students to answers.

Monolingual glossaries have several advantages over bilingual glossaries and bilingual dictionaries. They serve the needs of students of all native language groups. Also, they may be especially helpful for students who are taught in English. On the other hand, the monolingual glossary may not be as effective as the bilingual glossary. Students may not always be able to infer meaning from an English word, whereas the bilingual glossary immediately provides the equivalent word in the native language (Rivera & Stansfield, 1998).

However, for some, bilingual dictionaries may be more useful than glossaries. Students who use bilingual dictionaries in their classrooms on a regular basis may feel more comfortable with the dictionaries. They may have a better grasp of how to use an accommodation with which they are already familiar. Also, students who regularly use bilingual dictionaries may feel that a necessary tool of access has been withdrawn when they are not allowed to use it during an assessment.

Linguistic Modification of Test Items as Accommodation

Assessments with linguistically modified test items may also facilitate students' negotiation of language barriers. Linguistic modification involves altering the language of a text while keeping the content the same. This may be accomplished by shortening sentences, removing unnecessary expository material, using familiar or frequently used words, using grammar thought to be more easily understood—including using present tense—and using concrete rather than abstract formats (Abedi, Lord, & Plummer, 1997).

Analyses based on the linguistic complexity of items (Abedi et al., 1997) showed significant differences with respect to language background between student scores on complex and less complex items. According to Abedi and Lord (2001), it appears that modifying the linguistic structures in math word problems can improve ELL student performance. Students indicated preferences for linguistically modified items during focus group interviews and also scored higher, on average, on those items.

To accurately assess knowledge within content areas, students must comprehend what the items are asking and understand the response choices. However, in analyzing Grades 3 and 11 standardized math and science assessments, Imbens-Bailey and Castellon-Wellington (1999) found that two thirds of the items included general vocabulary words that were uncommon or used in an atypical manner. One third of the items included syntactic structures that were evaluated as complex or unusual in their construction.

Table 3 summarizes research findings of Abedi et al. (1997) accompanied by practical recommendations from Shuard and Rothery (1984).

Investigating Bilingual Dictionaries and Linguistic Modification

The constant need for ELL students to reach higher standards prompted us to investigate proper accommodations for them in large-scale assessments. Since existing accommodation data do not allow extensive evaluation, and accommodation research specifically for ELL students is limited, we decided to conduct a study evaluating bilingual dictionaries and linguistic modification of test items, which we will now describe.

Table 3

Research Findings Within the Field of Linguistic Complexity and Practical Recommendations for Linguistically Modifying Texts

Research findings	Practical recommendations
Words that are short (simple morphologically) tend to be more familiar and, therefore, easier.	Use simple words; use high-frequency words.
Passages with words that are familiar (simple semantically) are easier to understand.	Use familiar words. Omit or define words with double meanings or colloquialisms.
Longer sentences tend to be more complex syntactically and, therefore, more difficult to comprehend.	Retain Subject-Verb-Object structure for statements. Begin questions with question words. Avoid clauses and phrases.
Long items tend to pose greater difficulty.	Remove unnecessary expository material.
Complex sentences tend to be more difficult than simple or compound sentences.	Keep to the present tense, use active voice, avoid the conditional mode, and avoid starting with sentence clauses.

Sources: Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance.* (CSE Tech. Rep. No. 429). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing. Shuard, H., & Rothery, A., (Eds.). (1984). *Children reading mathematics.* London: J. Murray.

Methodology

This investigation was a study of the use of accommodation by ELL students on a test comprised of National Assessment of Education Progress (NAEP) science questions. The study was conducted between February 2000 and June 2000 with school sites in California, Hawaii, and Texas. The objective of the study was to test the instruments and determine the effectiveness of our procedures in administering accommodation for ELL students.

Participants

A total of 611 students (421 Grade 4 and 190 Grade 8 students) and 24 teachers at 11 school sites participated in the study. Out of the 611 students, 317 students (52%) were identified as ELL. Of those, 241 (76%) belonged to the target language groups sought in individual classrooms. Target language groups chosen were Spanish, Chinese, Filipino, and Korean. Teachers or administrators determined the English proficiency designation of students based on their schools' records. The 294 non-ELL students either were native English speakers or had become proficient enough in English to be redesignated. For the sake of this study, these students were combined into our "non-ELL" category. (Table A1 in Appendix A provides the distribution of participants across grades, schools, classes and designations.)

ELL students were studying science in different settings: in a bilingual program, in an English as a Second Language (ESL) science class, or in a mainstream class. Occasionally, a non-ELL class was tested in a school in order to balance another class comprised of ELL students. In one case, Filipino ELL students were drawn from five science classes in order to provide a significant number of ELL-designated participants who used the target language.

Region, school and class selection. Target languages in this study were chosen based on the largest second language groups in the United States. Research was then conducted to determine specific locations where there were communities belonging to these target groups. After specific areas were chosen, schools in each area were selected by determining the percentage of ELL students in Grade 4 and Grade 8, the percentage of students in those schools belonging to target language backgrounds, and the percentage of these students still classified as ELL. Permission was obtained from each participating school district and principal to conduct the study.

The principal or designated site coordinator generally chose two classes for testing so that, when possible, a significant portion of the participants would be ELL-designated students. Of those, as many ELL-designated students as possible represented a single target language population. The initial goal for class selection was to use Grade 4 and Grade 8 science classrooms with an equal distribution of ELL (from target languages) students and non-ELL students. The reality of classroom demographics, however, required us to be more flexible and, at times, to use more ingenuity to locate and recruit significant numbers of ELL students from the target languages and/or their non-ELL counterparts. In districts where ELL students were enrolled in ELL-only classes, both all-ELL and all-non-ELL classes were tested.

Instrumentation

For the study, Grade 4 and Grade 8 students were assessed on their understanding of science concepts and their reading comprehension. Each assessment was modeled after assessments administered by NAEP. The science tests incorporated a variety of multiple-choice and open-ended questions on earth, physical, and life science concepts that Grade 4 and Grade 8 students are expected to have been taught by that time in the school year. The Grade 4 reading test focused on assessing expository ability and narrative understanding through a variety of multiple-choice and open-ended questions.

The questionnaires for students, teachers, and schools were adaptations of existing tools or were newly developed. The science test candidate items for Grades 4 and 8 were based on the *NAEP Assessment and Framework Specifications* (National Assessment Governing Board, n.d.). The final selection was based on advice received from Grade 4 and Grade 8 science teachers. The science teachers evaluated the items' language and content difficulty. Items were eliminated from the selection pool if language was extremely complex, or the material was not likely to have been taught in Grade 4 or Grade 8, or if the items measured more recall than understanding, reasoning, or investigation. The following are details about the instruments.

Standardized science achievement tests. Subscales of standardized achievement tests in science were used to provide measures of dependent variables for this study. The science tests asked a variety of open-ended and multiple-choice questions from NAEP Grade 4 and Grade 8 science assessments. Students were assessed on their ability to demonstrate understanding of physical, earth, and life science concepts.

Grade 4 science test. Eight life science questions were taken from Section 2.1 of the 1996 NAEP Grade 4 Science assessment. Students were given 30 minutes to complete this section of multiple-choice and open-ended questions. The second section merged questions from the 1996 NAEP Science assessment sections 1 and 2.2. Students were given 15 minutes to answer 10 multiple-choice questions in life science, earth science and physical science. Of the 19 items in the Grade 4 science test, 8 were open-ended and 11 were multiple-choice.

Grade 8 science test. The Grade 8 science test contained a total of 30 multiple-choice and open-ended questions in order to assess understanding of various physical, earth, and life science concepts. Students were given 45 minutes to complete the exam, which incorporated 24 multiple-choice and 6 open-ended questions. Two versions of the Grade 8 science test, Booklet A and Booklet B, were created, which shared the same open-ended items; all items were presented in a different order in each booklet to discourage cheating. (A third version was created when Booklet A was linguistically modified. However, no non-ELL Grade 8 students were available to take the linguistically modified version.) The questions in these tests came from the 1988, 1990, and 1996 NAEP Grade 8 Science assessments.

Reading proficiency tests. An English reading efficiency/proficiency test was built for each grade level from two 25-minute intact blocks of the 1994 NAEP standardized reading assessment to measure the reading ability of both ELL and non-ELL students. One class period (approximately 55 minutes) was allocated to the reading proficiency test. Each reading proficiency test contained two reading passages, one narrative and one expository. The Grade 4 passages were followed by 5 to 6 multiple-choice test items and 5 open-ended test items each. The Grade 8 passages were followed by 3 to 4 multiple-choice test items and 7 to 8 open-ended test items each. The passages and questions for both grades were complete blocks taken from the 1994 NAEP Reading Assessment.

Student background questionnaire. The study included a student background questionnaire, used to determine whether background affected performance on the tests. The questionnaire included questions pertaining to language background, such as country of origin, length of time in the U.S., and language other than English spoken in the home. It also asked students to self-assess their English and native language proficiency. The questionnaire included items selected from both the 1996 NAEP administration and an earlier language background study conducted by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Different background questionnaires were used for Grades 4 and 8. Even though the content of the background questionnaires was similar across the two grades, the structure and wording of the questions for Grade 4 were simpler. Some questions had fewer response categories for Grade 4 students than for students in Grade 8. For example, Question 3 for both grades asks students to indicate at what grade they started their schooling in the U.S. For Grade 4 students, the response options were from preschool to Grade 4. For Grade 8 students, response options for Grades 5, 6, 7, and 8 were added.

Several new questions were added for Grade 8, querying the student's level of English proficiency (for speaking, reading, and writing) when the student first started school in the U.S. (questions 4, 5, and 6). Another new question asked about the student's level of readiness to do an oral presentation, write a report, or take a multiple-choice test.

Teacher and school questionnaires. The teacher questionnaire included questions regarding educational background and experiences. The school questionnaire contained questions about the school population and its science and English as a Second Language (ESL) resources.

Follow-up questionnaire. Students were asked to respond to a follow-up questionnaire in order to determine whether the dictionary helped them during the test and how the language in the test could have been made easier to understand.

Test booklets. At the beginning of the study, three in-class instruments were used: a science test, a reading test and a student background questionnaire. After administration at the third school, the follow-up questionnaire was created and appended, along with the student background questionnaire, to the science test booklet. There were two versions (A and B) of the Grade 8 science test throughout the study. After administration at the first three schools, the linguistically modified versions of the Grade 4 and Grade 8 science tests were added for the rest of the study. Early in the study, test administrators observed students cheating off their neighbors' tests, so a second version of each reading test was created by switching the order of the reading blocks and their questions.

Dictionaries. The English language dictionary used was a hardcover *Merriam Webster's Intermediate Dictionary*. Bilingual Spanish, Korean, Chinese, and Ilocano dictionaries were used for this study. We selected the bilingual dictionaries by consulting librarians, teachers, and linguists and then examining their recommendations. The dictionaries' contents were compared to each science test's lexicon. Because the English-Ilocano dictionary was located in the last third of a book, it was marked with a bookmark. The dictionaries contained non-science content words and science content words found in the tests. The English dictionary also contained definitions of both content and non-content words in the tests. Appendix B illustrates the content differences among the dictionaries used in the study. (See also Appendix C for a reference list of the dictionaries used. See Tables B3 and B4 for the test words found in each of them.)

Linguistically modified test version. In an effort to test whether linguistic modification of science test items reduced the performance gap between ELL and non-ELL students, linguistically modified versions of the Grade 4 and Grade 8 science tests were prepared. Words and sentences were amended or deleted to reduce the linguistic complexity, leaving the content of the question and content of the multiple-choice responses intact.

First, prior research on the effect of linguistic complexity on the performance of ELL students in content area assessment was thoroughly reviewed. Using linguistic modification guidelines developed at CRESST and considering other linguistic

features that contribute to difficulty in reading comprehension, we revised many of the Grade 4 and Grade 8 NAEP Science assessment items. As a result, the potentially challenging linguistic features were removed, reduced, or recast. Scientific vocabulary and concepts were preserved; only non-content vocabulary was changed.

The features most often modified included unfamiliar words, complex sentences, unnecessary expository material, abstract (vs. concrete) presentations, and passive voice. Questions that did not begin with a question word were also modified.

An example of an original item and its modified version is presented below.

Original version:

> One day Ms. Brown brought a bucket of pond water to her fourth grade class. In the bucket were several clumps of frogs' eggs, and there were many eggs in each clump, as you can see in Picture 1. "We'll put these eggs and the pond water into the fish tank on the table in the back of the room," said Ms. Brown, "and soon these eggs will hatch into tadpoles. Then we can watch as the tadpoles grow and change into frogs."
>
> Today, two weeks later, all of the eggs that are going to hatch have hatched and the fish tank is full of tadpoles. The last eggs hatched yesterday. As you can see in Picture 2, all the tadpoles do not look alike.
>
> Draw a circle around each of the tadpoles that hatched yesterday.

Modified version:

> Two weeks ago, Ms. Brown brought a bucket of pond water to her science class. In the bucket, there were several clumps of frogs' eggs. There were many eggs in each clump, as you can see in Picture 1. Ms. Brown said, "We'll put these eggs and the pond water into the fish tank. Soon these eggs will hatch into tadpoles. Then we can watch the tadpoles grow and change into frogs."
>
> Today, it is two weeks later. All of the eggs have hatched, and the fish tank is full of tadpoles. The last eggs hatched yesterday. As you can see in Picture 2, the tadpoles do not look alike.
>
> Draw a circle around each of the tadpoles that hatched yesterday.

Changes:

PARAGRAPH 1

- Idiomatic and abstract "one day" changed to more concrete "two weeks ago."
- Long noun phrase "fourth grade class" shortened to "science class."
- Understood subject of second sentence expressed with "there."
- Compound with "and" changed to two shorter sentences.
- Quotation introduced at beginning of sentence.
- Compound sentence in quote changed to two shorter sentences.

PARAGRAPH 2

- Two introductory items made into a separate, simple sentence.
- Unnecessary relative clause "that are going to hatch" removed.
- Indefinite pronoun "all" omitted.

See "Linguistic Modification Concerns" in Appendix D for a list of commonly revised linguistic features.

Design and Procedure

Science tests containing 19 (Grade 4) and 30 (Grade 8) NAEP items were administered in four forms to ELL and non-ELL students in Grades 4 and 8. One form contained original items with no accommodation. The remaining forms included one of three accommodations: an English dictionary, a bilingual dictionary, or a version with linguistically modified test items. All students had extra time to complete the science test. In addition to the science test, a reading assessment and questionnaires for each grade level were administered. Based on our observations during testing sessions, adaptations were made to the design and procedure throughout the study (see Appendix E for further detail).

Distribution of accommodations. A process was developed to ensure that the test materials and accommodations were distributed efficiently and randomly, yet as evenly as possible, among both the ELL and non-ELL students. When possible, schools sent rosters of the participating classes, which were examined to determine whether the class indeed contained enough ELL students with the specified home

language. After the students' names were entered into a database, they were divided into ELL and non-ELL categories. The ELL students belonging to the specified home language were noted. A specified number of these students were randomly assigned bilingual dictionaries; then other accommodations or no accommodations were assigned randomly among the remaining ELL and non-ELL students. In most classes in the study, a few students received linguistically modified science tests, the latest of the accommodations on trial. Figures 1 and 2 show examples of accommodation distributions.

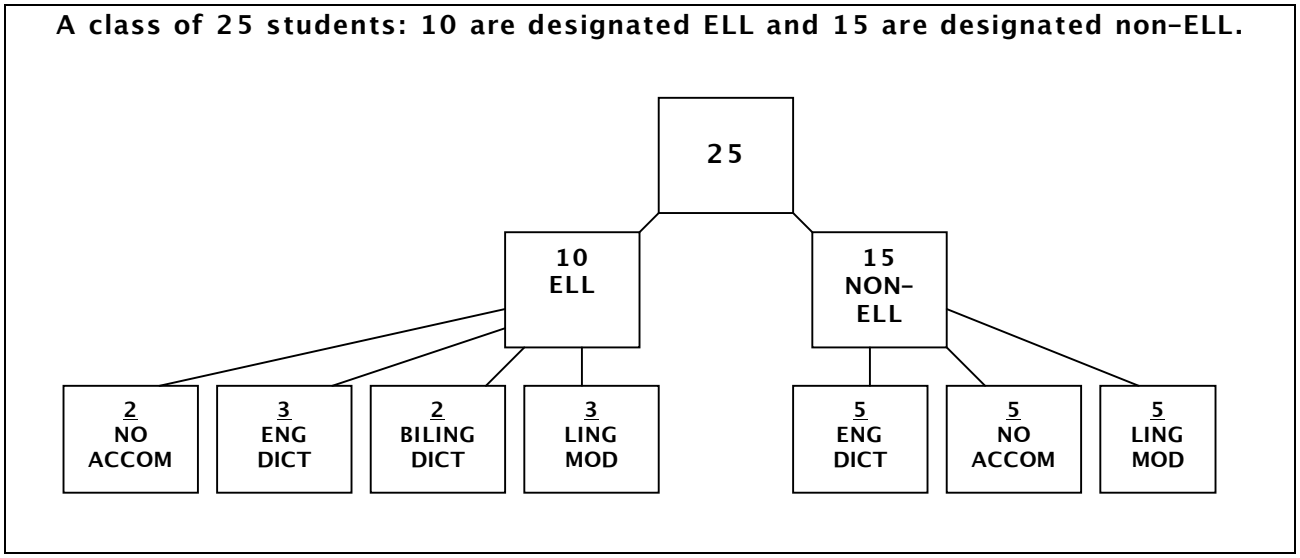

Figure 1. Example of accommodation distribution of three possible accommodations and no accommodation.

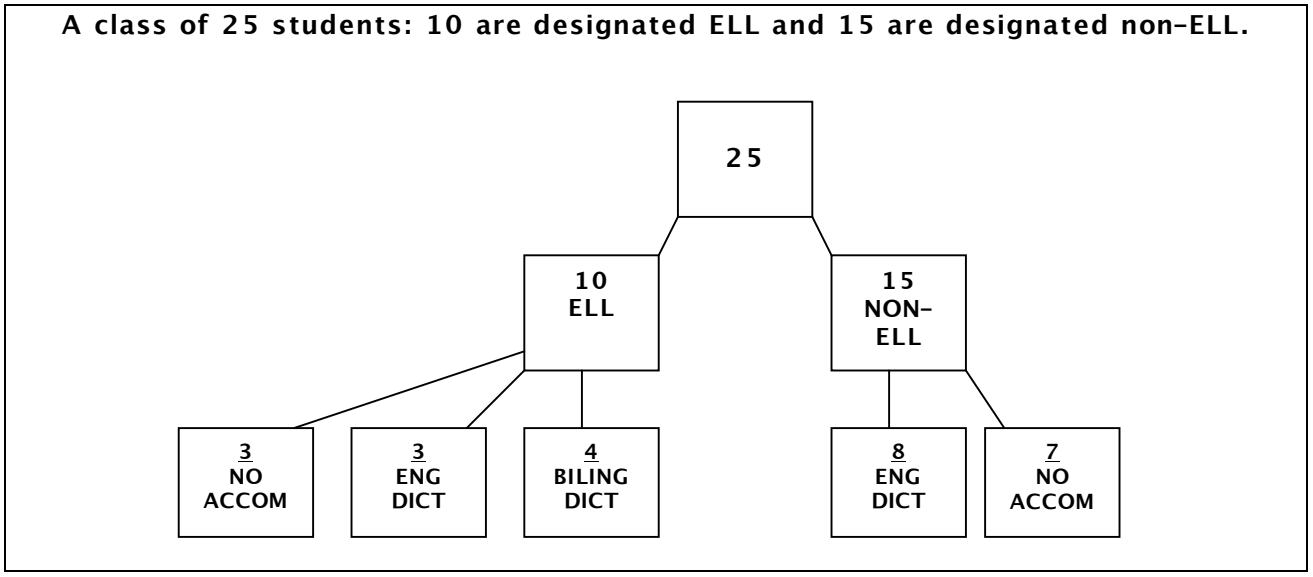

Figure 2. Example of accommodation distribution of two possible accommodations and no accommodation.

Administration of tests and questionnaires. There were two testing sessions per class, scheduled in the morning whenever possible. At the beginning of the first session, on Day 1, the science test was administered, followed by the accommodation follow-up questionnaire (beginning with the fourth school) and then the background questionnaire. Before the Grade 4 testing, test directions were read aloud. Beginning with the fourth school, students reviewed sample questions in both multiple-choice and open-ended formats.

Three accommodations were distributed randomly among the ELL students, and two of the accommodations (not bilingual dictionaries) were distributed randomly among the non-ELL students. Both ELL and non-ELL groups also contained students who received no accommodation, except for the extra time that was allotted to everyone.

To ensure consistent testing situations in the different classrooms, scripts for test administrators were prepared, tested early in the study, and revised after testing at the third school. There were scripts for each grade level and each day of testing. Test administrators were asked to observe the students, answer any of their questions, and write down the students' questions or comments throughout testing.

For administration of the questionnaires, instructions and items were read aloud to Grade 4 students. Only test instructions were read aloud to Grade 8 students during the study. To ensure accurate responses to the questionnaires, students with questionable or confusing responses were asked to clarify or correct them.

On Day 2, the NAEP reading assessment was administered in two 25-minute blocks. Directions were read aloud to each class, but no other accommodation was made. In one case, a class was not going to be available on Day 2 and took the reading assessment after a lunch break on Day 1.

Test administration personnel. Test administrators for the study were identified locally and consisted of off-track teachers, retired teachers, and graduate student researchers. All were trained by project staff to assure a standardized administration of the reading proficiency test and the accommodated standardized science test. They were compensated for their time and mileage accrued in traveling to test sites.

Scoring of open-ended items. All open-ended test items were scored by at least two teachers who were trained by the project staff. Open-ended science items

were scored by middle school science teachers or Grade 4 teachers, depending on the student's grade level. Middle school language arts teachers or Grade 4 teachers scored open-ended reading items depending on the grade level. Scorers were trained in the use of the NAEP scoring rubric. They were debriefed after each scoring session.

Results

As discussed earlier, several research questions guided the design and analyses of this study. The following research questions address issues concerning the effectiveness, validity, and differential impact of the accommodations.

Research Questions

1. Do accommodation strategies help reduce the performance gap between ELL and non-ELL students? (Effectiveness)

2. Do accommodation strategies impact the performance of non-ELL students on content-based assessment? (Validity)

3. Do student background variables impact performance on the accommodated assessments? (Differential impact)

4. Are accommodations easy to implement or use? (Feasibility)

Null Hypotheses and Alternative Hypotheses

The *null* hypotheses related to the research questions are:

H_{01}: In the science assessment, ELL students do not benefit from any of the accommodations used in this study. (Effectiveness)

H_{02}: Accommodations do not impact performance of non-ELL students on science tests. (Validity)

H_{03}: Student background variables do not impact performance on the accommodated science assessments. (Differential impact)

The *alternative* hypotheses corresponding to the null hypotheses are:

H_{11}: Some forms of accommodation are more effective than others in reducing the science performance gap between ELL and non-ELL students.

H_{21}: Accommodations impact performance of non-ELL students on science tests. The impact of accommodation on non-ELL students is the main concern with respect to the validity of accommodation. If there is a significant change in the performance of non-ELL students (increase or

decrease in their performance), then the outcome of the accommodated assessment may be confounded with the accommodation effects. That is, accommodation may actually alter the construct under measurement.

H_{31}: Student background variables impact performance on the accommodated assessments. If this is the case, then these background variables must be taken into consideration in making decisions about which accommodation to use with which students.

To test the above hypotheses concerning the use of accommodation, ELL and non-ELL students were tested under four levels of accommodation: an English dictionary, a bilingual dictionary, a linguistically modified version of the test items, and a standard condition in which no accommodation was provided.[2] Accommodations were randomly assigned to the ELL and non-ELL students within each classroom. Eight comparison groups were possible: 4 levels of accommodation by 2 levels of ELL status. Since there was no practical reason to give a bilingual dictionary to a non-ELL student, this group was excluded in the design. Table 4 illustrates the design for Grade 4, and Table 5 presents the design for Grade 8.

As Tables 4 and 5 show, there are two independent factors that may impact the outcome of the science assessment**:** type of accommodation (Factor A), and student ELL status (Factor B). Examining the main effect of Factor A will determine whether the accommodation strategies used in this study have any significant impact

Table 4

Grade 4 Design and Sample Size by Accommodation and ELL Status[a]

	ELL status		
Accommodation	ELL	Non-ELL	Total
Standard condition	N = 62	N = 85	N = 147
English dictionary	N = 59	N = 88	N = 147
Bilingual dictionary	N = 64	N/A	N = 64
Linguistically modified items	N = 20	N = 23	N = 43
Total	N = 205	N = 196	N = 401

Note. All conditions included extra time.

[a] The total number of students in Grade 4 was 406 and in Grade 8 the total number was 197. But due to missing data, the effective sample sizes depend on the analysis.

[2] All students had the same amount of extra time on the science assessment.

Table 5

Grade 8 Design and Sample Size by Accommodation and ELL Status[a]

	ELL status		
Accommodation	Non-ELL	Non-ELL	Total
Standard condition	N = 22	N = 33	N = 55
English dictionary	N = 25	N = 36	N = 61
Bilingual dictionary	N = 17	N/A	N = 17
Linguistically modified items	N = 13	Empty[b]	N = 13
Total	N = 77	N = 69	N = 146

Note. All conditions included extra time.

[a] The total number of students in Grade 4 was 406 and in Grade 8 the total number was 197. But due to missing data, the effective sample sizes depend on the analysis.

[b] Due to logistical problems, the linguistically modified version of the test was not provided to students in Grade 8.

on the outcome of assessment (science test score). Testing the main effect of Factor B will provide information on the performance difference between ELL and non-ELL students. Testing the interaction between Factors A and B will provide information about some of the hypotheses of this study (effectiveness and validity). A two-factor analysis of variance (ANOVA) model may be applied in this case. However, as Tables 4 and 5 show, the design (for both grades) is not fully crossed. For Grade 4, non-ELL students were not given a bilingual dictionary (Table 4). For Grade 8, there was an additional empty cell. Due to logistical problems, the linguistically modified version of the test could not be provided to non-ELL students in Grade 8. This resulted in two empty cells for the Grade 8 design.

Due to the design limitations discussed above, we were not able to use a two-factor fully-crossed ANOVA model. Instead, we used two one-factor models. To test the hypothesis concerning the validity of accommodation, we performed a one-factor analysis of covariance (ANCOVA). The ANCOVA model compared the mean of non-ELL students under different forms of accommodation (English dictionary, linguistically modified test version, and standard condition). To test the hypothesis concerning the effectiveness of accommodation, a similar analysis of covariance (ANCOVA) model compared the mean of ELL students under different forms of accommodation (English dictionary, bilingual dictionary, linguistically modified test version, and standard condition). To control for possible differences in reading

proficiency among students with different forms of accommodation, a reading score was used as a covariate in each model.

To locate the source of differences, we conducted multiple comparisons. The results of overall *F*-ratio and multiple comparisons will be presented later in this section.

Instruments

The focus of this study was the impact of accommodation on students' performance in science. Therefore, for each grade level, a test measuring students' science content knowledge provided data on the dependent variable. In addition to the science test, a reading comprehension test score was used as a covariate. Students' ELL status and type of accommodation were the main independent variables. Students' responses to background and accommodation follow-up questions were used as additional independent variables.

As indicated earlier, science test items were selected from the pool of released items from several main NAEP Science assessments. A science test booklet consisting of 19 items was constructed for Grade 4, and a test booklet of 30 items was constructed for Grade 8. Of the 19 items in the Grade 4 booklet, 11 were multiple-choice and 8 were open-ended format. Similarly, of the 30 items in the Grade 8 booklet, 24 were multiple-choice and 6 were open-ended format.

The selection of science test items for Grades 4 and 8 was made based on the NAEP item specifications and on the recommendations that we received from Grades 4 and 8 science teachers.

In addition to the science tests, we included a reading comprehension test for each grade. The reading comprehension test consisted of two 25-minute intact blocks of NAEP reading test items that were used in the 1994 NAEP Reading assessment (for Grade 4 or Grade 8 as appropriate). One class period (approximately 55 minutes) was allocated for the reading test. The reading test was used to obtain a measure of students' English language reading comprehension. The reading test scores were used as a covariate, to control for variation in the level of reading comprehension among the ELL and non-ELL students. There were multiple-choice and open-ended questions in the reading tests, for both Grades 4 and 8.

The science test booklets for students in both grades also included a background questionnaire and an accommodation follow-up questionnaire. The

background questionnaire contained questions on the students' language background, in addition to some demographic and opportunity-to-learn (OTL) questions.

The open-ended science test items were scored by experienced science teachers who were trained by the project staff. Two science teachers scored each open-ended science item. Tables 6 and 7 provide information on the type of instruments and number of items/questions in each instrument, for Grades 4 and 8 respectively. As Table 6 indicates, the science test for Grade 4 had 19 items, 11 of which were multiple-choice and 8 that were open-ended. The Grade 8 science test consisted of 30 items, of which 24 were multiple-choice and 6 were open-ended format.

Analyses of Open-Ended Questions, Grade 4 and Grade 8

As indicated earlier, each Grade 4 open-ended science item was scored independently by two Grade 4 teachers. Interrater reliability indices (percent of exact and within one-point agreement, PM correlation, intraclass correlation, kappa, and alpha coefficients) were computed using the Interrater Test Reliability System (Abedi, 1996). Table 8 summarizes the data on interrater reliability of open-ended science items for Grade 4.

Table 6

Grade 4 Test Booklets

	No. of items	No. of multiple-choice	No. of open-ended
NAEP science test	19	11	8
NAEP reading test	21	11	10
Background questionnaire	16	16	0
Follow-up questionnaire	7	6	1

Table 7

Grade 8 Test Booklets

	No. of items	No. of multiple-choice	No. of open-ended
NAEP science test	30	24	6
NAEP reading test	20	7	13
Background questionnaire	20	20	0
Follow-up questionnaire	7	6	1

Table 8

Grade 4 Interrater Reliability for Open-Ended Science Items

Item no.	Rater combination	No. of students	Kappa	Alpha	Exact agreement
1	1,2	406	.75	.87	87%
2	1,2	406	.42	.62	67%
3A	1,2	406	.57	.77	77%
3B	1,2	406	.41	.61	73%
5	1,2	406	.68	.81	82%
6	1,2	406	.47	.68	77%
7	1,2	406	.46	.73	68%
8	1,2	406	.30	.47	85%

In Table 8 we report kappa, alpha and percent of exact agreement. As data in Table 8 show, for some of the items, there are large discrepancies between the three interrater reliability indices. This is expected because the underlying theory and computational approaches are different for the different indices (see Abedi, 1996, for a discussion of differences between the different indices).

The main difference between percent of agreement and kappa is that percent of agreement is influenced by chance agreement, while kappa controls the variation due to chance agreement.

For the eight Grade 4 open-ended science items, percent of agreement ranged from a low of 67% (for item 2) to a high of 87% (for item 1). Kappa coefficient ranged from a low of .30 (for item 8) to a high of .75 (for item 1). Alpha coefficient ranged from a low of .47 (for item 8) to a high of .87 (for item 1). Looking at a combination of interrater reliability coefficients, one may conclude that some of the items were more difficult to score than others. In our future studies, for items with low interrater reliability, we plan to have more extensive training, or add more raters, or both.

For the Grade 4 science test, there was a wide range in the interrater reliability statistics among the test items and also a large discrepancy between different statistics on the same item. We indicated earlier that the discrepancy between the different statistics may be due to theoretical bases and computational approaches of the different statistics.

Table 9 presents interrater reliability statistics for the Grade 4 reading test. The individual interrater reliability statistics have a wide range across the reading test

Table 9

Grade 4 Interrater Reliability for Open-Ended Reading Items

Item no.	Rater combination	No. of students	Kappa	Alpha	Exact agreement
Blue Crab 1	1,2	406	.53	.69	77%
Blue Crab 4	1,2	406	.35	.74	58%
Blue Crab 6	1,2	406	.78	.88	93%
Blue Crab 8	1,2	406	.71	.83	86%
Blue Crab 10	1,2	406	.67	.80	83%
Hungry Spider 2	1,2	406	.71	.83	88%
Hungry Spider 4	1,2	406	.61	.76	81%
Hungry Spider 6	1,2	406	.48	.79	69%
Hungry Spider 8	1,2	406	.65	.79	83%
Hungry Spider 10	1,2	406	.67	.80	84%

items. Percent of exact agreement ranged from 58% for item Blue Crab 4 to 93% for item Blue Crab 6. Kappa coefficient ranged from .35 for item Blue Crab 4 to .78 for item Blue Crab 6, and alpha coefficient ranged from .69 for item Blue Crab 1 to .88 for item Blue Crab 6.

In general, interrater reliability coefficients were relatively high and suggest that the open-ended scoring was objective.

Table 10 presents the interrater reliability coefficients for the Grade 8 open-ended science items and Table 11 presents the interrater reliability results for the open-ended reading items. Once again, different interrater statistics provided different results. As the data in Table 10 suggest, most of the interrater reliability coefficients were high and indicate high agreement between the raters.

Table 10

Grade 8 Interrater Reliability for Open-Ended Science Items

Item no.	Rater combination	No. of students	Kappa	Alpha	Exact agreement
1	1,2	134	.85	.96	92%
2	1,2	133	.75	.86	84%
3	1,2	134	.86	.96	92%
4	1,2	134	.79	.89	88%
5	1,2	134	.88	.94	94%
6	1,2	134	.69	.77	83%

Table 11

Grade 8 Interrater Reliability for Open-Ended Reading Items

Item no.	Rater combination	No. of students	Kappa	Alpha	Exact agreement
Flying Machine 1	1,2	178	.66	.90	78%
Flying Machine 2	1,2	178	.72	.91	82%
Flying Machine 3	1,2	178	.81	.93	88%
Flying Machine 4	1,2	178	.59	.90	75%
Flying Machine 5	1,2	178	.72	.95	83%
Flying Machine 6	1,2	178	.66	.89	84%
Flying Machine 7	1,2	178	.69	.91	85%
Anasazi 1	1,2	178	.49	.85	67%
Anasazi 2	1,2	178	.47	.88	61%
Anasazi 3	1,2	178	.57	.91	70%
Anasazi 4	1,2	178	.60	.92	72%
Anasazi 5	1,2	178	.69	.93	80%
Anasazi 6	1,2	178	.67	.90	81%

Similarly, the results of interrater reliability analyses for the reading test in Grade 8 also suggest that there was a high level of agreement between the raters in scoring the open-ended reading items.

Examining the Internal Consistency of the Science and Reading Tests, Grade 4 and Grade 8

In classical test theory, if all the items on a test measure a single underlying construct, the test is uni-dimensional. In this case, the items should exhibit high internal consistency. To test the internal consistency of the reading and science tests, we computed Cronbach's coefficient alpha for both tests. Table 12 presents the internal consistency results for the reading and science tests for Grade 8. As the results in Table 12 show, the internal consistency coefficient for the entire set of Grade 8 reading items was .91**,** a very high internal consistency coefficient. This high coefficient suggests that the reading test is uni-dimensional and measures only one factor (reading comprehension). The two reading subscales also show a high level of internal consistency. For the first subscale (a text entitled "Flying Machine"), alpha was .86, and for the second subscale (a text entitled "Anasazi"), alpha was .85. For the science test, however, the internal consistency coefficient was low (.61), suggesting that the science test may be multidimensional.

Table 12

Grade 8 Internal Consistency Coefficients for Reading and Science Tests

Test	No. of items	No. of students	Alpha
Reading test			
Flying Machine—Reading passage	11	178	.86
Anasazi—Reading passage	9	178	.85
Total reading test	20	178	.91
Science test	30	146	.61

Table 13 presents the internal consistency results for the Grade 4 reading and science tests. These results are similar to those found for Grade 8. The reading test shows a higher level of internal consistency than the science test. However, compared with the internal consistency results for Grade 8, the Grade 4 reading test had lower internal consistency coefficients (see Table 13). The science test again had low internal consistency, suggesting that it may be multidimensional.

Testing Hypotheses Concerning Effectiveness and Validity of Accommodation

To test the effectiveness hypothesis, we compared the performance of ELL students who were provided the English dictionary, bilingual dictionary, or linguistically modified items accommodation in science with the performance of those ELL students who were tested under the standard NAEP condition. In this study, a significantly higher performance under any of the first three accommodations would indicate the effectiveness of that particular accommodation.

To test the validity hypothesis, we compared the performance of non-ELL students under the English dictionary, bilingual dictionary, and linguistically modified items accommodations with the performance of those non-ELL students who were tested under the standard NAEP condition. Any significant difference in

Table 13

Grade 4 Internal Consistency Coefficients for Reading and Science Tests

Test	No. of items	No. of students	Alpha
Reading test			
Blue Crabs—Reading passage	10	389	.74
Hungry Spider—Reading passage	11	389	.82
Total reading test	21	389	.87
Science test	19	320	.66

the performance of non-ELL students would suggest an impact of accommodation on the construct, thus creating concerns over the validity of accommodation.

Results for Grade 4 Students

Table 14 presents descriptive statistics for the Grade 4 science scores for each type of accommodation by ELL subgroup. As the table shows, ELL students had lower science test scores (M = 11.17, SD = 3.67, n = 205) than non-ELL students (M = 12.73, SD = 3.35, n = 196). There were differences between ELL and non-ELL performance under different forms of accommodation. For ELL students, the two dictionary accommodations and the linguistic modification accommodation seemed to make a difference. Comparing the performance of ELL students under those accommodations with that of ELL students under the standard condition, ELL students scored better under the two dictionary and the linguistic modification accommodations. For example, the mean science score for ELL students under the English dictionary condition was 11.97 (SD = 3.47, n = 59), compared with a mean of 10.04 (SD = 3.66, n = 62) for ELL students under the standard condition.

For non-ELL students, the kind of accommodation did not seem to make a difference. For students tested under the English dictionary condition, the mean score was 12.94 (SD = 3.54, n = 88). For students tested under the linguistically modified condition, the mean was 12.22 (SD = 3.37, n = 23), compared with a mean of 12.64 (SD = 3.16, n = 85) for students tested under the standard condition.

Table 14

Grade 4 NAEP Science Achievement Scores, Descriptive Statistics

Accommodation provided	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	10.04 (SD = 3.66; n = 62)	12.64 (SD = 3.16; n = 85)	11.54 (SD = 3.61; n = 147)
English dictionary	11.97 (SD = 3.47; n = 59)	12.94 (SD = 3.54; n = 88)	12.55 (SD = 3.54; n = 147)
Bilingual dictionary	11.72 (SD = 3.73; n = 64)		11.72 (SD = 3.73; n = 64)
Linguistically modified items	10.55 (SD = 3.37; n = 20)	12.22 (SD = 3.37; n = 23)	11.44 (SD = 3.44; n = 43)
Column total	11.17 (SD = 3.67; n = 205)	12.73 (SD = 3.35; n = 196)	11.93 (SD = 3.60; n = 401)

Note. 26 points possible. All conditions included extra time.

Table 15 presents the descriptive statistics for the Grade 4 reading test. The reading score was used as a covariate in the model comparing students' science scores under different forms of accommodation.

Effectiveness. To test the hypotheses concerning effectiveness of accommodation, we performed a one-factor analysis of variance. The ANOVA model compared the mean of ELL students under different forms of accommodation (English dictionary, bilingual dictionary, linguistically modified version, and standard condition). To control for a possible initial difference between students at different levels of accommodation, the reading score was used as a covariate. An *F*-ratio of 2.40 ($df = 3, 194$; $p = .07$), which was not statistically significant, suggested that the accommodation strategies that were used in this study did not have significant impact on students' performance. However, since the probability of a Type I error for this model (.07) was close to the .05 critical value, we performed multiple comparison analyses. Table 16 presents a summary of multiple comparison analyses. Of the three comparisons made, two were significant. There was a significant improvement in the score of ELL students over the standard condition when they were provided with an English dictionary. Also, ELL students who received a bilingual dictionary performed significantly higher than their peers under the standard condition (see Table 16).

Validity. To test the validity of accommodation, the performance of non-ELL students under accommodation was compared with the performance of students

Table 15

Grade 4 NAEP Reading Achievement Scores, Descriptive Statistics

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	9.20 ($SD = 5.62$; $n = 62$)	11.95 ($SD = 4.87$; $n = 81$)	10.76 ($SD = 5.37$; $n = 143$)
English dictionary	11.18 ($SD = 4.94$; $n = 57$)	13.16 ($SD = 5.54$; $n = 82$)	12.35 ($SD = 5.37$; $n = 139$)
Bilingual dictionary	10.36 ($SD = 5.37$; $n = 61$)		10.36 ($SD = 5.37$; $n = 61$)
Linguistically modified items	9.68 ($SD = 6.05$; $n = 19$)	11.57 ($SD = 5.69$; $n = 22$)	10.70 ($SD = 5.86$; $n = 41$)
Column total	10.17 ($SD = 5.42$; $n = 199$)	12.44 ($SD = 5.29$; $n = 185$)	11.26 ($SD = 5.47$; $n = 384$)

Note. 22 points possible.

Table 16

Grade 4 ELL Mean NAEP Science Achievement Scores Adjusted for Reading Achievement

Accommodation	ELL adjusted mean	Contrast with standard condition
Standard condition	10.49 (SE = .33; n = 62)	NA
English dictionary	11.45 (SE = .35; n = 57)	p = .05
Bilingual dictionary	11.62 (SE = .34; n = 61)	p = .02
Linguistically modified items	10.67 (SE = .60; n = 19)	p = .79

Note. 26 points possible. All conditions included extra time. SE = Standard error.

under the standard condition. Once again, to control for students' level of English proficiency, the reading score was used as a covariate. A nonsignificant F-ratio of .774 (df = 2, 181; p = .46) indicated that accommodation strategies did not change the performance of non-ELL students. That is, accommodation did not affect the validity of assessment.

Table 17 presents the adjusted means, standard errors, and numbers of students for the one-factor ANCOVA testing the validity of accommodation. Table 18 also presents level of significance for the multiple comparison tests, comparing different accommodations with the standard condition for non-ELL. Since the overall F test was not significant, these multiple comparisons do not provide very useful information.

Table 17

Grade 4 Non-ELL Mean NAEP Science Achievement Scores Adjusted for Reading Achievement

Accommodation	Non-ELL adjusted mean	Contrast with standard condition
Standard condition	12.99 (SE = .29; n = 54)	NA
English dictionary	12.65 (SE = .29; n = 57)	p = .42
Linguistically modified items	12.25 (SE = .57; n = 16)	p = .25

Note. 26 points possible. All conditions included extra time. SE = Standard error.

Results for Grade 8 Students

Table 18 presents the descriptive statistics for the Grade 8 science scores. On average, non-ELL students (M = 12.73, SD = 4.21, n = 69) outperformed ELL students (M = 10.94, SD = 3.61, n = 72) by about 2 points. Among the ELL students, the type of accommodation made a difference in test scoring. Students under the linguistically modified condition scored the highest (M = 13.27, SD = 3.04, n = 11), followed by students under the English dictionary condition (M = 11.52, SD = 3.53, n = 23) and the standard condition (M = 10.32, SD = 3.99, n = 22). Students under the bilingual dictionary condition scored the lowest (M = 9.38, SD = 2.69, n = 16). Among the non-ELL sample, students under the English dictionary accommodation (M = 12.64, SD = 4.29, n = 33) scored about the same as students under the standard condition (M = 12.83, SD = 4.19, n = 36).

Table 18

Grade 8 NAEP Science Achievement Scores, Descriptive Statistics

Accommodation provided	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	10.32 (SD = 3.99; n = 22)	12.83 (SD = 4.29; n = 33)	11.83 (SD = 4.32; n = 55)
English dictionary	11.52 (SD = 3.53; n = 23)	12.64 (SD = 4.19; n = 36)	12.20 (SD = 3.95; n = 59)
Bilingual dictionary	9.38 (SD = 2.69; n = 16)	N/A	9.38 (SD = 2.69; n = 16)
Linguistically modified items	13.27 (SD = 3.04; n = 11)	Empty	13.27 (SD = 3.04; n = 11)
Column total	10.94 (SD = 3.61; n = 72)	12.73 (SD = 4.21; n = 69)	11.82 (SD = 4.00; n = 141)

Note. 36 points possible. All conditions included extra time.

Table 19 presents the descriptive statistics for the Grade 8 reading test. The reading score was used as a covariate in the model comparing students' science scores under different forms of accommodation.

Effectiveness. To test the effectiveness hypothesis for Grade 8, the performance of ELL students under accommodation was compared with the performance of ELL students under the standard NAEP condition, using a one-factor ANOVA model. In this model, reading score was used as a covariate. An F-ratio of 2.88 (df = 3, 67; p = .04), which was significant at the .05 nominal level, suggested that some

Table 19

Grade 8 NAEP Reading Achievement Scores, Descriptive Statistics

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	4.14 (SD = 3.52; n = 22)	10.12 (SD = 5.90; n = 33)	7.73 (SD = 5.85; n = 55)
English dictionary	4.89 (SD = 3.14; n = 23)	9.93 (SD = 5.94; n = 36)	7.97 (SD = 5.58; n = 59)
Bilingual dictionary	4.13 (SD = 2.45; n = 16)	N/A	4.13 (SD = 2.45; n = 16)
Linguistically modified items	5.18 (SD = 5.03; n = 11)	Empty	5.18 (SD = 5.03; n = 11)
Column total	4.53 (SD = 3.43; n = 72)	10.02 (SD = 5.88; n = 69)	7.22 (SD = 5.51; n = 141)

Note. 27 points possible.

accommodations significantly impacted ELL students' performance. Table 20 presents a summary of multiple post hoc comparison analyses. As the data show, the only accommodation that improved Grade 8 ELL students' performance was the linguistic modification of test items. The adjusted mean science score for ELL students under the linguistically modified condition, was 13.00 (SE = .95, n = 11), compared with a mean of 10.49 (SE = .67, n = 22) under the standard condition.

Validity. To test the validity hypothesis for the Grade 8 data, we compared the performance of non-ELL students under the different accommodations (see Table 21). Once again, provision of accommodation should not alter the construct (i.e.,

Table 20

Grade 8 ELL Mean NAEP Science Achievement Scores Adjusted for Reading Achievement

Accommodation	ELL adjusted means	Contrast with standard condition
Standard condition	10.49 (SE = .67; n = 22)	N/A
English dictionary	11.37 (SE = .66; n = 23)	p = .35
Bilingual dictionary	9.55 (SE = .79; n = 16)	p = .37
Linguistically modified items	13.00 (SE = .95; n = 11)	p = .04

Note. 36 points possible. All conditions included extra time. SE = Standard error.

Table 21

Grade 8 Non-ELL Mean NAEP Science Achievement Scores Adjusted for Reading Achievement

Accommodation	Adjusted means	Contrast with standard condition
Standard condition	12.80 (SE = .65; n = 36)	N/A
English dictionary	12.67 (SE = .62; n = 33)	p = .89

Note. Both conditions included extra time. SE = Standard error.

should not affect the performance of the native speakers of English). Comparing the accommodated performance of non-ELL students (i.e., Dictionary and Extra Time) with the standard NAEP condition yielded an F-ratio of .020 (df = 1, 66; p = .89), which was not statistically significant. The results indicate that the accommodation did not affect the validity of the science measure.

Background Questionnaire

As discussed earlier, in addition to NAEP science and reading tests, we asked students to respond to a set of background questions. Student responses to these questions provided additional information for our research hypotheses. We will present the results of the background questions for Grade 4 students first and then for Grade 8 students.

Grade 4 background questions. Tables 22 and 23 present frequencies of responses to the different background questions for all students in Grade 4.

As data in Table 22 show, in response to Question 1, a majority of students (70.9%) indicated that they had been born in the U.S. whereas 29.1% reported that they had been born in other countries. Consistent with the response pattern for Question 1, in response to Question 2, 69% of the students indicated that they had lived in the U.S. "All my life" while 31% reported otherwise.

Of the total sample, 55.9% indicated that they initially attended preschool in the U.S. and 29% indicated that they initially attended kindergarten in the U.S. Only 15% responded that they initially attended Grade 1 or higher in the U.S. In response to the question of whether they have been to a school outside the U.S. (Question 4), a large majority of students (80.2%) responded "No, never."

Table 22

Grade 4 Frequencies and Percentages for Student Background Questionnaire, Questions 1–4

Questions and responses	Frequency	%
1. Country of birth:		
Cambodia	0	0
China	10	2.5
Cuba	0	0
Korea	10	2.5
Mexico	20	4.9
Puerto Rico	0	0
Taiwan	1	0.2
United States	288	70.9
Other	77	19.0
2. Time lived in US:		
Less than 1 year	8	2.0
1 year	4	1.0
2 years	12	3.0
3 years	14	3.4
4 years	14	3.4
More than 4 years	74	18.2
All my life	280	69.0
3. Initial grade attended in US:		
Preschool	226	55.9
Kindergarten	117	29.0
1st grade	26	6.4
2nd grade	14	3.5
3rd grade	8	2.0
4th grade	13	3.2
4. Have been to a school outside the US:		
No, never	324	80.2
Yes, in the country of birth	63	15.6
Yes, in the country not of birth	17	4.2

To explore the pattern of language use at home, students were asked to indicate what language(s) (including English) they spoke at home before they started school (Table 23). In this question, students were allowed to select multiple languages. Thus, frequencies for this question are larger than the total *n*. A majority of respondents (68.7%) selected English as the language spoken in the home. Spanish was the next largest category with 28.6% of the responses. Other languages such as Chinese (12.8%), Korean (10.6%), and Tagalog (8.9%) were also named.

Table 23

Grade 4 Frequencies and Percentages for Student Background Questionnaire, Questions 5–10

Questions and responses	Frequency	%
5. Before starting school, language spoken at home: (you may choose more than one language)[a]		
Chinese	52	12.8
English	279	68.7
Khmer	1	0.2
Korean	43	10.6
Spanish	116	28.6
Tagalog	36	8.9
Other	104	25.6
7. What language other than English do you speak at home now?		
None	93	23.0
Chinese	39	9.7
Khmer	2	0.5
Korean	41	10.1
Spanish	115	28.5
Tagalog	19	4.7
Other	95	23.5
	M	*SD*
6. How well can you understand spoken English at school?	3.55	0.65
8. How well do you speak the other language at home?	3.39	0.76
9. How well do you read the other language at home?	2.91	1.03
10. How well do you write the other language at home?	2.75	1.06

[a] Since selecting multiple responses was permissible, the sum of frequencies and percentages will be higher that the totals reported for other questions.

Since the language currently spoken in the home was of interest, students were asked to indicate the language other than English they spoke at home. As the data in Table 23 show, the response pattern to this question is consistent with the pattern in Question 5. Of the languages other than English, Spanish had the highest frequency of use (28.5%) followed by Korean (10.1%), Chinese (9.7%) and Tagalog (4.7%).

Table 23 also shows the means and standard deviations for the Likert-type questions that we asked both Grade 4 and Grade 8 students. We asked ELL students to rate their level of understanding spoken English at school on a 4-point scale ranging from 4 (*very well*) to 1 (*not at all*). The mean rating for this question (Question 6) was 3.55 (SD = .65) out of a maximum of 4, indicating that ELL students in this study felt that they understood spoken English at school relatively well.

Using the same Likert scale, we asked ELL students to indicate how well they spoke the other language at home (Question 8). The mean rating for this question was 3.39 (*SD* = .76), suggesting that ELL students spoke the other language well at home while maintaining a good understanding of English at school, as presented by the data for Question 6.

Similar to Question 8, Question 9 asked ELL students how well they read the other language at home, and Question 10 asked how well they wrote the other language at home. The mean rating for Question 9 was 2.91 (*SD* = 1.03), which is still above the scale midpoint of 2.5, but was not as high as the mean was for Question 8. The mean rating for Question 10 (*M* = 2.75, *SD* = 1.06) also was not as high as for Question 8. These data suggest that ELL students spoke, read, and wrote the other language at home relatively well, but they had a better ability to speak the language than to read or write it.

To examine the response pattern across the ELL categories, responses of ELL and non-ELL students were compared. Because the frequencies in some of the response categories, such as Asian languages, were small and would be even smaller if divided across the two ELL categories, we combined some of the response categories. Table 24 presents frequencies and percentages by ELL status. As data in Table 24 suggest, ELL and non-ELL students have significantly different response patterns on some questions. For example, time lived in the United States was significantly different across students' ELL status. Ninety-five percent of non-ELL students indicated that they have lived in the U.S. for more than 4 years or their entire life as compared to 79.5% of ELL students.

Similarly, the data in Table 24 indicate ELL and non-ELL differences in the initial grade in which students attended school in the United States. More than 92% of the non-ELL students indicated that they initially attended preschool in the U.S. as compared to about 78% of ELL students. When both ELL and non-ELL students were asked whether they had been to a school outside the U.S., more non-ELL students (88.4%) responded "No" than ELL students (72.1%).

In response to the question "What language other than English do you speak at home now?" (Question 7), more non-ELL students selected the "None" category (40.1%) than ELL students (6.8%). As expected, more ELL students selected Spanish (39.5%) than non-ELL students (17.3%). Similarly, more ELL students (35.1%) selected an Asian language than non-ELL students (14.2%).

Table 24

Grade 4 Student Background Questionnaire by ELL Status

Questions and responses	ELL		Non-ELL	
	Frequency	%	Frequency	%
1. Country of birth:				
U.S.	128	62.4	158	79.4
Others	77	37.6	41	20.6
2. Time lived in US:				
1 year or less	11	5.4	1	0.5
2 – 4 years	31	15.1	9	4.5
More than 4 years (all my life)	163	79.5	189	95.0
3. Initial grade attended school in US:				
Preschool	159	77.6	182	92.4
Kindergarten	27	13.2	13	6.6
1st grade—4th grade	19	9.3	2	1.0
4. Have you been to a school outside the US?				
No, never	147	72.1	175	88.4
Yes	57	27.9	23	11.6
7. What language other than English do you speak at home now?				
None	14	6.8	79	40.1
Spanish	81	39.5	34	17.3
Asian (Chinese, Korean, Khmer, Tagalog)	72	35.1	28	14.2
Other	38	18.5	56	28.4

Grade 8 background questions. Table 25 presents frequencies and percentages for all Grade 8 students (ELL and non-ELL). As shown in Table 25, 57.2% of students indicated that they were born in the U.S., followed by 14.5% in Korea and 8.1% in Mexico. A number of students (15.6%) marked "other" countries (Question 1).

Consistent with the responses to Question 1, a majority of Grade 8 students indicated that they had lived in the U.S. their entire life (57.2%). In fact, the percentage of students marking U.S. as their country of birth (57.2%) in Question 1 was identical with the percentage of students indicating they have lived in the U.S. "All my life" (57.2%) in Question 2. This would provide some evidence of high parallel form reliability. In response to the question about "initial grade attended school in the U.S.," a large number of respondents (45.6%) indicated that they had been attending school in the U.S. since preschool. The next largest response category was kindergarten with 15.2%. Comparing these percentages with the responses from Questions 1 and 2, once again, reliability or consistency of the responses can be

Table 25

Grade 8 Frequencies and Percentages for Student Background Questionnaire, Questions 1–3, 8

Questions and responses	Frequency	%
1. Country of birth:		
Cambodia	0	0
China	8	4.6
Cuba	0	0
Korea	25	14.5
Mexico	14	8.1
Puerto Rico	0	0
Taiwan	0	0
United States	99	57.2
Other	27	15.6
2. Time lived in US:		
Less than 1 year	7	4.0
1 year	18	10.4
2 years	16	9.2
3 years	8	4.6
4 years	6	3.5
More than 4 years	19	11.0
All my life	99	57.2
3. Initial grade attended school in US:		
Preschool	78	45.6
Kindergarten	26	15.2
1st grade	3	1.8
2nd grade	3	1.8
3rd grade	6	3.5
4th grade	2	1.2
5th grade	10	5.8
6th grade	14	8.2
7th grade	21	12.3
8th grade	8	4.7
8. Have been to a school outside the US:		
No, never	107	60.5
Yes, in the country of birth	55	31.1
Yes, in the country not of birth	9	5.1

observed. A total of 67 students (39.2%) selected other grades (Grades 1 to 8) as their initial grade of school attended in the U.S. (see Table 25).

Responses to Question 8 (see Table 25) show that a majority of respondents (60.5%) indicated that they had never been to a school outside the U.S., whereas 36.2% of the respondents said that they had studied in schools outside the U.S.

As discussed in the literature review, students' language background factors have substantial impact on their performance in content-based subject areas. The type and amount of language spoken in the home are among the most influential language background variables. Due to the importance of this factor, we included some questions about it on the questionnaire (Table 26). We asked students to identify what language they spoke at home before they started school (Question 9) and what language they speak at home now (Question 11).

In response to Question 9, a large group of students (42.4%) indicated that they spoke Spanish at home before starting school. About the same percentage of students (41.1%) said that they spoke English at home before they started school and 33.4% of the students selected other languages. (Note that students selected more than one response to this question; thus, the total of the percentages is greater than 100%.) ELL students were also asked to report any language other than English currently spoken in their home (Question 11). Of the 167 respondents to this question, 75 (44.9%) selected Spanish, 39 (23.4%) said "None" (i.e., no language other than English being spoken in their home), 25 (15.0%) selected Korean, and a total of 28 (16.8%) selected other languages.

To understand students' impressions of the importance of language factors in assessment, we asked (Question 10), "After reading a book at school, which would you be able to do?" We directed students to respond to all three of the choices: (a) an oral book report (in which oral language proficiencies would be required), (b) a written book report (which needs writing skills), and (c) take a multiple-choice test (which might be easier than the two other formats). Of the respondents to Question 10, 76 (43.9%) indicated that they would be able to do an oral book report, 109 (63%) said they would be able to do a written book report, and 83 (48%) said they would be able to take a multiple-choice test. Once again, for this question, multiple responses were selected.

Finally, students were asked to self-report their level of English proficiency (speaking, reading, writing, and understanding) when they first started school (Questions 4–6) and at their current grade in school (Question 7). They were also asked to self-report their level of proficiency in the other language, if spoken at home (Questions 12–14). These questions were all in Likert format ranging from 1 (*not at all*) to 4 (*very well*). The midpoint for this range is 2.5. As data in Table 26

Table 26

Grade 8 Frequencies and Percentages for Student Background Questionnaire, Questions 4–7 and 9–14

Question and responses	Frequency	%
9. Before starting school, language spoken at home: (you may choose more than one language)		
Chinese	9	5.1
English	73	41.1
Khmer	4	2.3
Korean	26	14.7
Spanish	75	42.4
Tagalog	1	0.6
Other	19	10.7
10. After reading a book at school, which would you be able to do?		
Oral book report	76	43.9
Written book report	109	63.0
Multiple-choice test	83	48.0
11. What language other than English do you speak at home now?		
None	39	23.4
Chinese	4	2.4
Khmer	1	0.6
Korean	25	15.0
Spanish	75	44.9
Tagalog	1	0.6
Other	22	13.2
	M	*SD*
4. How well did you speak English when you first started school in the U.S.?	2.79	1.07
5. How well did you read English when you first started school in the U.S.?	2.63	1.01
6. How well did you write English when you first started school in the U.S.?	2.56	1.03
7. How well can you understand spoken English at school?	3.26	0.91
12. How well do you speak the other language at home?	3.36	0.75
13. How well do you read the other language at home?	3.16	0.90
14. How well do you write the other language at home?	3.01	0.87

show, mean scores for all of the responses are above the midpoint of 2.5.[3] However, some questions have relatively lower means. For example, the means for self-reported English proficiency (Questions 4, 5, 6, and 7) are generally lower than the means for proficiency in the other language (Questions 12, 13, and 14). This may reflect the fact that a large number of ELL students were included in this sample.

[3] Even though the numerical values for these questions are on an ordinal scale of measurement, we treated them as continuous scales and computed the means and standard deviations.

Table 27 presents frequencies and percentages for the responses to the background questions by student ELL status. As indicated earlier in reporting the Grade 4 results, due to the small number of participants, we combined some of the categories. In general, the response patterns for ELL and non-ELL students are very different on the questions related to language background or place of residency. Following are some general findings:

1. More ELL students are born outside the U.S. (71.3%) than non-ELL students (18.3%).

2. A higher percentage of non-ELL students indicated that they had lived in the U.S. more than 4 years or "All my life" (93.5%) than ELL students (38.8%).

3. More non-ELL students indicated that they initially started with preschool in the U.S. (73.1%) than ELL students (12.8%).

4. Most of the non-ELL students indicated that they had not been to a school outside the U.S. (88%) as compared with a smaller percentage of ELL students (32.9%).

5. More non-ELL students indicated that they would be able to do an oral book report (Non-ELL: 57%, ELL: 28.8%); a written book report (Non-ELL: 73.1%, ELL: 51.3%), and take a multiple-choice test (Non-ELL: 62.4%, ELL: 31.3%) after reading a book at school.

6. Non-ELL students self-reported a higher level of English proficiency than ELL students.

Background Variable Impact on Science Performance

Results for Grade 4. To examine the impact of background variables on science performance and to identify variables with a greater level of impact, multiple regression analysis was used. Several models were created to test the power of background variables in predicting students' performance in science under different accommodation conditions. Hence, the science test score was used as the criterion variable and background variables as predictor variables. Table 28 lists background variables used in the multiple regression models as predictors.

Two multiple regression models were used: one for students under the standard NAEP condition and one for the English dictionary condition. Table 29 summarizes the results of multiple regression for the two models. As Table 29 shows, the percent of variance explained by the model (R^2) is larger for students under the standard NAEP condition ($R^2 = .301$) than for students under the English

Table 27

Grade 8 Student Background Questionnaire by ELL Status, Questions 1–8 and 10–14

Question and response(s)	ELL		Non-ELL	
	Frequency	%	Frequency	%
1. Country of birth:				
U.S.	23	28.8	76	81.7
Others	57	71.3	17	18.3
2. Time lived in US:				
1 year or less	24	30.0	1	1.1
2 – 4 years	25	31.3	5	5.4
More than 4 years (all my life)	31	38.8	87	93.5
3. Initial grade attended school in US:				
Preschool	10	12.8	68	73.1
Kindergarten	11	14.1	15	16.1
1st Grade – 4th Grade	11	14.1	3	3.2
5th Grade – 8th Grade	46	59.0	7	7.5
8. Have been to a school outside the US:				
No, never	26	32.9	81	88.0
Yes	53	67.1	11	12.0
11. What language other than English do you speak at home now?				
None	6	7.8	33	36.7
Spanish	30	39.0	45	50.0
Asian (Chinese, Korean, Khmer, Tagalog)	27	35.1	4	4.4
Other	14	18.2	8	8.9
10. After reading a book at school, which would you be able to do?				
Oral book report	23	28.8	53	57.0
Written book report	41	51.3	68	73.1
Multiple-choice test	25	31.3	58	62.4
	M	*SD*	*M*	*SD*
4. How well did you speak English when you first started school in the U.S.?	2.25	1.01	3.26	0.89
5. How well did you read English when you first started school in the U.S.?	2.18	0.84	3.02	0.98
6. How well did you write English when you first started school in the U.S.?	2.13	0.85	2.92	1.03
7. How well can you understand spoken English at school?	2.79	0.92	3.66	0.65
12. How well do you speak the other language at home?	3.39	0.78	3.33	0.71
13. How well do you read the other language at home?	3.21	0.88	3.10	0.92
14. How well do you write the other language at home?	3.06	0.83	2.95	0.92

Table 28

Grade 4 Student Background Variables Used in Multiple Regression Models

Question
Time lived in U.S.
Language other than English spoken at home was Chinese
Language spoken at home was English
Language other than English spoken at home was Korean
Language other than English spoken at home was Spanish
Language other than English spoken at home was Tagalog
Language other than English spoken at home was "Other"
How well can you understand spoken English at school?
How well do you speak, read, and write the other language at home? (composite)
ELL status

Table 29

Grade 4: Predicting Science Test Scores From Background Variables

Accommodation	R^2	*F*	*DF*	Significance
Standard NAEP condition	.301	4.646	10, 108	<.001
English dictionary	.133	1.876	9, 110	.063

Note. Both conditions included extra time.

dictionary condition (R^2 = .133). This finding suggests that the English dictionary accommodation may remove some of the language barriers that confound science content assessment.

Results for Grade 8. Regression models similar to those for Grade 4 were used in Grade 8 to test the power of background variables in predicting science performance. Separate multiple regression models were used to test the prediction under the standard NAEP condition and under the English dictionary condition. Table 30 lists background variables used as predictors for Grade 8 students.

Table 31 summarizes the results of the multiple regression analysis for Grade 8. The R^2 for students under the standard NAEP condition (R^2 = .279) is slightly higher than the R^2 under the dictionary accommodation (R^2 = .250). These results are consistent with those reported for Grade 4. However, because of sample size in Grade 8, the results did not reach the standard .05 significance level. We will examine this issue in a future study, where we will have a larger number of Grade 8 students.

Table 30

Grade 8 Student Background Variables Used in Multiple Regression Models

Question
Time lived in U.S.
Language spoken at home was English only
Language other than English spoken at home was Khmer
Language other than English spoken at home was Korean
Language other than English spoken at home was Spanish
Language other than English spoken at home was Tagalog
Language other than English spoken at home was Chinese
Language other than English spoken at home was "Other"
How well do you speak, read, and write the other language at home? (composite)
ELL status

Table 31

Grade 8: Predicting Science Test Scores From Background Variables

Accommodation	R^2	*F*	*DF*	Significance
Standard NAEP condition	.279	1.661	7, 30	.157
English dictionary	.250	1.234	10, 37	.302

Note. Both conditions included extra time.

Follow-Up Questionnaire Results

To obtain data on accommodations from students' perspectives, we developed and used accommodation follow-up questionnaires. The same accommodation follow-up questions were used for each of the four conditions, and the purpose was to collect data on students' reactions to the provision of accommodations. We will present the results of the follow-up questions for Grade 4 students first and then for Grade 8 students.

Grade 4 follow-up questions. Grade 4 students were asked if they had been given a dictionary. If they received a dictionary, they then were asked to indicate how effective and how useful the dictionary had been. Table 32 presents the follow-up questions used for Grade 4 students.

To examine the pattern of responses across the ELL categories (comparing responses of ELL students with non-ELL students), frequencies of responses to the follow-up questions were obtained separately for each group. The first question

Table 32

Grade 4 Follow-Up Questions

Question
1. Were you given a dictionary or glossary for this test?
2. In the science test, were there words that you did not understand?
3. Did you use the dictionary during the test to find words?
4. Did the dictionary or glossary help you understand the questions?
5. Would it help if the test explained words in another language?
6. Would it help if the test used easier words?

asked "Were you given a dictionary or glossary for this test?" The response to this question was either "dictionary or glossary" or "none." If no accommodation were provided to the student, he or she would select the "None" category and would not respond to the rest of follow-up questions.

Follow-up Question 2 asked students if there were words that they did not understand on the science test. Response options to this question were "No," there were not any words that I did not understand; "Yes, some," there were a few words that I did not understand; and "Yes, many," there were many words that I did not understand on the science test. These response options were in a Likert format, which allowed us to compute a mean rating for the questions. Thus, the mean for each question would range between 1 (*no difficulty*) and 3 (*many difficult words*), with a midpoint of 2. To compare the response patterns of ELL and non-ELL students, we present the data for both groups.

Table 33 presents data for follow-up Question 2. The overall mean for non-ELL students was 1.50 (SD = .57, n = 156), and for ELL students the mean was 1.71 (SD = .59, n = 185). The mean for ELL students was higher than the mean for non-ELL students, which suggests that ELL students found more words that they did not understand. The mean difference between ELL and non-ELL students varies across the different accommodation conditions. The largest difference between the means for ELL and non-ELL students was on the standard condition (more than .6 standard deviation) where no accommodation was provided. There was also a difference between ELL and non-ELL students on the dictionary condition (.39 standard deviation). However, the means for the ELL and non-ELL students are identical for the linguistically modified version of the test. This result supports the accuracy of the procedure used to linguistically modify the test items.

Table 33

Grade 4 Descriptive Statistics for "Were there words that you did not understand?"

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	1.84 (*SD* = .63; *n* = 56)	1.47 (*SD* = .56; *n* = 64)	1.64 (*SD* = 62; *n* = 120)
English dictionary	1.74 (*SD* = .59; *n* = 53)	1.51 (*SD* = .58; *n* = 70)	1.61 (*SD* = .60; *n* = 123)
Bilingual dictionary	1.61 (*SD* = .56; *n* = 56)	NA	1.61 (*SD* = .56; *n* = 56)
Linguistically modified items	1.55 (*SD* = .51; *n* = 20)	1.55 (*SD* = .60; *n* = 22)	1.55 (*SD* = .55; *n* = 42)
Column total	1.71 (*SD* = .59; *n* = 185)	1.50 (*SD* = .57; *n* = 156)	1.61 (*SD* = .59; *n* = 341)

Note. 1= No; 2 = Yes, some; 3 = Yes, many. All conditions included extra time.

To examine the significance of the difference between ELL and non-ELL responses to Question 2, analysis of variance was used. Because there was an empty cell in the two-way cross-classification (see Table 33), we were not able to use a two-factor fully-crossed ANOVA model. Instead, we used a one-way ANOVA to compare the seven group means, with no particular attention to the main effects of type of accommodation and students' ELL status.

For Table 33, the results of the overall one-way ANOVA with 7 groups indicated that there are differences among the groups (F = 2.887; df = 6, 334; p = .009). ELL students found more words on the science test that they could not understand than did non-ELL students (t = –3.284; df = 339; p = .001). Among the ELL students the differences in the means for the various accommodation groups were not significant (F = 2.014; df = 3, 181; p = .114,). It should be noted, however, that the difference observed between the standard condition (M = 1.84) and the linguistically modified items condition (M = 1.55) may very well be significant with an increase in the number of participants.

The third follow-up question asked Grade 4 students whether they used a dictionary or glossary to find difficult words. Response options were "No," did not use dictionary; "Yes, some," used dictionary sometimes; and "Yes, a lot," used dictionary a lot. Table 34 presents descriptive statistics for this question across the ELL categories. Responses for this question were coded from 1 (*no dictionary use*) to 3 (*used dictionary a lot*). As Table 34 shows, the overall mean for this question was 1.56,

Table 34

Grade 4 Descriptive Statistics for "Did you use the dictionary?"

Accommodation provided	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
English dictionary	1.58 (*SD* = .54; *n* = 52)	1.52 (*SD* = .56; *n* = 69)	1.55 (*SD* = .55; *n* = 121)
Bilingual dictionary	1.59 (*SD* = .56; *n* = 56)	NA	1.59 (*SD* = .56; *n* = 56)
Column total	1.58 (*SD* = .55; *n* = 108)	1.52 (*SD* = .56; *n* = 69)	1.56 (*SD* = .55; *n* = 177)

Note. 1 = No; 2 = Yes, some; 3 = Yes, a lot. Both conditions included extra time.

which is smaller than the midpoint of 2, indicating that students in this sample rarely used a dictionary during the science test. Under the English dictionary accommodation, ELL students had a slightly higher mean (M = 1.58, SD = .54, n = 52) than non-ELL students (M = 1.52, SD = .56, n = 69). For the bilingual dictionary accommodation, there was no comparison group. There was no significant difference in the amount of dictionary usage as self-reported by the three groups of students in Table 34 (F = .267; df = 2, 174; p = .766,). Only those students who actually received a dictionary are included in this analysis.

Follow-up Question 4 asked Grade 4 students if the dictionary or glossary helped them understand the science questions. Table 35 shows the means for the responses to this question. Only students who received some type of dictionary are included in the analysis. The mean for ELL students under the dictionary accommodation (M = 1.77, SD = .82, n = 53) was higher than the mean for non-ELL students (M = 1.64, SD = .71, n = 67). This result suggests that among those students who used a dictionary, ELL students believed that it helped them more so than non-ELL students. However, the difference between the means of the ELL and non-ELL groups did not reach statistical significance (F = .946; df = 2, 170; p = .390).

Table 36 reports the data on whether Grade 4 students believed that an explanation of words in another language would have benefited them on the science test. The overall mean was 1.61 (SD = .71, n = 333) which is lower than the midpoint of 2 for this question. These results suggest that, in general, students in this study believed that explaining words in another language would not help their performance on the science test. However, there were relatively large gaps between the means for ELL and non-ELL students. For example, the mean for ELL students

Table 35

Grade 4 Descriptive Statistics for "Did the dictionary help?"

Accommodation provided	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
English dictionary	1.77 (*SD* = .82; *n* = 53)	1.64 (*SD* = .71; *n* = 67)	1.70 (*SD* = .76; *n* = 120)
Bilingual dictionary	1.83 (*SD* = .80; *n* = 53)	NA	1.83 (*SD* = .80; *n* = 53)
Column total	1.80 (*SD* = .81; *n* = 106)	1.64 (*SD* = .71; *n* = 67)	1.74 (*SD* = .77; *n* = 173)

Note. 1 = No; 2 = Yes, some; 3 = Yes, a lot. Both conditions included extra time.

under the standard condition was 2.07 (*SD* = .79, *n* = 54) as compared with a mean of 1.37 (*SD* = .52, *n* = 59) for non-ELL students, a difference of more than one standard deviation. ELL students believed that explaining words in another language would help. ELL students tested under other accommodations also showed higher means for this question.

The results of the overall one-way ANOVA with seven groups indicated that there are differences among the groups ($F = 8.60$; $df = 6, 328$; $p < .001$). As one would expect, ELL students felt that explanation in another language would be of more benefit than did non-ELL students ($t = –6.319$; $df = 333$; $p < .001$). Among the ELL students there were differences in the means for the various accommodation groups

Table 36

Grade 4 Descriptive Statistics for "Would explanation in another language help?"

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	2.07 (*SD* = .79; *n* = 54)	1.37 (*SD* = .52; *n* = 59)	1.71 (*SD* = .75; *n* = 113)
English dictionary	1.77 (*SD* = .75; *n* = 53)	1.43 (*SD* = .63; *n* = 70)	1.74 (*SD* = .44; *n* = 121)
Bilingual dictionary	1.88 (*SD* = .72; *n* = 56)	NA	1.88 (*SD* = .72; *n* = 56)
Linguistically modified items	1.55 (*SD* = .51; *n* = 20)	1.39 (*SD* = .58; *n* = 23)	1.47 (*SD* = .55; *n* = 43)
Column total	1.87 (*SD* = .74; *n* = 183)	1.40 (*SD* = .58; *n* = 152)	1.61 (*SD* = .71; *n* = 333)

Note. 1= No; 2 = Maybe; 3 = Yes. All conditions included extra time.

(F = 8.60; df = 6, 328; p < .001). ELL students who received English dictionaries (M = 1.77, SD = .75, n = 53) and those who received linguistically modified items (M = 1.55, SD = .51, n = 20) felt that explanation in another language would be of less help to them, in contrast to ELL students under the standard condition (M = 2.07, SD = .79, n = 54).

"Would it help if the test used easier words?" was the sixth follow-up question (Table 37). The responses to this question were "Yes, definitely" or "No." "No" was coded 1 and "Yes" was coded 2. For ease of discussion and interpretation, we also computed the mean for this question. The mean ranged from 1 (easier words do not help students' performance on the science test) to 2 (easier words definitely help students' performance).

Table 37 presents descriptive statistics for the responses to Question 6. The overall mean was 1.73 (SD = .45, n = 340). The mean was higher than the midpoint of 1.5, which indicates that students in general believed that using easier words would help to improve their performance on the science test. However, the results of a one-factor analysis of variance model suggested that there were no significant differences among the groups of students with regard to this question (F = 1.19; df = 6, 333; p = .312).

Interesting trends can be seen in the follow-up data from the Grade 4 questionnaire. Following are a few summary statements based on the results presented in Tables 33 to 37.

Table 37

Grade 4 Descriptive Statistics for "Would easier words help?"

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	1.78 (SD = .42; n = 55)	1.63 (SD = .52; n = 65)	1.70 (SD = .48; n = 120)
English dictionary	1.79 (SD = .41; n = 52)	1.70 (SD = .46; n = 69)	1.74 (SD = .44; n = 121)
Bilingual dictionary	1.70 (SD = .46; n = 56)	NA	1.70 (SD = .46; n = 56)
Linguistically modified items	1.85 (SD = .37; n = 20)	1.78 (SD = .42; n = 23)	1.81 (SD = .39; n = 43)
Column total	1.77 (SD = .43; n = 183)	1.68 (SD = .48; n = 157)	1.73 (SD = .45; n = 340)

Note. 1= No; 2 = Yes, definitely. All conditions included extra time.

- ELL students, more than non-ELL students, indicated that there were words in the science test that they did not understand (see Table 33).
- ELL and non-ELL students equally used dictionaries (see Table 34).
- ELL students, more than non-ELL students, believed that a dictionary helped them with the test (see Table 35).
- ELL students, more than non-ELL students, indicated that explanation in another language would help (see Table 36).
- Both ELL and non-ELL students strongly suggested that easier words would help them with the test (see Table 37).

Grade 8 Follow-Up Questions. The follow-up questionnaire for Grade 8 was similar to the Grade 4 questionnaire, with a few minor differences. Results of analyses of responses from Grade 8 students are consistent with the results that were reported for students in Grade 4. Tables 38 through 42 show student responses to the follow-up questionnaire for Grade 8. The small numbers of participants responding to these questions render statistical analysis impractical for this section of the study. Descriptive data are presented here primarily to show the trends and also to provide ideas for possible analysis with larger numbers of participants in a future study.

Table 38

Grade 8 Descriptive Statistics for "Were there words that you did not understand?"

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	2.25 (SD = .68; n = 16)	2.00 (n = 1)	2.24 (SD = 66; n = 17)
English dictionary	2.00 (SD = .66; n = 24)	1.83 (SD = .39; n = 12)	1.94 (SD = .58; n = 36)
Bilingual dictionary	2.31 (SD = .60; n = 16)	NA	2.31 (SD = .60; n = 16)
Linguistically modified items	2.17 (SD = .58; n = 12)		2.17 (SD = .58; n = 12)
Column total	2.16 (SD = .64; n = 68)	1.85 (SD = .38; n = 13)	2.11 (SD = .61; n = 81)

Note. 1= No; 2 = Yes, some; 3 = Yes, many. All conditions included extra time.

Table 39

Grade 8 Descriptive Statistics for "Did you use the dictionary?"

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
English dictionary	1.58 (*SD* = .65; *n* = 24)	1.58 (*SD* = .67; *n* = 12)	1.58 (*SD* = .65; *n* = 36)
Bilingual dictionary	1.63 (*SD* = .50; *n* = 16)	NA	1.63 (*SD* = .50; *n* = 16)
Column total	1.60 (*SD* = .59; *n* = 40)	1.58 (*SD* = .67; *n* = 12)	1.60 (*SD* = .60; *n* = 52)

Note. 1= No; 2 = Yes, some; 3 = Yes, a lot. Both conditions included extra time.

Table 40

Grade 8 Descriptive Statistics for "Did the dictionary help?"

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
English dictionary	1.61 (*SD* = .72; *n* = 23)	1.83 (*SD* = .83; *n* = 12)	1.69 (*SD* = .76; *n* = 35)
Bilingual dictionary	1.63 (*SD* = .50; *n* = 16)	NA	1.63 (*SD* = .50; *n* = 16)
Column total	1.63 (*SD* = .63; *n* = 39)	1.83 (*SD* = .83; *n* = 12)	1.67 (*SD* = .68; *n* = 51)

Note. 1= No; 2 = Yes, some; 3 = Yes, a lot. Both conditions included extra time.

Table 41

Grade 8 Descriptive Statistics for "Would explanation in another language help?"

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	2.43 (*SD* = .76; *n* = 14)	2.50 (*SD* = .71; *n* = 2)	2.24 (*SD* = 73; *n* = 16)
English dictionary	2.25 (*SD* = .61; *n* = 24)	1.42 (*SD* = .99; *n* = 12)	1.97 (*SD* = .77; *n* = 36)
Bilingual dictionary	2.14 (*SD* = .95; *n* = 14)	NA	2.14 (*SD* = .95; *n* = 14)
Linguistically modified items	2.08 (*SD* = .79; *n* = 12)		2.08 (*SD* = .79; *n* = 12)
Column total	2.23 (*SD* = .75; *n* = 64)	1.57 (*SD* = .85; *n* = 14)	2.12 (*SD* = .81; *n* = 78)

Note. 1= No; 2 = Maybe; 3 = Yes. All conditions included extra time.

Table 42

Grade 8 Descriptive Statistics for "Would easier words help?"

Accommodation group	ELL status		Row total (ELL + Non-ELL)
	ELL	Non-ELL	
Standard condition	1.88 (*SD* = .34; *n* = 16)	2.00 (*SD* = .00; *n* = 2)	1.89 (*SD* = 32; *n* = 18)
English dictionary	1.82 (*SD* = .39; *n* = 22)	1.83 (*SD* = .39; *n* = 12)	1.82 (*SD* = .39; *n* = 34)
Bilingual dictionary	1.75 (*SD* = .77; *n* = 16)	NA	1.75 (*SD* = .77; *n* = 16)
Linguistically modified items	1.83 (*SD* = .39; *n* = 12)		1.83 (*SD* = .39; *n* = 12)
Column total	1.82 (*SD* = .49; *n* = 66)	1.86 (*SD* = .36; *n* = 14)	1.83 (*SD* = .47; *n* = 80)

Note. 1= No, 2= Yes, definitely. All conditions included extra time.

Interesting trends can be seen in the follow-up data from the Grade 8 questionnaire. The following are a few summary statements based on the results presented in Tables 38 through 42.

- ELL students, more than non-ELL students, indicated that there were words in the science test that they did not understand (see Table 38).
- ELL and non-ELL students equally used dictionaries (see Table 39).
- Non-ELL students, more than ELL students, believed that a dictionary helped them with the test (see Table 40).
- ELL students, more than non-ELL students, indicated that explanation in another language would help (see Table 41).
- Both ELL and non-ELL students strongly suggested that easier words would help them with the test (see Table 42).

Discussion

Recent legislation mandated inclusion of all students—including English language learners—in large-scale national and state assessments. To assure a fair assessment system for everyone and to hold schools accountable for equal opportunity education for all, some legislation—such as the No Child Left Behind Act (2002)—requires achievement reporting by subgroups. The subgroup reporting includes ELL students. However, studies have shown a substantial test performance gap between ELL and non-ELL students. Results of studies on the assessment of

ELL students have clearly demonstrated that these performance gaps for ELL students are partly due to the impact of language factors on assessment. That is, ELL students perform poorly on content-based assessments mainly because they may not understand the language of test items, language that is unrelated to the content being assessed. To help ELL students overcome the problem of language in content-based assessments, some accommodation strategies have been suggested.

While provision of accommodation may help ELL students perform higher on a content-based assessment, it raises a new set of considerations. Among them are issues concerning effectiveness, validity, differential impact, and feasibility of accommodations used for ELL students. To address these major issues, we used four research questions to guide the analyses and reporting of our study, and these will be the basis for discussion of the results. The questions are:

1. Do accommodation strategies help reduce the performance gap between ELL and non-ELL students? (Effectiveness)

2. Do accommodation strategies impact the performance of non-ELL students on content-based assessment? (Validity)

3. Do student background variables impact performance on the accommodated assessments? (Differential impact)

4. Are accommodations easy to implement or use? (Feasibility)

In this study, we assessed both ELL and non-ELL students in science under four accommodation conditions: (a) a bilingual dictionary, (b) an English dictionary, (c) linguistic modification of items, and (d) a standard testing condition (NAEP) with no accommodations. We included multiple forms of accommodations to enable us to compare the effectiveness of different accommodation approaches. Students from different language and cultural backgrounds were included to examine any possible cross-cultural/cross-language factors that might impact the outcome of accommodated assessment.

In the U.S., the effectiveness of a language accommodation, to a great extent, depends on a student's English language background. We tested both ELL and non-ELL students because examining the validity of accommodated assessment would be impossible without observing the effects of accommodation on a general student population. We included a measure of English proficiency to be used as a covariate because we believe both ELL and non-ELL groups are not homogeneous within themselves. English reading ability is a desirable covariate in a study that compares

test performance with and without language accommodation. Students were tested from a wide array of schools and school districts and thus had English language proficiency designations based on different criteria or timetables. It was therefore necessary to compare science test ability with a current and comparable measure of reading ability before making observations about the effect of accommodations designed to aid reading and understanding. A reading test administered to all students in the study provided a more trustworthy covariate than ELL designation.

A student background questionnaire was developed to examine the impact of a student's background on an accommodated assessment. This information allowed us to test whether a student's background impacted performance on the science assessment and whether this impact differed under the various accommodation conditions.

During the early stages of the study, questions arose as to whether the students were effectively using the dictionary accommodation. As a result, an accommodation follow-up questionnaire was developed to measure students' self-reporting of the effectiveness of the various accommodated conditions. This information helped confirm and put into context the results of the accommodation analysis.

Findings

The results of this study showed that some of the accommodation strategies were effective in increasing the performance of ELL students and reducing the performance gap between ELL and non-ELL students. The results suggested that the effectiveness of accommodation may vary across grade level. Some forms of accommodation strategies were shown to be effective for Grade 4 students but not for Grade 8 students.

In brief, the English dictionary was among the effective accommodations for ELL students in Grade 4. For Grade 8 ELL students, however, linguistic modification of test items seemed to be more effective. These results seem reasonable since the content assessment for students in the higher grade was more linguistically complex.

The performance of non-ELL students did not show any significant improvement under any of the accommodations used for either grade in this study. This finding is encouraging since it suggests that the validity of the assessments was not compromised by the use of accommodation.

The impact of background variables on the results was studied. In Grade 8, there was an inadequate number of participants to allow enough statistical power for tests of significance on the background variables and the follow-up questions. In Grade 4, many background variables were significantly related to performance on the science assessment. These variables included time in the U.S., initial grade attended in the U.S., having attended school outside the U.S., primary home language of Korean or Spanish, and ability to understand spoken English at school. As fourth-grade ELL students have spent a large part of their lives functioning in a language other than English, the significance of these variables is not surprising.

In Grade 4, student background variables had less impact on science performance for students who received an English dictionary than for students under the standard condition. The student background variables that had the greatest impact on science performance for students under the standard condition were ELL status, time in the U.S., and Korean home language. There were no student background variables that had a significant impact on science performance for students receiving the English dictionary accommodation.

In the accommodation follow-up questionnaire, there were no significant differences in the amount of dictionary use reported or in the rating of the helpfulness of dictionaries by ELL and non-ELL students. Not surprisingly, ELL students more often than non-ELL students stated that explanation in another language would benefit them. What was interesting was that, among ELL students, those who received English dictionaries or linguistically modified test versions rated the helpfulness of explanation in another language lower than did ELL students testing under the standard condition. There were no significant differences in how students rated the helpfulness of "easier words on the test."

We will now discuss the results of this study in response to the four research questions. With respect to Question 1, on the effectiveness of accommodations, different forms of accommodation showed different levels of effectiveness across grade levels. For Grade 4 students, the English dictionary was an effective form of accommodation in terms of increasing the performance of ELL students. In Grade 8, however, linguistic modification of test items was the only form of accommodation that helped ELL students improve their performance.

Question 2 concerns the validity of accommodations used in this study. The results indicated that none of the accommodation strategies used in this study

improved the performance of non-ELL students. This provides assurance on the validity of accommodations that were used in this study. A lack of impact on the performance of non-ELL students suggests that the accommodation used did not change the construct being measured.

Question 3 raises the possibility of differential impact of accommodation due to the interaction of student background with the assessments. The results indicated that variables such as the length of time in the U.S. impacted the assessment results. This finding is consistent with our previous studies indicating the impact of language background on students' performance (see, for example, Abedi & Lord 2001; Abedi et al., 2000).

The last research question, Question 4, concerns the feasibility and practicality of the accommodations. If the most effective and valid accommodations are not applicable in large-scale assessment, then those accommodations may not have much practical value. Accommodations such as one-on-one testing, extension of testing time and/or reading the test items or test directions aloud may not be feasible in large-scale assessments. As indicated earlier in this report, we tried to use more practical accommodations that do not require additional efforts by teachers, school personnel, or other test administrators.

The two accommodations used that were shown to be effective and valid were the English dictionary (for Grade 4) and the linguistic modification of test items (for Grade 8). While both are language accommodations, they differ in other respects. The linguistic modification of test items was an easy accommodation to administer, as its implementation occurred at the test instrument development stage. We first selected test items and then modified them by removing any unnecessary linguistic complexity. Linguistic modification of test items did not pose any difficulty in the test administration stage since it did not require additional administration time or special administration procedures.

The use of the English dictionary posed some challenges. First, it took a substantial amount of time for the research team to select dictionary editions appropriate for the age groups in this study (Grade 4 and Grade 8). Second, providing dictionaries at testing sites required effort that was not justified by the observed lack of use of that accommodation. Third, dictionary use itself is a skill that ELL students and their classmates may not yet have acquired. Therefore, the dictionary accommodation did not meet our feasibility criterion.

Challenges

The study was designed with the intention to test Grade 4 and Grade 8 ELL students and their classmates in an approximately 50-50 split between ELL and non-ELL students. However, this design did not take into account the kind of dilution of language groups that takes place in junior high and middle school. An elementary school might contain a large number of students from the same language group, but a middle school may draw students from many language groups in the area, creating a science or ESL class that contains only a few students from the language groups being studied. This was especially true when searching for classes containing significant numbers of Chinese, Korean, or Filipino ELL students.

Another difficulty in obtaining Grade 8 ELL participants is the likelihood that by Grade 8, a student has been in the U.S. for enough years to have been re-designated an English-proficient student.

Further comments on each type of dictionary are warranted. Bilingual dictionaries are designed for English speakers who are learning another language (usually for conversation purposes) and contain less academic language than standard English language dictionaries. The English language dictionary used in our study contained more content terms and more "academic" language than the bilingual dictionaries.

Implications for Policy, Practice, and Research

In considering these results, no matter how feasible it may seem to provide ELLs students with a published bilingual dictionary, since it didn't reduce the performance gap between ELL and non-ELL students, it may not have a practical use as a language accommodation for testing the populations in this study. After early pilot testing revealed that students could not always find unknown test words in the bilingual dictionaries, we examined the number of words in the tests that were actually in each dictionary. The appearance of the tests' non-content words varies by dictionary and is discussed in Appendix B and listed in Table B4. In the same dictionaries, science content words were also available. The number of content words ranged from a maximum of 70 (English dictionary) to a minimum of 25 (Ilocano dictionary; see Table B1). The difference in ratios of content words to non-content words is also a consideration when selecting a pre-published accommodation tool.

The English dictionary seemed effective for fourth graders, but if its use is unfamiliar and/or its size unwieldy, it may not be very useful either, especially in large-scale assessments. The accommodations that require the student to look something up (and possibly not find it) might be utilized less than those that are more pre-packaged, such as the linguistic modification of test items.

Fortunately, none of the accommodations seemed to affect the construct of the science test. So, with regard to accommodation validity, even if districts or schools fall behind in re-designating ELL students from "LEP" to "RFEP," the language accommodations used in this study would not significantly improve the scores of non-ELL students in a content assessment.

An important feature of this research was the assessment of each student's English language reading ability. Testing both ELL and non-ELL students with a single instrument posed a challenge because reading assessments such as the National Assessment of Education Progress are designed for the general population and do not discriminate well among ELLs. To provide a measure of reading comprehension for students in future studies, instruments for ELL students and those fluent in English could be combined in a three-part test. The combination test could include an assessment of word recognition, a section of the Language Assessment Scales (LAS; 1990) test that seems to discriminate among all types of English language learners, and one block of the NAEP reading assessment (used in the present study).

As the best accommodation strategies may differ by grade level, by home language/culture, and by other background variables, promising language accommodation strategies—such as reducing the language complexity of the content tests themselves—merit further research. For example, further research could examine the performance of ELL students and their classmates on those science questions with a higher language load.

References

Abedi, J. (1996). The interrater/test reliability system (ITRS). *Multivariate Behavioral Research, 31,* 409-417.

Abedi, J., Courtney, M., & Leon, S. (2001). *Language accommodation for large-scale assessment in science.* Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data*. 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. (2001). *Students' performance differences in standardized achievement tests and background factors: Analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

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

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation on assessment of English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

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

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bailey, A. (2000). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *Assessing English language learners* (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., & Castellon-Wellington, M. (2000). *Students' concurrent performance on tests of English language proficiency and academic achievement* (Final Deliverable to

OERI/OBEMLA, Contract No. R305B60002, pp. 51-83). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

California Department of Education. (2000). *English learners in California*. Retrieved August 25, 2000, from http://www.cde.ca.gov/ccpdiv/Eng_Learn/CCR2000-EL/index.htm

Durán, R. P. (1989). Assessment and instruction of at-risk Hispanic students. *Exceptional Children, 56*, 154-158.

Gandara, P., & Merino, B. (1993). Measuring the outcomes of LEP programs: Test scores, exit rates, and other mythological data. *Educational Evaluation and Policy Analysis, 15,* 320-328.

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

Goals 2000: Educate America Act, Pub. L. No. 103-227, 108 Stat. 125 (1994).

Goldstein, A. A. (1997, March). *Design for increasing participation of students with disabilities and limited English proficient students in the National Assessment of Educational Progress.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Imbens-Bailey, A., & Castellon-Wellington, M. (1999, September). *Linguistic demands of test items used to assess ELL students*. Paper presented at the annual conference of the National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Improving America's Schools Act of 1994. Pub. L. No. 103-382. 108 Stat. 3518 (1994).

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services: 2000-2001 Summary report*. Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

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*(1), 55-75.

Language Assessment Scales. (1990). Monterey, CA: CTB/McGrawHill

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.

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, National Center for Education Statistics.

National Assessment Governing Board, NAEP Science Consensus Project. (n.d.). *Science framework for the 1996 and 2000 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board. Retrieved 25 October 2005 from http://www.nagb.org/pubs/96-2000science/toc.html

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

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

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

North Central Regional Educational Laboratory. (1996b). *The status of state student assessment programs in the United States: Annual report.* Oakbrook, IL: North Central Regional Educational Laboratory and Council of Chief State School Officers.

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

O'Sullivan, C. Y., Reese, C. M., & Mazzeo, J. (1997, May). *NAEP 1996 science report card for the nation and the states* (NCES 97-497). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Rivera, C., & Stansfield, C. W. (1998). Leveling the playing field for English language learners: Increasing participation in state and local assessments through accommodations. Retrieved August 25, 2000, from http://ceee.gwu.edu/standards_assessments/researchLEP_accommodcase.htm

Rivera, C., & Stansfield, C. W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998-1999.* Arlington, VA: The George Washington University, Center for Equity and Excellence in Education.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). *Statewide assessment program policies and practices for the inclusion of limited English*

proficient students (EDO-TM-97-02). Washington, DC: ERIC Clearinghouse on Assessment and Evaluation.

Saville-Troike, M. (1991). Teaching and testing for academic achievement: The role of language development. *NCBE Focus: Occasional Papers in Bilingual Education, 4*. Retrieved August 25 2000, from http://www.ncbe.gwu.edu/ncbepubs/focus/focus4.html

Shuard, H., & Rothery, A., (Eds.). (1984). *Children reading mathematics.* London: J. Murray.

Texas Education Agency. (n.d.). *Administrative Rules.* Retrieved August 25, 2000, from http://www.tea.state.tx.us/

Appendix A

Table A1

Sampling of Participants and School Sites

Date		GR	State	Language	Class	ELL in language	Total ELL	Total students
2/28-29	01	4	California 1	Spanish	01401	5	14	32
4/4-5	01	4	California 1	Spanish	01402	2	9	31
4/4-5	01	4	California 1	Spanish	01403	3	10	32
3/7-8	02	4	California 1	Spanish	02401	5	5	21
3/7-8	02	8	California 1		02801		0	29
3/7-8	02	8	California 1	Spanish	02802	1	1	29
3/22-23	05	8	California 1		05801		0	34
3/22-33	05	8	California 1	Spanish	05802	29	31	31
5/16-17	07	4	California 1	Korean	07401	10	10	30
5/16-17	07	4	California 1	Korean	07402	16	18	29
5/18-19	08	4	California 2	Chinese	08401	12	16	30
5/18-19	08	4	California 2	Chinese	08402	28	29	31
5/18-19	09	8	California 2	Chinese	09801	10	24	22
5/23-24	10	4	Texas	Spanish	10401	27	27	27
5/23-24	10	4	Texas	Spanish	10402	27	27	27
5/25-26	12	4	Texas		12401		0	17
5/25-26	12	4	Texas	Spanish	12402	22	24	24
5/30-31	13	8	California 1	Korean	13801	27	31	29
5/30-31	13	8	California 1	Korean	13802	0	16	16
6/5-6	14	4	Hawaii	Ilocano	14401	8	2	27
6/5-6	14	4	Hawaii	Ilocano	14402		5	27
6/5-6	15	4	Hawaii	Ilocano	15401		5	24
6/5-6	15	4	Hawaii	Ilocano	15402	9	12	12
					15403			
					15404			
					15405			
Total						241	316	611

Note. Twelve Filipino students from 4 different classes were brought into one room for testing. Nine were Ilocano speakers.

Appendix B

Dictionary Contents

After early pilot testing revealed that students could not always find unknown test words in the bilingual dictionaries, we examined the number of words in the tests that were actually in each dictionary (see Table B1 and B2). Table B3 and B4 illustrate the content differences between the English dictionary and the bilingual dictionaries used.

Table B1

Grades 4 and 8 Number of Content Words in Science Tests Found in Each Dictionary Provided as an Accommodation

Chinese	Ilocano	Korean	Spanish	English	Total
55	25	41	34	70	71

Table B2

Grades 4 and 8 Number of Non-Content Words in Science Tests Found in Each Dictionary Provided as an Accommodation

Chinese	Ilocano	Korean	Spanish	English	Total
67	52	67	53	71	71

The difference in ratios of content words to non-content words is also a consideration when selecting an accommodation tool.

In addition, the significant difference between use of an English dictionary and a bilingual dictionary as an accommodation should be noted. A bilingual dictionary usually offers one or a few words as a simple translation of the unknown item. For example:

experiment *n* : experimento *v* : experimentar

water *n* : agua

The definition provided by a non-compact English dictionary often offers more than a synonym of the unknown item:

water *n* : the liquid that descends from the clouds as rain, forms streams, lakes, and seas, and is a major part of *all living material* and that is an odorless and tasteless compound having two atoms of hydrogen and one atom of oxygen per molecule.

fuel *n* : a material from which atomic energy can be produced especially *in a reactor*; a source of energy.

Notice how the definitions of *water* and *fuel* directly lead a student to possibly correctly answer the following NAEP science questions:

Which of the following is found in every living cell?

1. alcohol
2. cellulose
3. chlorophyll
4. hemoglobin
5. water

At the present time, where does most of the energy used in this country come from?

6. nuclear reactors
7. hot springs
8. solar batteries
9. burning of fuels
10. don't know

Table B3

Content Words in Science Tests and Their Appearance in Accommodation Tools

Word in test	Location in Test	Grade of test	Chinese dictionary	Ilocano dictionary	Korean dictionary	Spanish dictionary	English dictionary
Alcohol	Answer	8A	Y	Y	Y	Y	Y
Aspen	Question	8A	N	N	N	N	Y
Atmosphere	Answer	4,8A,B	Y	N	Y	Y	Y
Atom	Question	8B	Y	N	Y	N	Y
Caterpillar	Question	4	Y	Y	N	Y	Y
Cell	Question	8A,B	Y	N	Y	Y	Y
Cellulose	Answer	8A	N	N	N	Y	Y
Celsius	Question	8A,B	Y	N	N	N	Y
Centigrade	Question	8B	Y	N	Y	N	Y
Chlorophyll	Answer	8A	Y	N	N	N	Y
Cross-section	Question	8A,B	Y	N	N	N	Y
Crust	Q/A	8A,B	Y	Y	Y	Y	Y
Dial	Question	8A	Y	N	Y	Y	Y
Digestive	Question	8A,B	Y Digest	Y Digest	Y	Y Digest	Y
Earth	Question	4	Y	Y	Y	Y	Y
Electrical	Question	8A,B	Y Electric	Y Electric	Y	Y Electric	Y Electric
Energy	Question	4	Y	N	Y	Y	Y
Energy	Answer	8B	Y	N	Y	Y	Y
Enzyme	Question	8B	Y	N	N	N	Y
Evidence	Question	4	Y	Y	Y	Y	Y
Frequency	Answer	8A,B	Y	Y Frequent	Y	Y	Y
Fuel	Answer	8B	Y	Y	Y	Y	Y Answer
Gauge	Question	8A	Y	N	Y	Y	Y
Geology	Question	8A,B	Y	Y Geologist	Y	N	Y
Grasshopper	Question	4	Y	Y	Y	N	Y

Table B3 (continued)

Word in test	Location in Test	Grade of test	Chinese dictionary	Ilocano dictionary	Korean dictionary	Spanish dictionary	English dictionary
Gravitational	Answer	8A,B	Y	N	Y	Y Gravity	Y Answer
Half-life	Question	8A	N	N	N	N	Y Answer
Hemoglobin	Answer	8A	N	N	N	N	Y
Human	Question	4	Y	Y	Y	Y	Y
Insect	Question	4	Y	Y	Y	Y	Y
Larva	Question	4	Y	Y	N	N	Y Answer
Lunar Eclipse	Question	8A,B	Y Lunar	N	Y Lunar	N	Y Answer
Magnetic pole	Answer	8A,B	P Magnetic	N Magnet	P Magnetic	P Magnetic	Y
Magnifying	Question	8B	Y Magnify	Y Magnify	Y Magnify	Y	Y
Mealworm	Question	4	N	N	N	N	Y
Microscope	Question	8B	Y	N	Y	Y	Y
Mitochondrion	Question	8B	N	N	N	N	Y Answer
Mucus	Question	8B	Y	Y	N	N	Y Answer
Muscles	Question	8A	Y Muscle	Y Muscle	Y Muscle	Y Muscle	Y Muscle
Nebula	Question	8A,B	N	N	N	N	Y
Newborn	Question	4	N	N	N	Y	Y
North Star	Answer	8B	P North	P North	P North	P North	Y
Nuclear	Answer	8A,B	Y	N	Y	N	Y
Nuclear	Answer	4	N	N	N	N	Y Nuclear
Nuclear reactors	Answer	8B	Y	N	N	N	Y
Nutrient	Question	8A,B	Y	N	N Nutrition	N	Y
Organ	Question	8A	Y	Y	Y	Y	Y
Organ	Answer	8B	Y	Y	Y	Y	Y
Organism	Answer	8B	Y	N	Y	N	Y
Oxygen	Answer	8A,B	Y	N	Y	N	Y
Oxygen	Question	4	Y	N	Y	N	Y
Pepsin	Question	8B	N	N	N	N	Y

Table B3 (continued)

Word in test	Location in Test	Grade of test	Chinese dictionary	Ilocano dictionary	Korean dictionary	Spanish dictionary	English dictionary
Power plant			P Power, Power-station	N Power	Y	N Power	Y
Protein	Question	8B	Y	N	N	N	Y
Pupa	Question	4	N	N	N	N	Y
Reactors	Question	8B	Y	N React	N React	N React	N React
Reading		8	Y	Y Read	Y	Y Read	Y
Reproduce	Answer	4	Y	Y	Y	N	Y
Salamander	Question	4	N	Y	N	N	Y
Solar system	Question	8A,B	Y	N	P Solar	N	Y Advantage
Species	Answer	4	Y	Y	Y	Y	Y
Tadpole	Story	4	Y	N	N	N	Y Answer
Tectonic plate	Answer	8A	N	N	N	N	Y Advantage?
Theory	Question	8A,B	Y	N	Y	Y	Y
Thermometers	Question	8A	Y	Y	Y	Y	Y
Tissue	Question	8A	Y	N	N	Y	Y
Tissue	Answer	8B	Y	N	N	Y	Y
Trait	Question	8A,B	Y	N	Y	N	Y
Vitamins	Question	8	Y Vitamin	N	Y	Y	Y
Volcano	Answer	4	Y	N	Y	Y	Y
Volume	Question	8A	Y	Y	Y	Y	Y

Note. Y = Yes, appears in dictionary. N = No, does not appear. P = Part of an expression appears.

Table B4

Non-Content Words in Science Tests and Their Appearance in Accommodation Tools

Word in test	Location in test	Grade of test	Chinese dictionary	Ilocano dictionary	Korean dictionary	Spanish dictionary	English dictionary
Accurately	Question	8A,B	Y	Y	Y	Y	Y
Adult	Question	4	Y	Y	Y	Y	Y
Arthritis	Question	8A,B	Y	N	N	N	Y
Article	Directions	4	Y	Y	Y	Y	Y
Batteries	Answer	4	Y	N	N	Y	Y
Blond	Question	4	Y	Y	Y	Y	Y
Boundary	Question	8A,B	Y	Y	Y	Y	Y Advantage
Broadcasting	Answer	4	Y	Y	Y	Y	Y
Cavities	Question	8A	N	Y	Y	N	Y
Clump	Story	4	Y	N	Y	N	Y
Coal	Answer	4	Y	Y	Y	Y	Y
Consistent	Answer	8A,B	Y	N	Y	Y	Y
Continent	Question	8A	Y	Y	Y	Y	Y
Cycle	Question	4	Y	N	Y	N	Y
Data	Question	4	Y	N	Y	Y	Y
Deposit	Answer	4	Y	Y	Y	Y	Y
Diagram	Question	8A,B	Y	N	Y	Y	Y
Different	Question	4	Y	Y	Y	N Differ	Y
Dock	Question	4	Y	N	Y	Y	Y
Dune	Answer	4	Y	N	Y	N	Y
Equation	Question	8A	Y	N	N	N Equate	Y
Explanation	Question	4	Y Explain	Y	Y	Y	Y
Factories	Answer	4	Y	Y	Y	Y	Y
Gasoline	Question	4	Y	Y Gas	Y	Y	Y Advantage
Glowing	Answer	4	Y	Y Glow	Y	Y	Y

Table B4 (continued)

Word in test	Location in test	Grade of test	Chinese dictionary	Ilocano dictionary	Korean dictionary	Spanish dictionary	English dictionary
Graph	Question	4	Y	N	Y	Y	Y
Hatch	Story	4	Y	Y	Y	Y	Y
Headaches	Question	8A,B	P Head	P Head	Y Headache	Y Headache	Y Headache
Hillside	Question	8B	Y Hill	Y	Y	Y Hill	Y
Ice Cap	Answer	4	P Ice	P Ice	Y	P Ice	Y
Indigestible	Answer	8B	Y Indigestion	Y Indigestion	Y Indigestion	Y Indigestion	Y
Inherit	Question	8A,B	Y	Y	Y	Y	Y Advantage
Insulation	Answer	8A,B	Y	N	Y Insulate	N	Y Answer
Interaction	Question	8A	Y	N	Y	N	Y
Joint	Diagram	8A,B	Y	Y	Y	Y	Y
Labeled	Question	4	Y Label	Y Label	Y Label	Y Label	Y Label
Lining	Question	8B	Y	Y	Y	N (clothes; brakes)?	Y Advantage
Live	Question	4	Y	Y	Y	Y	Y
Machine	Question	4	Y	Y	Y	Y	Y
Mirror	Question	8A,B	Y	N	Y	Y	Y Answer
Moon	Question	4	Y	Y	Y	Y	Y Answer
Multiplied	Question	8A	Y Multiply	Y Multiply	Y Multiply	Y Multiply	Y Multiply
Oil	Answer	4	Y	Y	Y	Y	Y Answer
Oval	Direction	4,8A,B	Y	Y	Y	N	Y
Plot	Question	8A,B	Y	Y	Y	Y	Y
Portable	Question	4	Y	Y	Y	Y	Y Answer
Predict	Question	8A,B	Y	Y	Y	Y	Y
Procedures	Question	8A	Y Procedure	Y Procedure	Y Procedure	N Proceed	Y Procedure
Process	Question	8A,B	Y	Y	Y	Y	Y
Prompt	Question	8A,B	Y	Y	Y	Y	Y Answer

Table B4 (continued)

Word in test	Location in test	Grade of test	Chinese dictionary	Ilocano dictionary	Korean dictionary	Spanish dictionary	English dictionary
Property	Question	8A,B	Y	Y	Y	Y	Y
Represent	Question	4	Y	Y	Y	Y	Y
Resources	Question	8A,B	Y	Y	Y	N	Y
Rheumatism	Question	8A,B	Y	Y	Y	Y	Y
Ripple	Question	4	Y	Y	Y	Y	Y
Same	Question	4	Y	Y	Y	Y	Y
Sand	Answer	8A	Y	Y	Y	Y	Y Answer
Sediment	Answer	8A,B	Y	Y	Y	N	Y
Similar	Question	4	Y	Y	Y	Y	Y
Smog	Question	4	Y	N	Y	N	Y Answer
Solar	Answer	8B	Y	N	Y	N	Y
Strike	Diagram	8A	Y	Y	Y	Y	Y
Source	Question	4	Y	Y	Y	Y	Y
Storm window	Answer	8A,B	P Storm	P Storm	P Storm	N	Y Advantage
Substance	Answer	8B	Y	Y	Y	Y	Y
System	Answer	8B	Y	Y	Y	Y	Y Answer
Toss	Question	4	Y	Y	Y	Y	Y
Transport	Answer	8B	Y	Y	Y	Y	Y
Types	Question	4	Y Type	Y Type	Y Type	Y Type	Y Type
Variable	Question	8A,B	Y	Y	Y	Y	Y
Window	Question	8A,B	Y	Y	Y	Y	Y

Note. Y = Yes, appears in dictionary. N = No, does not appear. P = Part of an expression appears.

Appendix C

Methodology Appendix

English and Bilingual Dictionaries

Butterfield, A. S. (1993). *Spanish-English, English-Spanish dictionary.* New York: Hippocrene Books.

Concise English-Chinese, Chinese-English dictionary. (1999). Hong Kong: Oxford University Press (China).

Korean-English, English-Korean dictionary. (1992). New York : Hippocrene Books.

Merriam Webster's Intermediate Dictionary. (1986). Springfield, MA: Merriam-Webster, Inc.

Rubino, R. G. (1998). *Ilocano: Ilocano-English/English/Ilocano dictionary and phrasebook*. New York: Hippocrene Books.

Spanish and English Student dictionary. (1999). Chicago: NTC Publishing Group.

Appendix D
Linguistic Modification Concerns

VOCABULARY (LEXICON)

- false cognates
- unfamiliar words (idioms, phrasal verbs, infrequently used words, words containing cultural assumptions, words containing unfamiliar contexts)
- overuse of synonyms
- long words
- phrases specific to the content area
- word sound

GRAMMAR (SYNTAX)

- long phrases in questions (no question word at the beginning)
- compound sentences (coordinating conjunctions, conjunctive adverbs)
- complex sentences (subordinating clauses)
- logical connectors (conditional clauses)
- unfamiliar tenses (conditional verbs, modals)
- long noun phrases
- relative clauses
- unclear or missing antecedents of pronouns
- negation, especially negative questions, negative terms, grammatical double negatives
- comparative construction and added complications
- prepositional phrases, especially when separating subject and verb
- verb phrases
- misplaced adjective phrases

STYLE OF DISCOURSE

- long problem statements; unnecessary expository material
- abstract (vs. concrete) presentation of problem
- passive voice
- complex arrangement of parts of speech
- paragraphs not unified in style (multiple changes in style of discourse, missing transitions)

CONCERNS SPECIFIC TO SCIENCE PROBLEMS

- phrasing that confuses the sequence of events
- words with both technical and non-technical meanings
- science keywords misinterpreted
- derivatives of content words

Appendix E

Adaptations to the Procedures

Some observations during early testing sessions spurred immediate changes to the test administration procedures or to the accommodations themselves.

1. Initially, the test administration instructions asked the students to refrain from asking questions. When this admonition was removed, the students' questions began to inform the design of the study.
2. When early testing showed little use of the dictionaries, a brief questionnaire was added in order to find out whether the students found the words difficult and whether the dictionary was useful. The test administration script was amended to include a reminder that the dictionaries could be used by anyone who had received one.
3. When it seemed that not many Hispanic students were using the dictionaries, an additional, more passive, accommodation was added to the study, the linguistically modified test version.
4. When we realized that the time in a class period barely allowed for an introduction, students writing their names, test directions and the test and questionnaire, the test administration scripts were streamlined and the booklet distribution pre-planned to save class time. As long as class rosters were submitted to CRESST ahead of time, each test booklet contained the student name and accommodation type. The instructor helped distribute the pre-assigned booklets. Then test administrators passed out the English and bilingual dictionaries. With the in-class set-up time minimized, needed sample questions were inserted between the introduction and the directions.
5. To discourage cheating, as well as to control for order effects, the reading test was easily made into two different "versions" by switching the order of the two passages and their questions.
6. The student background questionnaires were rewritten into unfinished statement form to rid items of inverted word order as well as save time.
7. Since Ilocano speakers from the Philippines outnumbered Tagalog speakers, Ilocano dictionaries were used as an accommodation when testing Filipino students in Grade 4.
8. When the contents of the selected dictionaries were compared to the science test lexicons, a more comprehensive Chinese dictionary replaced the first one selected.
9. Florida data collection in Jacksonville was because of the limited number of Spanish speakers among the ELLs. When Dade County was contacted for access to their more appropriate population, the students there were deemed too "overtested" by that time of the year.