Effectiveness and Validity of Accommodations for English Language Learners in Large-Scale Assessments

CSE Report 608

Jamal Abedi, Mary Courtney, and Seth Leon

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

September 2003

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 4.2 Validity of Assessment and Accommodations for English Language Learners
Jamal Abedi, Project Director, CRESST/UCLA

The work reported herein was supported 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.

Acknowledgments

This study required the participation of 40 schools, their administrators, faculty, and staff. Their hospitality and patience is much appreciated, and their dedication to educating young people is greatly admired. We cannot thank the students enough for their diligence in taking the tests.

Many people generously contributed to the development of this study. We are especially indebted to Joan Herman for her insightful comments and advice, and to Eva Baker for her continuous support of our research.

Richard Durán participated in this study as an advisory member of the research team. His comments on the pilot report improved the design of the main study, and his close reading improved this report's content. Ann Mastergeorge generously shared her thorough teacher survey on accommodation use.

Frances Butler of UCLA/CRESST and Alison Bailey of UCLA provided excellent suggestions on the pilot study report. Their insight helped shape the design of the main study. Each read the draft of this report with an eye on how to better communicate what we learned and thought. We are also grateful to professors Richard Durán of UCSB, Carolyn Hofstetter of UC Berkeley, and Carol Lord of CSULB for their review of this report and for their valuable suggestions, feedback and advice.

Jennifer Vincent, project assistant, provided essential logistical support to every aspect of the study. Her creation of the automated data input forms was skillful and her data record-keeping meticulous.

We appreciate Jim Mirocha's assistance in data analysis. Jennifer Goldberg led the scoring of the study's open-ended test items at hours most convenient to the hard-working raters. We are also indebted to the many raters who set aside so much time to sit and read student work.

Student clerks Yvette Cuenco and Amy Gallatin are remembered for their patience with the revision process. Graduate student researchers Melissa Cantu, Jennifer Goldberg, and Ani Shabazian assisted in the literature review.

Many people helped us land on our feet in unfamiliar school districts, especially Leon Belcher, Carolyn Hofstetter, Elham Kazemi, Martin O'Callaghan,

Henry Solomon, and Harriet Sturgeon. Michael O'Connell and Patrick Lee took some of the data provision load off of the school administrators in their districts.

We also thank Jenny Kao for her generous contributions to the preparation of this manuscript, including revisions and updates to the introductory sections.

We appreciate the dedication, travel, hard work, and observations of our test administrators:

Site 1	**Site 2**	**Site 3**	**Site 4**
Mary H. Cong	Irene Lupe Araujo	Erin Adams	Lori Colliander
Liz Galvin	Gabino Arredondo	Antonio Cerda	Kathryn Donald
Tina Henderson	Alejandra Flores	Andrea Higgins-Simon	Takashi Furuhata
John Iwanaga	Lana Hau	Christopher Jones	Patricia Guevara
Hyun Uk Kim	Maria Leung	Caroline Kalil	Raquel Guevara
Lillian Neville	Heather MacLeod	Valisha LaNear	Nicholas Leininger
Louise Nixon	Van Nguyen	Esmeralda Lopez	Amina Rahman
Tammy Shel	Alejandro Pena	Dan P. Morgan	Thomas Youn
Tara Watford	Jorge Solis	Mario Nevarez	
	Michael Sova	Alma Grant Swisher	—and:
	Mieka Valdez		Isimemen Johnson
	Kelly Ken Li Wong		Gisela Torres

A Special Acknowledgement

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

Julia Lara, Council of Chief State School Officers (CCSSO)

Carol Lord, California State University, Long Beach

John Mazzeo, Educational Testing Service

John Olson, Council of Chief State School Officers (CCSSO)

Charlene Rivera, George Washington University

Lorrie Shepard, University of Colorado, Boulder

Catherine Snow, Harvard University

Charles Stansfield, Second Language Testing, Inc.

EFFECTIVENESS AND VALIDITY OF ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNERS IN LARGE-SCALE ASSESSMENTS

Jamal Abedi, Mary Courtney, and Seth Leon
CRESST/University of California, Los Angeles

EXECUTIVE SUMMARY

As the population of English language learners (ELLs) in U.S. public schools continues to grow, issues concerning their instruction and assessment are steadily among the top national priorities in education. Recent legislation mandates inclusion of *all* students in a school's assessment and accountability system (see the Improving America's Schools Act of 1994 and the No Child Left Behind Act of 2001). However, research on and practice in the instruction and assessment of ELL students has raised a new set of issues and concerns. Among the most important are that language factors may confound test results in content areas such as math and science where language should not play a role (Abedi & Lord, 2001; Bailey, 2000; Durán, 1989; Garcia, 1991; Mestre, 1988).

To reduce the impact of language factors and to make a fair assessment of the content knowledge of ELL students, some educational researchers and practitioners recommend the use of accommodation (Mazzeo, Carlson, Voelkl, & Lutkus, 2000; O'Sullivan, Reese, & Mazzeo, 1997; Rivera & Stansfield, 1998). Accommodation for ELL students aims to help "level the playing field" with regard to English language comprehension. However, recent research— including studies by researchers at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST)—has demonstrated that accommodation may sometimes do more than originally intended. An accommodation may alter the construct under measurement, giving an unfair advantage to those receiving the accommodation, and thereby negatively impacting the validity of assessment (see especially Abedi, Lord, Hofstetter, & Baker, 2000; Olson & Goldstein, 1997).

Even if an accommodation is found to be both effective and valid, it may not be feasible to implement. Some forms of accommodation may cause an additional burden, either financially or logistically, 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: effectiveness, validity, differential impact, and feasibility:

1. Do accommodation strategies help to reduce the performance gap between ELL and non-ELL students by removing language barriers? (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 the accommodations easy to prepare and use? (Feasibility)

Methodology

A total of 1,854 Grade 4 students and 1,594 Grade 8 students in 132 classes at 40 school sites participated in this study.[1] The English proficiency designation of students was determined based on school records. Out of 3,448 students, 1,712 (49.7%) students were identified as being ELL.[2] The other 1,736 students were designated either as speaking English as a home language or as having become proficient enough in English to be re-designated. For the main study, these latter students were combined into our non-ELL category. Three language categories were targeted for the main study: Spanish, Chinese, and "other Asian languages."[3] Of the 1,712 ELL students, 1,614 (94.3%) students belonged to one of the three target language categories.

The focus of the research was to study the impact of accommodation on students' performance in science. We selected science because it is a content area in which language ability may interfere with the measurement of science achievement. To measure science knowledge, a science test of multiple-choice and open-ended questions was administered in four forms. One form contained original NAEP (National Assessment of Educational Progress) science items and a few TIMSS (Third International Mathematics and Science Study) multiple-choice items with no language accommodation.[4] The remaining three forms each included a language accommodation (either a Customized English Dictionary, a Bilingual/English Glossary, or a Linguistic Modification[5] of the test) that addressed the challenge of understanding the English lexicon and, possibly, its syntax. These three forms also allowed additional time (50%). In addition to the science test, an English language reading battery was also administered to determine students' reading proficiency levels.

---

[1]A small number of students were excluded from the study because they were completely non-English speaking, were enrolled in a different grade, or were administered an inappropriate accommodation. The number of participating students may not always match the number of students included in the analysis (by a small margin).

[2]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.

[3]The more than 200 Asian students combined in this category spoke languages from East Asia, Southeast Asia, the Philippines, and the Pacific Islands. Notable among these are, in order of frequency: Korean, Vietnamese, Tagalog, Mien, Khmer, Lao, Ilocano and Samoan.

[4]The only difference between the standard condition in NAEP and in our study is that we gave additional time (50%).

[5]The linguistic complexity of the science items—but not the science content—was simplified.

The reading measure was an essential part of the accommodation study because students at different levels of reading proficiency may benefit differently from any accommodation received in this study, regardless of their ELL status. The English reading efficiency / proficiency battery consisted of the Fluency section of the Language Assessment Scales (LAS), one intact block of the 1994 NAEP Reading assessment (released items), and an experimental word identification section. This compilation of language assessment tools was devised to measure the reading ability of both ELL and non-ELL students. The LAS section asked students to choose the word missing from a sentence. The word identification section asked students to identify English words that were listed among nonsense words that mimicked English word structure. The NAEP block contained open-ended and multiple-choice items based on a reading passage.

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 a student's background affected his / her performance on the tests. The questionnaire queried, among other things, language background, country of origin, and length of time in the United States, and also asked students to self-assess their proficiency both in English and in their home language. The accommodation follow-up questionnaire asked students, when applicable, whether the dictionary / glossary helped them during the science test and, for all students, how the language in the test could have been made easier to understand. The teacher questionnaire included questions regarding educational background and experiences. The school questionnaire contained questions about the school population and the science and English as a Second Language (ESL) resources.

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

Most open-ended test items were scored by at least two raters who were trained by the project staff. The NAEP guidelines and scoring rubrics were followed for double-scoring the open-ended science and language items.

Results

Null hypotheses. The null hypotheses related to the research questions mentioned earlier are:

- $\mathbf{H_{01}}$: In the science assessment, ELL students do not benefit from any of the accommodations used in this study. (Effectiveness)
- $\mathbf{H_{02}}$: Accommodation does not impact performance of non-ELL students on the science test. (Validity)
- $\mathbf{H_{03}}$: Student background variables do not impact performance on accommodated science assessment. (Differential impact)

To test these hypotheses concerning the use of accommodation, ELL and non-ELL students were tested under several accommodation conditions (a Customized English Dictionary, a Bilingual/English Glossary, and a Linguistic Modification version of the test items) or with no accommodation (henceforth Standard condition)

The focus of this study was on the impact of different accommodation types and students' ELL status on their performance in science. That is, the two independent variables that may impact the outcome of the science assessment are the type of accommodation and ELL status. Examining the main effect of the accommodation type will determine whether the accommodation strategies used in this study have any significant impact on the outcome of assessment (science test score). Testing the main effect of student ELL status will provide information on the performance difference between ELL and non-ELL students. Testing the interaction between accommodation type and ELL status will provide information about two of the main hypotheses of this study (effectiveness and validity).

Reading efficiency/proficiency was used as a covariate in this study. Thus, a two-factor analysis of covariance was deemed suitable for analyzing the results of the study. However, since we were interested in testing particular hypotheses, we conducted a series of a priori or planned tests. Instead of using a two-way model to test effectiveness and validity, we used a different one-way ANCOVA for testing each. To test the effectiveness hypothesis, we compared student performance under each accommodation with the Standard condition. For testing the validity hypothesis, we conducted planned tests to compare accommodated and non-accommodated outcomes. To test differential impact, we conducted a two-way ANCOVA to examine whether the impact of accommodation on science performance differed by primary home language.

Outcome variables. A test measuring students' science content knowledge provided data on the outcome (dependent variable). As a covariate, we combined reading measures from existing mainstream and ELL testing instruments. One block of NAEP items (released) provided multiple-choice and open-ended questions about a reading passage. From the LAS, the Fluency subscale was given.

The NAEP Reading block was selected from the 1994 NAEP main assessment in reading based on the results of our pilot phase of this study. The LAS Fluency subscale (rather than the complete LAS scale) was selected since the results of an earlier study suggested that the Fluency subscale has a better discrimination power when compared to other LAS subscales. In addition, we administered an experimental word recognition test lasting 2 minutes. However, the word recognition test data were not used because we discovered that this measure requires further design revisions.

We created a latent composite of all the components that measured different aspects of students' reading proficiency. This latent composite was the common variance across the NAEP multiple-choice subscale, the NAEP open-ended subscale,

and the LAS Fluency subscale. A simple-structure confirmatory factor analysis was used to create the latent composite of the various measures.

Analyses of open-ended questions. As indicated earlier, each open-ended science and reading item was scored independently by two raters. Interrater reliability indices (percent of exact and within one-point agreement, P.M. correlation, intraclass correlation, and kappa, and alpha coefficients) were computed.

For the eight open-ended science items in Grade 4, percent of agreement ranged from a low of 69% to a high of 97%. The kappa coefficient ranged from a low of .42 to a high of .94. The alpha coefficient ranged from a low of .59 to a high of .98. These results suggest that some of the open-ended items were more difficult to score reliably than others. To increase the interrater reliability, new raters, rating on consecutive days, re-scored items that had lower than average interrater reliability. Major improvement in the interrater reliability indices was observed as a result of re-scoring. For example, the kappa coefficient for item 2 improved from the original .54 to .76. For item 3a, the kappa increased from .71 to .97 and for item 5, the kappa increased from .43 to .92. A similar trend of increase was seen for all items and with the different interrater statistics.

Examining the internal consistency of the reading and science tests. The internal consistency coefficient for the entire set of Grade 8 reading items was .78. The multiple-choice reading test items had a higher alpha (.73) than the open-ended items (.56). The internal consistency coefficient for the science test (.69) was lower than the coefficient for the reading test (.78). The internal consistency coefficient was higher for the science multiple-choice items (.69) than for the science open-ended items (.45). This low internal consistency coefficient suggests that the science test may be multi-dimensional. (We improved the alpha coefficient by removing items. See the Results section.)

For Grade 4, the overall internal consistency coefficient for reading was .82. As was the case for the Grade 8 reading test, the coefficient was higher for the multiple-choice items (.79) than for the open-ended items (.60). For the science test, the overall alpha coefficient was .71.

Testing hypotheses concerning effectiveness and validity of accommodation. To test the effectiveness hypothesis, we compared the performance of ELL students who were provided accommodation in science with the performance of those ELL students who were tested under the Standard condition. A significantly higher performance under any accommodation in this study would support the effectiveness of that particular accommodation.

To test the validity hypothesis, we compared the performance of non-ELL students under accommodation with the performance of those non-ELL students who were tested under the Standard condition (i.e., no accommodation). Any

significant difference in the performance of non-ELL students may suggest an impact of accommodation on the construct, thus creating concerns about the validity of accommodation.

Results for Grade 4 students. As indicated earlier, a latent score was computed for reading and was used as a covariate. A latent score for science was computed and was used as the outcome variable. Latent scores were transformed to T-scaled scores with a mean of 50 and standard deviation of 10 (see Linn & Gronlund, 1995, pp. 454-455).

ELL students in Grade 4 had lower science test scores (T-scores *M* = 47.64, *SD* = 9.39; raw scores *M* = 7.72, *SD* = 3.18; *n* = 957) than did non-ELL students (T-scores *M* = 52.67, *SD* = 10.00; raw scores *M* = 9.37, *SD* = 3.30; *n* = 846).[6] There were slight differences in the performance of both ELL and non-ELL students under different forms of accommodation. However, these differences did not reach statistical significance. Comparing the performance of ELL students under accommodation with the performance of students under the Standard condition, Grade 4 ELL students scored slightly lower under some of the accommodations. For example, the mean T-science score for Grade 4 ELL students was 48.37 (*SD* = 9.75; raw scores *M* = 7.89, *SD* = 3.31; *n* = 270) under the Customized Dictionary condition, 45.62 (*SD* = 8.19; raw scores *M* = 7.05, *SD* = 2.74; *n* = 135) under the Bilingual/English Glossary condition, and 47.36 (*SD* = 9.48; raw scores *M* = 7.70, *SD* = 3.19; *n* = 284) under the Linguistic Modification condition, as compared to an T-mean of 48.23 (*SD* = 9.38; raw scores M = 7.93, *SD* = 3.21; *n* = 282) for ELL students under the Standard condition.

For non-ELL students in Grade 4, accommodation did not seem to have an effect. The T-mean scores were 52.81 (*SD* = 10.23; raw scores *M* = 9.41, *SD* = 3.28; *n* = 247) under the Customized Dictionary condition, 52.46 (*SD* = 9.75; raw scores *M* = 9.17, *SD* = 3.26; *n* = 101) under the Bilingual/English Glossary condition, and 52.54 (*SD* = 10.57; raw scores *M* = 9.46, *SD* = 3.52; *n* = 257) under the Linguistic Modification condition, as compared to an T-mean of 52.74 (*SD* = 9.29; raw scores *M* = 9.31, *SD* = 3.10; *n* = 241) under the Standard condition.

Similar to the science scores, the latent reading scores were transformed to scaled scores on a scale with a mean of 50 and standard deviation of 10. Consistent with the results of earlier studies, non-ELL students in Grade 4 obtained higher reading scores (T-scores *M* = 52.50, *SD* = 10.16; raw scores *M* = 11.94, *SD* = 4.20; *n* = 846) than did ELL students (T-scores *M* = 47.79, *SD* = 9.32; raw scores *M* = 9.87, *SD* = 3.82; *n* = 957). The trend of higher reading scores for non-ELL students holds across the categories of accommodations. That is, under all four accommodation conditions, non-ELL students showed higher mean reading scores than did ELL students. However, there were small differences in the mean scores across the

[6]Statistical significance results will be reported later in this section.

accommodation categories for both ELL and non-ELL groups. To control for initial differences in reading, we adjusted the science test scores for the effect of language proficiency (the students' reading scores). We then compared accommodation outcomes based on the adjusted science scores. We also controlled for Spanish being the home language, due to the makeup of the Bilingual/English Glossary accommodation group.

Effectiveness—Grade 4. To test the hypothesis concerning effectiveness of accommodation, we conducted a series of a priori tests. We were only interested in comparisons of accommodated assessments with the Standard condition. We conducted three planned comparisons, one for each form of accommodation. In the first test, we compared the science mean score (that was adjusted by reading score) under the Customized Dictionary accommodation (48.03) with the science mean score under the Standard condition (47.91). ELL students under the Customized English Dictionary accommodation scored slightly higher than under the Standard condition, but the difference did not reach the .05 statistical significance level (p = .83). In the second test we compared ELL students' performance under the Bilingual/English Glossary condition (47.28) with their performance under the Standard condition (47.91). ELL students scored slightly lower under the Bilingual/English Glossary condition than under the Standard condition. Lastly, we compared the performance of ELL students under the Linguistic Modification condition. ELL students performed slightly lower under the accommodation (47.87) than under the Standard condition (48.11). However, the difference did not reach statistical significance at the .05 level (p = .21).

Validity—Grade 4. To test the validity of accommodation, the performance of non-ELL students under accommodation was compared to the performance of non-ELL students under the Standard condition. Once again, to control for students' level of English proficiency, the science test scores were adjusted by reading scores. Three planned comparisons were conducted. In these comparisons, the performance of students under each of the three accommodations was compared to the performance of students under the Standard condition. None of the comparisons were significant. The probability for a Type I error was above .05 in all three comparisons. These results suggest that the accommodation strategies used in this study did not affect the performance of non-ELL students; therefore, the accommodations did not alter the construct under measurement.

Results for Grade 8 students. Similar to the data reported for students in Grade 4, on average, non-ELL students in Grade 8 (T-scores M = 53.12, SD = 9.59; raw scores M = 12.61, SD = 3.91; n = 856) outperformed ELL students (T-scores M = 46.35, SD = 9.21; raw scores M = 9.44, SD = 3.62; n = 733) by about 7 points on the science test. Among the ELL students, the type of accommodation made a difference in test scoring. Those who received the Linguistic Modification condition scored the highest (T-scores M = 47.63, SD = 9.53; raw scores M = 9.94, SD = 3.88; n = 209), followed by students under the Customized English Dictionary condition (T-scores

M = 46.68, *SD* = 9.00; raw scores *M* = 9.36, *SD* = 3.51; *n* = 206) and the Standard condition (T-scores *M* = 45.73, *SD* = 9.41; raw scores *M* = 9.30, *SD* = 3.70; *n* = 199). Students under the Bilingual/English Glossary condition scored the lowest (T-scores *M* = 44.58, *SD* = 8.38; raw scores *M* = 8.93, *SD* = 3.11; *n* = 119). Among the non-ELL sample, students in Grade 8 under the Bilingual/English Glossary condition performed the lowest (T-scores *M* = 50.73, *SD* = 8.58; for raw scores *M* = 11.60, *SD* = 3.87; *n* = 129). Under the Customized English Dictionary condition (T-scores *M* = 53.17, *SD* = 9.84; raw scores *M* = 12.73, *SD* = 3.81; *n* = 241) and the Linguistic Modification condition (T-scores *M* = 54.00, *SD* = 8.97; raw scores *M* = 12.90, *SD* = 3.76; *n* = 241), non-ELL students performed about the same as those under the Standard condition (T-scores *M* = 53.48, *SD* = 10.26; raw scores *M* = 12.75, *SD* = 4.10; *n* = 245).

On the reading assessment of Grade 8 students, ELL students performed substantially lower (T-scores *M* = 45.93, *SD* = 9.16; raw scores *M* = 9.37, *SD* = 2.84; *n* = 733) than non-ELL students (T-scores *M* = 53.49, *SD* = 9.36; raw scores *M* = 12.33, *SD* = 2.75; *n* = 856). There were also some differences in the reading test scores across the accommodation categories. For example, ELL students who took the science test under the Customized English Dictionary condition had the highest reading scores (T-scores *M* = 48.78, *SD* = 9.10; raw scores *M* = 9.54, *SD* = 2.70; *n* = 206), but the highest reading scores for non-ELL students were the scores of students who took the science test under the Linguistic Modification condition (T-scores *M* = 54.34, *SD* = 8.74; raw scores *M* = 12.62, *SD* = 2.92; *n* = 241). However, because the reading scores were then used to adjust the science test scores, these initial differences did not affect the outcome of this study. (Again, we controlled for Spanish being the home language.)

Effectiveness—Grade 8. To test the effectiveness hypothesis, the performance of ELL students under accommodation was compared to the performance of ELL students under the Standard condition. The Linguistic Modification version was the only accommodation that significantly impacted the performance of Grade 8 ELL students, which was significant at the .05 nominal level (*p* = .03). The performance of ELL students was highest under this accommodation (*M* = 47.52, *SE* = .50, *n* = 209). The other two accommodations did not show any significant impact on the performance of ELL students. For the Customized English Dictionary condition, the probability of a Type I error was .91. For the Bilingual/English Glossary condition, the *p* value was .68.

Validity—Grade 8. The results for effectiveness of the Linguistic Modification accommodation are not enough to judge its appropriateness in the assessment of ELL students. We must also ensure that the accommodation did not help non-ELL students; otherwise, its validity would be compromised. To test the validity of accommodations for the Grade 8 data, we compared the performance of non-ELL students under the different accommodations. None of the comparisons were significant. That is, none of the accommodation strategies had any impact on non-

ELL students' performance in science. The mean science score for non-ELL students under the Linguistic Modification condition (M = 53.42, SE = .47, n = 241) was almost identical with the mean science of students under the Standard condition (M = 53.38, SE = .47, n = 245). These results suggest that the accommodation strategies we used did not impact the construct under measurement and can be used for ELL students without adversely affecting the validity of this accommodation.

Like the approach taken in Grade 4, the latent reading score was used as a covariate in the model comparing students' science scores under the different forms of accommodation. That is, students' science scores were adjusted by their reading proficiency scores.

Impact of primary home language. To determine whether student background variables—such as primary home language—impacted performance on the accommodated assessments, we examined the students' primary home languages and science test results under various accommodation conditions. Noticeable differences occurred under the Linguistic Modification condition. Under this accommodation, students whose primary home language was English performed slightly lower on science than those under the Standard condition. Students with Spanish as a primary home language performed slightly higher with the Linguistic Modification accommodation than under the Standard condition. Students with other non-English home languages performed substantially higher on science under the Linguistic Modification condition than under the Standard condition. In other words, it appears that students whose primary home language is neither English nor Spanish benefited the most from the Linguistic Modification version of the test.

Item-level analysis. As discussed earlier, linguistic modification of test items was the only effective accommodation in this study. To further investigate the performance of students under this accommodation, we performed analyses at the item level. The results of our analyses showed that the difference in the p value (proportion of correct responses) between ELL and non-ELL students differed greatly across the science test items according to the item's level of linguistic complexity. The higher the level of linguistic complexity, the larger the performance difference between ELL and non-ELL students. The larger the performance difference between ELL and non-ELL students, the more linguistic modification of test items helped reduce the performance gap.

Analyses indicated that when compared with the Standard condition, ELL students under the Linguistic Modification condition showed more improvement than those under other accommodation conditions. For example, out of 30 science items, students performed better on 22 items under the Linguistic Modification condition as compared with the Customized English Dictionary condition (performed better on 14 items) and the Bilingual/English Glossary condition (performed better on 13 items). There were 13 items on which non-ELL students outperformed ELL students with a p-value difference of 0.11 or greater. For all these

items, students who received the Linguistic Modification version of the science test outperformed those who received the Standard condition.

The correlation between the *p*-value differences of ELL and non-ELL students and the *p*-value differences of accommodated students and those under the Standard condition was the greatest for Linguistic Modification ($r = -.60$, $p < .01$). This high negative correlation also suggests that linguistic modification of test items reduced the performance gap between ELL and non-ELL students more than the other accommodations.

Discussion

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

Both ELL and non-ELL students in Grades 4 and 8 were tested in science under accommodation or under a standard testing condition. Language accommodation strategies (Customized English Dictionary, Bilingual/English Glossary, and Linguistic Modification of test items) were selected based on frequency of usage, nationwide recognition, feasibility, and first-language literacy factors. Students were sampled from different language and cultural backgrounds. We also included a measure of English reading proficiency to control for any initial differences in reading ability.

The results of this study show that some of the accommodation strategies used were effective in increasing the performance of ELL students and reducing the performance gap between ELL and non-ELL students. The results suggest that the effectiveness of accommodation may vary across the grade level. In general, accommodations did not have a significant impact on students' performance in Grade 4. Neither ELL nor non-ELL Grade 4 students benefited from any of the three accommodation strategies that were used. We believe this may be because language demand in textbooks and tests may be different in lower than in higher grades, and all three accommodation strategies used were specifically language related. With an increase in the grade level, more complex language may interfere with content-based assessment. Though language factors still have impact on the assessment of ELL students in lower grades, other factors such as poverty and parent education may be more powerful predictors of students' performance in lower grades. Another consideration is that Grade 4 students may be less familiar with glossary and dictionary use, as well as less exposed to science.

The lack of significant impact on Grade 4 non-ELL students is an encouraging result because it suggests that the accommodation did not alter the construct under measurement.

The findings of this study on the effectiveness of accommodation for Grade 8 students were different from the findings for Grade 4 students. The Linguistic Modification accommodation was shown to have a significant impact on the ELL students' performance. This impact was in the expected direction. That is, this accommodation helped ELL students to increase their performance while the accommodated performance of non-ELL students was unchanged. A nonsignificant impact of the linguistically modified test on the non-ELL group assures the validity of this accommodation. As for feasibility, this accommodation requires up-front preparation, but is easy to implement in the field; therefore, it is feasible for large-scale assessments. For further considerations, please see the Discussion section of the full report.

CONTENTS

EFFECTIVENESS AND VALIDITY OF ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNERS IN LARGE-SCALE ASSESSMENTS

Jamal Abedi, Mary Courtney, and Seth Leon
CRESST/University of California, Los Angeles

Introduction

Every test that uses language is, in part, a language proficiency test. One must understand the language of the test properly in order to respond accurately. However, this is problematic for testing those who have yet to acquire a firm grasp of that language. Test results may not accurately reflect an individual's knowledge in a content-based assessment if performance is affected by language proficiency. Thus, special attention must be made to ensure that assessments are accurately measuring content knowledge.

Recent policy mandates, including the Improving America's Schools Act of 1994 and the No Child Left Behind Act of 2001, illustrate the continued interest in education and assessment. Since assessment results often shape curriculum and instruction, it is only fair that accurate assessments are made. However, for English language learners (ELL),[1] content-based tests may often inadvertently function as language proficiency tests. It is therefore imperative that we seek means of accommodation to reduce construct-irrelevant factors in assessments, especially with the continuing increase in numbers of ELL students. However, national research in the areas of assessment and accommodation of ELL students suggests that there may not be a simple solution to these national issues (e.g., Rivera, Stansfield, Scialdone, & Sharkey, 2000).

---

[1]The terms *English language learner* (ELL) and *limited English proficient* (LEP) are both 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 with 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, in contrast, conveys that the student has a deficit or a "limiting" condition. Since accommodations are specifically intended for use with the LEP population of ELL students, instances of the term ELL in this report generally refer to this LEP population.

Research supports the notion that language is a crucial factor in the assessment of ELL students. In their study of English language proficiency and academic achievement, Butler and Castellon-Wellington (2000) indicated that there is a strong relationship between ELL students' level of language proficiency and their performance on content assessments. By examining the language demand in content assessment, Bailey (2000) hypothesized that the test items most differentiating the performance of ELL and non-ELL students are those with greater language difficulty.

In their analyses of extant data, Abedi and Leon (1999) found that language was a significant factor in content-based assessment for ELL students. They also found structural differences between the performance of ELL and non-ELL students. They indicated that language is a significant source of measurement error for ELL students, especially for those at lower levels of English proficiency.

Butler and Castellon-Wellington (2000) found evidence that language factors confounded the content assessment of ELL students. Thus, for ELL students, the outcome of content-based assessment may not be a fair assessment of what they know in content-based areas.

To help reduce the effect of language factors in the assessment of ELL students, accommodations have been suggested and used in large-scale national and local assessments (see, for example, Mazzeo, Carlson, Voelkl, & Lutkus, 2000; Olson & Goldstein, 1997; Rivera et al., 2000). Accommodations were provided for students with limited English proficiency in the National Assessment of Education Progress (NAEP) test administrations. The 1996 NAEP assessment provided the first series of studies evaluating testing accommodations and their effectiveness, using oversampling of Grades 4, 8, and 12 ELL students (Goldstein, 1997; Mazzeo, 1997).

However, issues have been raised regarding the use of accommodation for ELL students. We must ensure not only that accommodations are effective, but also that they are valid and feasible to implement. In evaluating the NAEP accommodation data, a major limitation was the lack of control or comparison groups; the same limitation applies to all other national large-scale accommodation data. To test the validity of accommodation—that is, to find out whether accommodation actually impacts the constructs being assessed—both ELL and non-ELL student groups should be tested under accommodation and no-accommodation conditions.

Abedi, Hofstetter, Lord, and Baker (1998) found that some forms of accommodation improved the performance of non-ELL students more than that of ELL students. That is, the performance gap between ELL and non-ELL students was increased by the provision of accommodation, rendering it invalid.

Feasibility is another major issue in the use of accommodation in the assessment of ELL students. Some forms of accommodation are difficult to implement, especially in large-scale assessments. For example, one-on-one testing may be difficult or expensive, or both, and in some situations even impossible. English and bilingual dictionaries are frequently used as a form of accommodation for ELL students. However, their effectiveness is dependent on the students' familiarity with dictionaries and their inclination to take advantage of published language tools. The pilot study for our project demonstrated the logistical difficulty of using commercial English and bilingual dictionaries (see Abedi, Courtney, Mirocha, Leon, & Goldberg, 2001).

With support from the Office of Educational Research and Improvement (OERI) and the Office of Bilingual Education and Minority Languages Affairs (OBEMLA), researchers at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) designed a study to examine issues concerning the provision of accommodation for ELL students. The following research questions guided this study:

- Do accommodation strategies help reduce the performance gap between ELL and non-ELL students by removing language barriers? (Effectiveness)
- Do accommodation strategies impact the validity of the assessment, that is, do they change the content of the assessment? (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 concerning the use of accommodation, we tested both ELL and non-ELL students in Grades 4 and 8 in science under three accommodation conditions, or under a Standard condition in which no accommodation was provided. Students were sampled from different language and cultural backgrounds.

Accommodation strategies were selected based on frequency of usage, nationwide recognition, feasibility, and first-language literacy factors. Three forms of accommodation were used in this study: a Customized English Dictionary, an English-to-Spanish Glossary, and a Linguistic Modification (i.e., linguistically simplified) version of the test items. Each one can clearly function as an aid to the language needs of ELL students on large-scale assessments.

An accommodation was randomly assigned to ELL and non-ELL students within each classroom with consideration of home language in assigning the English-to-Spanish Glossary. Thus, 8 comparison groups were possible: 4 levels of accommodation by 2 levels of ELL status. However, since there was no practical reason to give a bilingual glossary to a non-ELL student, this group was given an English-to-English glossary that offered a simpler replacement word for any unfamiliar non-science word in the test.

Additionally, we included a measure of English reading efficiency / proficiency to control for initial English language differences, since the level of English ability among both the ELL and non-ELL students was different and may have affected the accommodation strategies.

Literature Review

The population of English language learners continues to grow rapidly in size. Between 1990 and 1997, the number of United States residents not born in the U.S. increased by 30% (Hakuta & Beatty, 2000). According to the 2000-2001 *Survey of the States' Limited English Proficient (LEP) Students* (Kindler, 2002), more than 4.5 million LEP students were enrolled in public schools.

The continuing increase in numbers of ELL students has prompted increased interest in upholding instruction standards and providing fair assessments. Subsequently, federal legislation in the last 10 years, including the No Child Left Behind Act of 2001 (NCLB), Goals 2000, and the Improving America's Schools Act of 1994 (IASA), has aimed to improve instruction and assessment by mandating inclusion of all students in large-scale assessments. Validity and equitability of inferences drawn from standardized assessments, especially for ELL students, have thus gained much attention. This consequently affects assessment design, delivery, interpretation, and use.

Standardized achievement tests play a pivotal role in education. They tend to shape instruction and student learning (Linn, 1995). They are used for accountability and grade promotion. They are also frequently used for assessment and classification of ELL students (Zehler, Hopstock, Fleischman, & Greniuk, 1994). It is therefore imperative that standardized assessments be fair and accurate for all students.

However, students' language background factors can reduce the validity and reliability of inferences drawn about their content-based knowledge. Content-based assessments (such as in science and math) are conducted in English and normed on native English speaking test populations, thereby inadvertently functioning as English language proficiency tests.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) reminds us that

> [f]or all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is not the language of the test. Test use with individuals who have not sufficiently acquired the language of the test may introduce construct-irrelevant components to the test process. In such instances, test results may not reflect accurately the qualities and competencies intended to be measured. (p. 99)

Findings of a series of studies conducted by CRESST on the impact of students' language background on their performance indicated that

- language background affects performance in content-based areas such as math and science;
- linguistic complexity of test items may threaten the validity and reliability of achievement tests, particularly for ELL students; and
- as the level of language demand decreases, the performance gap between ELL and non-ELL students decreases.

(See Abedi & Leon, 1999; Abedi, Leon, & Mirocha, 2001; Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000.)

Since exemption from assessments is granted only for specific cases, schools have turned to other alternatives, including providing accommodation to help students overcome language barriers. We will now discuss issues regarding assessment of ELL students' language accommodations used in this study.

Validity Issues for Assessing ELL Students

Some ELL students may have the content knowledge or the cognitive ability, or both, needed to perform successfully on assessment tasks, but may not yet be able to demonstrate in English what they know. Therefore, assessment procedures may not be equitable and may not yield valid results for ELL students (Gandara & Merino, 1993; LaCelle-Peterson & Rivera, 1994). The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) point out that whenever students who are still in the process of learning English are tested in English, regardless of the content or intent of the test, their proficiency in English will also be tested.

When testing academic achievement in content areas, assessments must provide valid information. Ideally, instruments will yield beneficial and accurate information about students. In order to provide the most meaningful data, a number of questions are addressed when evaluating assessments (LaCelle-Peterson & Rivera, 1994).

Technical / validity questions:

- Is the test valid for the school populations being assessed—including 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 demonstrate knowledge of 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 been provided that would give ELL students the same opportunity available to fluent English proficient students?

Performance Differences Between ELL and Non-ELL Students

Research has found that students' language background confounds their performance on math word problems (see, for example, Abedi & Lord, 2001). Language background may also confound scores on science tests if language comprehension, rather than content knowledge, is reflected in scores. Linguistic complexity of test items may affect students' ability to perform on a test. ELL students may be unfamiliar with the linguistically complex structure of questions, may not recognize vocabulary terms, or may misinterpret an item literally (Durán, 1989; Garcia, 1991). They may also perform less well on tests because they read more slowly (Mestre, 1988).

In addition to language difficulties, cultural variables may also affect test results. Such variables include attitudes toward competition, attitudes toward the importance of the individual versus the importance of the group or family, gender roles, attitudes toward the use of time, and attitudes toward the demonstration of knowledge (see Liu, Thurlow, Erickson, Spicuzza, & Heinze, 1997). Analyses of students' answers to background questions and their math and reading scores (Abedi, Lord, & Hofstetter, 1998) indicated that language-related background variables, length of time in the United States, overall grades, and the number of school changes are valuable predictors of ELL students' performance in math and reading.

According to Mazzeo et al. (2000), of the Grade 8 ELL students participating in the NAEP 1996 mathematics sample, 19% were receiving science instruction 2 or more years below grade level. Their teachers estimated that at least 34% of them were performing 2 years or more below grade level in science. Taking an assessment written for students who receive instruction at their corresponding grade level would thus be difficult for many ELL students (see Tables 1 and 2).

Defining Accommodation

Accommodations, sometimes referred to as *modifications* or *adaptations,* are intended to "level the playing field," so that students may provide a clearer picture of what they know and can do, especially with regard to content-based assessments (e.g., in mathematics and science), where performance may be confounded with

Table 1

Percentage Distribution of ELL Students' Level of 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.

Table 2

Percentage Distribution of ELL Students' Estimated Grade Level of Performance in Science 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.

their English or home language proficiency or other background variables. Accommodations are not intended to give ELL students an unfair advantage over students not receiving an accommodated assessment.

The umbrella term "accommodation" includes two types of changes: modifications of the test itself and modifications of the test procedure. The first type, changes in the test format, includes translated or adapted tests, for example:

1. assessment in the student's home language;
2. bilingual versions of the tests;

3. modification of linguistic complexity;

4. use of glossaries in the home language and/or English that are embedded into the test.

These accommodations directly address the linguistic needs of the student, but they must be designed with care to ensure that the accommodated format does not change the test content, that is, the construct being measured. For this reason, schools have more often employed accommodations of the second type: changes in the test procedure. Examples (from Rivera et al., 2000) include:

1. allowing English language learners to have extended time to take the test on the same day;

2. multiple testing sessions, small group or separate room administration, individual administration;

3. administration by a familiar test administrator;

4. use of published dictionaries or glossaries;

5. simplified directions;

6. repeated instructions;

7. translating the directions; and

8. reading the directions or questions aloud.

Assessment of ELL Students

The IASA of 1994 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 subjects other than English. . . ." The debate on what form such assessments should take continues. In search of accurate assessments for ELL students, some individual classrooms and schools have turned to alternative assessments, such as portfolios, interviews and oral testing. These, however, are not cost-effective and are too time-consuming for large-scale assessments.

Because large-scale assessments do not effectively assess the content knowledge of ELL students, many students have traditionally been exempted from exams. According to the 1999-2000 State Student Assessment Programs (SSAP)

annual survey summary report (Council of Chief State School Officers [CCSSO], 2001), official criteria for exemption are often based on one or more of the following:

- time in the United States;
- time in an English as a Second Language (ESL) program;
- formal assessments of English;
- informal assessments of English.

Most local and state assessments still allow for the exemption of some ELL students, but they also administer various test accommodations, based on cost considerations, political expediency, or feasibility of administration (Kopriva, 2000).

Local communities, school, districts, parents, or a combination of these decide assessment exemptions. (For the practices in specific states, refer to Roeber, Bond, and Connealy, 1998, and Rivera et al., 2000.) Exempting students from assessments, however, does not provide for a measurement of progress and may not allow students opportunities, such as additional instruction, that could be offered based on the assessments. Instead, assessments must provide more appropriate instruments to monitor and report the progress of ELL students across districts, states, and the nation.

State Policies on Accommodation

States vary on policies regarding the identification of ELL students and the role of accommodation on 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.

In California, one of the top 10 states having Spanish-speaking ELL students, students are determined to be ELL 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, ELL students are identified based on a home language survey, oral language proficiency testing, informal assessment through a teacher/parent interview, student interview or teacher survey, standardized achievement test scores, and classroom grades. Beginning in the 2000-01 school year, all Texas ELL

students will take the Texas Assessment of Academic Skills (TAAS) in English or Spanish unless a student is a recent unschooled immigrant enrolled in U.S. schools for 12 months or less. 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 home language and translating of some components of the test in a student's home language. School district officials are the decision makers for ELL accommodation.

The 1999-2000 SSAP annual survey (CCSSO, 2001) examined participation in state assessments by ELL students and found that 29 states allowed accommodations for ELL students in all assessments, 18 allowed accommodations with some assessments, and 4 did not permit any accommodations for ELL students. An alternate assessment, often used for the least English proficient students, who would have been excluded before, was available for ELL students in 16 states. A variety of accommodations were allowed that year: ELL students were assessed in a modified setting (44 states); with a modified format of presenting the assessment (43 states), such as directions read aloud, interpreted, repeated, etc.; with a change of timing or scheduling (41 states); and/or with a modified method of responding to the questions, such as marking responses in the booklet, using a computer, or having a scribe record their answers. The other accommodations listed in the report (permitted in 27 states) were word lists, dictionaries, or glossaries. According to Rivera et al. (2000), a survey of state assessment directors for 1998-1999 found 21 states that allowed bilingual dictionary accommodations on reading tests; 11 of the 21 allowed them for all parts of the assessment.

Rivera, Vincent, Hafner, and LaCelle-Peterson (1997) noted that 52% of states reported that they allowed test modifications for ELL 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, 1996a, 1996b; see also Liu et al., 1997) also found that half of the states reported that they did allow accommodation for ELL 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 for the test or an alternate test (Liu et al., 1997).

In general, state policies on the process of identifying ELL students contain some similarities, including collecting information from assessments and identifying home language. Not all states have specific accommodation policies, although all

states seem to be addressing concerns about including all students in large-scale assessments. (For a more detailed discussion and for information on other states, see Rivera et al., 2000.) More research, however, is needed to determine the most effective ways to accommodate ELL students.

Evaluating the Use of Accommodation

Accommodation may improve the accuracy of test scores by eliminating irrelevant obstacles for ELL students (Rivera & Stansfield, 1998). 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. These tests will be more of a measure of the individual's true ability in the subject being assessed than scores earned on tests without appropriate accommodation. The accommodation may also increase the comparability of scores.

Although appropriate test accommodation helps "level the playing field," it is important that accommodations not give an unfair advantage to students who receive them over students who do not (Rivera & Stansfield, 1998). For example, students who have access to standard published dictionaries during an assessment may be able to correctly respond to certain items only because the answer is contained within a definition. Providing extra time only for ELL students may give an unfair advantage if other students have lower scores simply due to lack of time to complete test items (see Abedi, Lord, et al., 1998; Abedi, Lord, Hofstetter, et al., 2000; Hafner, 2001).

Past NAEP assessments have included accommodations that can be examined. In the 1995 NAEP field test, accommodations for mathematics included extra testing time, modifications in the administration of sessions, and facilitation in the reading of directions. Also available were Spanish-English bilingual assessment booklets, and Spanish-only assessment booklets, which most ELL students chose to take. The results (Olson & Goldstein, 1997) indicated that the translated versions of some items may not have been parallel in measurement properties to the English versions. (No accommodations were tested for science items in the field test.)

The NAEP 1996 tests were designed with three samples of schools, using the 1996 inclusion criteria in the second and third samples and having assessment accommodations available in the third sample. ELL students were permitted any of these accommodations: one-on-one testing, small group testing, extended time, oral reading of directions, signing of directions, use of magnifying equipment, and the

use of an individual to record answers—plus a Spanish/English glossary of scientific terms. Students using the glossary were usually given extra time. Very few ELL students used the glossary (O'Sullivan, Reese, & Mazzeo, 1997).

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 NAEP 1996 science assessment were not developed, due to resource constraints and comparability questions about the results obtained with standard and translated versions of NAEP instruments.

The limitations of NAEP accommodation data prompted us to conduct further investigation regarding accommodation usage. Abedi, Lord, Hofstetter, et al. (2000) found that linguistically modified testing, extra time, and glossary plus extra time helped ELL students. Results suggested that the effectiveness of accommodation strategies might depend, to some extent, on the students' background variables, particularly their language background variables. These promising results encouraged us to focus specifically on language accommodations.

Effective, valid, and economical accommodations on national standardized tests will allow schools, districts, and states to be compared more reliably. We will now discuss a few accommodation strategies that we have evaluated in both our pilot studies and the present study.

Language Accommodation Strategies

Our study focused on accommodations that directly address the students' anticipated difficulty with the language of the text. This section summarizes several findings pertinent to the effectiveness, validity, and feasibility of the accommodation approaches used in the study. We will discuss commercially published English dictionaries, glossaries, linguistic modification of test items, and extra time.

Commercially published English dictionaries. For students with limited vocabularies in English, the provision of a commercially published English dictionary seems a practical accommodation alternative. This would be especially helpful for students who are already familiar with using dictionaries. However, this strategy may pose as a difficulty for students who are unfamiliar with dictionary usage. Furthermore, commercially published dictionaries may give an unfair advantage if answers to test items can be found within definitions.

Some "ESL" dictionaries have entries suitable for ELL students to understand (Kopriva, 2000). When commercially published ESL dictionaries were provided to urban middle school students in Minnesota during a reading test (Thurlow, 2001), the scores of 69 non-ELL students and 133 Hmong ELL students did not significantly differ, regardless of whether or not they had the dictionary. However, for those Hmong students who reported using the dictionary and had self-reported an intermediate English reading proficiency, the scores showed some mild improvement. Therefore, the dictionary accommodation may not be appropriate for the lower level ELL students, who did not seem to benefit from the accommodation.

In the pilot study for this project (Abedi, Courtney, et al., 2001), commercial dictionaries and glossaries were provided as forms of accommodation. However, they were determined to be ineffective in the analysis and seemed to be disregarded by many students. Providing a dictionary to each student also appeared to be logistically unfeasible for large-scale assessments.

Commercially published bilingual dictionaries (Glossaries). Some states that allow commercially published bilingual dictionaries as an accommodation, such as Ohio and Massachusetts, have approved lists of bilingual dictionaries (Rivera & Stansfield, 1998). These bilingual dictionaries actually function as bilingual glossaries by merely translating items rather than defining them. It is possible that the states want to ensure that larger, expanded dictionaries, which give an unfair advantage to students, are not used.

Customized glossary and dictionary use. In order to overcome the main disadvantages of commercial dictionary use as an accommodation (accidental provision of test content material, difficult format and language, the difficulty of providing dictionaries, and non-use), we created customized glossaries and dictionaries for our study that are defined and discussed next.

The concise glossaries created for this study provided the simplest and most item-appropriate translation/synonym for each difficult non-science word in the test. The customized English dictionary simulated the look of full entries of a dictionary without the bulk of the entire text or the unfair advantage of providing definitions for terms and concepts being tested.

A study of 422 students in Grade 8 science classes (Abedi, Lord, Kim, & Miyoshi, 2000) found that ELL students scored highest with a customized English dictionary accommodation, when compared with performance on NAEP Science

items in two other test formats: a test booklet in original format (no accommodation) and one with English glosses and Spanish translations in the margins. The customized English dictionary accompanied a test at the end of the test booklet and included only words that appeared in the test items. (The mean scores for the three formats were 10.18, 8.36, and 8.51, respectively, on a 20-item test). Although the accommodations helped the ELL students score higher, there was no significant difference between non-ELL students in the three test formats. This suggests that these accommodation strategies did not affect the construct.

Linguistic modification. Linguistic modification of test items involves modifying the language of the test text while maintaining the construct. Assessments that are linguistically modified may facilitate students' negotiation of language barriers. 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 (such as present tense), and using concrete rather than abstract formats (Abedi, Lord, & Plummer, 1997).

Abedi et al. (1997) found significant differences with respect to language background between student scores on complex items and less complex items. Abedi and Lord (2001) found that modifying the linguistic structures in math word problems can affect student performance. Students indicated preferences for items that were less linguistically complex in interviews and also scored higher on linguistically modified items. The linguistic modification had an especially significant impact for low-performing students and English language learners.

Content-based standardized achievement tests aim to measure students' knowledge of specific content areas. To accurately assess knowledge within content areas, students must comprehend what the items are asking and understand the response choices. However, analyses of mathematics and science subsections of 3rd- and 11th-grade standardized content assessments by Imbens-Bailey and Castellon-Wellington (1999) showed that two thirds of the items included general vocabulary considered uncommon or used in an atypical manner. One third of the items included complex or unusually constructed syntactic structures.

In *Ensuring Accuracy in Testing for English Language Learners,* the LEP Consortium of the CCSSO State Collaborative on Assessment and Student Standards (SCASS) gave seven recommendations for improving accessibility of text material (Kopriva, 2000). Table 3 summarizes research findings of Abedi et al. (1997)

Table 3

Linguistic Complexity: Research Findings and Practical Recommendations

Research findings[a]	Practical recommendations[b]
Short words (simple morphologically) tend to be more familiar and therefore easier.	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.	Use the present tense, use active voice, avoid the conditional mode, and avoid starting statements and questions with clauses.

[a]Source: Abedi et al. (1997).

[b]Source: Shuard and Rothery (1984).

accompanied by practical recommendations from Shuard and Rothery (1984) and Kopriva.

In studies examining the language of math problems, making minor changes in the wording of a problem affected student performance (Cummins, Kintsch, Reusser, & Weimer, 1988; DeCorte, Verschaffel, & DeWin, 1985; Hudson, 1983; Riley, Greeno, & Heller, 1983). Larsen, Parker, and Trenholme (1978) compared student performance on math problems that differed in sentence complexity and level of familiarity of the non-math vocabulary. Low-achieving Grade 8 students scored significantly lower on the items with more complex language.

Studies using items from NAEP compared student scores on actual NAEP items with parallel modified items in which the math task and math terminology were retained, but the language was simplified. One study (Abedi & Lord, 2001) of 1,031 Grade 8 students found small but significant score differences for students in low- and average-level math classes. Among the linguistic features that appeared to contribute to the differences were low-frequency vocabulary and passive voice verb constructions (see Abedi et al., 1997, for discussion of the nature of and rationale for the modifications).

Another study (Abedi, Lord, et al., 1998) of 1,394 Grade 8 students in schools with high enrollments of Spanish speakers showed that modification of the

language of the test items contributed to improved performance on 49% of the items; the students generally scored higher on shorter problem statements. A third study (Abedi, Lord, Hofstetter, et al., 2000) tested 946 Grade 8 students in math with different accommodations including modified linguistic structures, provision of extra time, and provision of a glossary. Among the different options, only the linguistic modification accommodation narrowed the score gap between ELL and non-ELL students.

Other studies have also employed language modification of test items. Rivera and Stansfield (2001) compared student performance on regular and modified Grade 4 and Grade 6 science items. Although the small sample size did not show significant differences in scores for ELL students, the study did demonstrate that linguistic simplification did not affect the scores of non-ELL students, indicating that linguistic simplification was not a threat to score comparability.

Extra time. Allowing more time to complete test sections than is normally allotted is a common accommodation strategy, possibly because it does not require changes to the test itself and is easier to implement. This accommodation may lead to higher scores for ELL students (Hafner, 2001; Kopriva, 2000), possibly because the extra time better permits the decoding of the academic English in the test.

There has been inconclusive research on extra time as an accommodation strategy for ELL students. In a study allowing extra time for samples of both ELL and non-ELL students, all students with the extra time condition had the highest scores (Hafner, 2001). While extra time helped Grade 8 ELL students on NAEP math tests, it also aided non-ELL students, thus putting to question its validity as an assessment accommodation for ELL students (Abedi, Lord, et al., 1998; Abedi, Lord, Hofstetter, et al., 2000). Therefore, when extra time is allotted, it should be given to all students.

Extra time is considered a necessary addition when time-consuming accommodations are provided. A study providing glossaries with extra time (Abedi, Lord, Hofstetter, et al., 2000) on Grade 8 math tests for 946 southern California students found that both ELL and non-ELL students performed significantly higher when extra time was provided along with the glossary. Provision of extra time only, or the glossary only, had lesser impact; in fact, for ELL students, provision of only the glossary resulted in slightly lower scores, probably a consequence of information overload.

Reading Assessment of Proficient and Non-Proficient Readers

Following is a brief discussion of findings on the LAS Fluency section, the NAEP Reading block, and the sight-word recognition portions of our assessment. We decided that measures of reading proficiency were necessary as a covariate for our analyses, since students' reading abilities are not homogenous, and students may benefit differently from different accommodations. Reading proficiency assessments are normed for either non-ELL students (such as NAEP's Reading Comprehension blocks) or for ELL students (such as the LAS test battery). There seems to be no single written assessment suitable for both types of readers. Consequently, we used a combination of measures for both ELL and non-ELL students.

LAS Fluency section and NAEP Reading block. The Fluency section of the Language Assessment Scales (LAS) showed a higher level of discrimination power in assessing reading ability among ELL students in a previous CRESST study, whereas intact blocks of NAEP Reading items provided a good distribution among non-ELL students in the pilot portion of this project (Abedi, Courtney, et al., 2001).

Word Recognition. In one second or less, a sight word is recognized without pausing to break it into parts (phonemic decoding). Once students have a large vocabulary of sight words, they are free to concentrate on constructing the meaning of text (Gough, 1996). Since word recognition is central to the reading process (Chard, Simmons, & Kameenui, 1998), word recognition tests may help determine reading levels. Testing word recognition can serve as an easier way of assessing reading levels. Although comprehensive reading assessments tend to be more valid in determining reading ability, word recognition tests still provide a valid estimate of student ability and can be given in less time than comprehensive assessments.

Vocabulary checklists are a type of word recognition test that has been used by various researchers (Read, 2000). The Eurocentres Vocabulary Size Test (EVST) (Meara & Buxton, 1987; Meara & Jones, 1988) has been used to estimate the vocabulary size of ELL students by using a graded sample of words covering numerous frequency levels. This test also uses non-words to provide a basis for adjusting the test takers' scores if they appear to be overstating their vocabulary knowledge. Because the EVST is administered by computer, some have viewed it as an efficient and accurate placement procedure, able to assign students to levels with minimal effort (Read).

EVST and other checklist tests can give a valid estimate of the vocabulary size of most ELL students (Read, 2000). Exceptions, however, include learners at low levels of proficiency and individual learners whose pattern of vocabulary acquisition has been unconventional. Despite these concerns, Meara (1996) expressed optimism that the problems with checklist tests can be overcome and that they can provide satisfactorily reliable estimates of vocabulary size. The great attraction of the checklist format is how simple it is, both for its construction and for the test takers to respond to. Its simplicity means that a large number of words can be covered within the testing time available, which is important for achieving the sample size necessary for making a reliable estimate (Read).

Oral or written word recognition assessments may be effective measures of reading ability. Although comprehensive reading assessments tend to be more valid in determining reading ability, word recognition tests still provide a valid estimate of a student's ability and are able to be given in a shorter period of time than the comprehensive assessments.

Methodology

This study investigated the use of accommodation by ELL students on a test comprised of NAEP Science questions and a few Third International Mathematics and Science Study (TIMSS) multiple-choice items. The study was conducted between October 2000 and June 2001 in four urban school districts within three states with large minority populations.

Participants

A total of 1,854 Grade 4 students, 1594 Grade 8 students, and 104 teachers (132 classes) at 40 school sites participated in the study.[2] Out of 3,448 students, 1,712 (49.7%) were identified as being ELL. Three language categories were targeted for the main study: Spanish, Chinese, and "other Asian languages."[3] Of the 1,712 ELL students, 1,614 (94.3%) belonged to one of the three target language categories.

[2]A small number of students were excluded from the study because they were completely non-English speaking, were enrolled in a different grade, or were administered an inappropriate accommodation. In one site, two students designated "NEP" for "non-English proficient" were excluded from the analysis. The number of participating students may not always match the number of students included in the analysis (by a small margin).

[3]The more than 200 Asian students combined in this category spoke languages from East Asia, Southeast Asia, the Philippines, and the Pacific Islands. Notable among these, in order of frequency in the study population, are Korean, Vietnamese, Tagalog, Mien, Khmer, Lao, Ilocano, and Samoan.

Teachers or administrators determined the English proficiency designation of students based on their school's records. The 1,736 non-ELL students either had English as a home language or had become proficient enough in English to exit the ELL program. For the sake of this study, these students were combined into our non-ELL category.

ELL students were studying science in different settings: either in bilingual programs, 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 totally of ELL students. In several cases, Grade 8 ELL students who were not in the same class but who had the same science teacher were assembled together in order to provide a significant number of ELL-designated participants from the target language group.

Languages and districts. In order to test a variety of ELL students, four test sites were targeted, two with a preponderance of Spanish-speaking ELL students (Sites 1 and 3), and two with many Asian-language-speaking ELL students (Sites 2 and 4). Sites 1, 2 and 3 were visited in the last part of 2000 to test Grade 4 students and in the first half of 2001 to test Grade 8 students. Site 4 was visited in May and June of 2001 to test both Grades 4 and 8 students.

Participants were identified not only by language proficiency designation, but also by their home language, classroom, school and district. We examined test results by each of these affiliations because linguistic, cultural, curricular, and institutional variables cannot be ignored. Test-taking experience varied from district to district. All Grade 8 students tested had recently taken standardized tests, albeit different ones, depending on the locale. The types of tests varied, too. For example, Grade 4 students in Site 4 had recently finished a state assessment consisting of 90 minutes of uninterrupted open-response writing. Grade 8 students in Site 3 had completed state assessments, and Grade 8 students in Sites 1 and 2 had completed Stanford 9 testing.

Region, school and class selection. Schools participating in this study were chosen based on the largest second language groups in the United States and then by research of specific locations where there were communities belonging to these groups. After specific areas were chosen, schools were selected by determining the percentage of Grades 4 and 8 ELL students, 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 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 Grades 4 and 8 science classrooms with an equal distribution of ELL (from target language) students and non-ELL students. The reality of classroom demographics, however, required more flexibility and, at times, ingenuity to get 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. The size of the classes ranged from 9 to 36 students.

Setting

Students took the test in their normal classroom setting, except when a particular science teacher's Asian Grade 8 ELL students could be collected from throughout the building and assembled in a library or spare classroom. In all cases, the classroom teacher or another school official was present.

Instrumentation

For the study, Grade 4 and Grade 8 students were assessed on their understanding of science concepts and their reading comprehension. The science tests incorporated a variety of multiple-choice and open-ended questions on earth, physical, and life science concepts that Grades 4 and 8 students are expected to have been taught in the first half of the school year. The reading tests focused on assessing expository ability and narrative understanding through a variety of multiple-choice and open-ended questions. Students wrote their responses in the test booklets.

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*. The final selection was based on advice received from Grade 4 and Grade 8 science teachers. The science teachers evaluated the item language and 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 if they measured more recall than understanding, reasoning or investigation. Tables 4 and 5 summarize the

Table 4

Grade 4 Test Booklets Administered: Item Summary

	No. of items	No. of multiple-choice	No. of open-ended
Science	20	13	7
Reading	20	15	5
Background questionnaire	15	15	0
Accommodation questionnaire	7	7	0

Table 5

Grade 8 Test Booklets Administered: Item Summary

	No. of items	No. of multiple-choice	No. of open-ended
Science	30	25	5
Reading	19	13	6
Background questionnaire	15	15	0
Accommodation questionnaire	7	7	0

numbers and kinds of items (multiple-choice and open-ended) for the different instruments used for students in Grades 4 and 8, respectively. A detailed description of the instruments is provided next.

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 used a variety of open-ended and multiple-choice questions from the NAEP Grade 4 and Grade 8 Science assessments and a few TIMSS multiple-choice items. Students were assessed on their ability to demonstrate understanding of physical, earth, and life science concepts.

Grade 4 science test. In this test, questions 1-8 were life science questions taken from the 1996 NAEP test. Students were given 30 minutes to complete this section of multiple-choice and open-ended questions. Section 2 merged the 1996 NAEP with some TIMSS items. All the items in Section 2 were multiple-choice questions. Students were then given 15 minutes to answer 12 multiple-choice questions in life, earth, and physical sciences. Of the 20 items in the Grade 4 science test, 7 were open-ended and 13 were multiple-choice.

Grade 8 science test. The Grade 8 science assessment asked a total of 30 multiple-choice and open-ended questions in order to assess understanding of various life, earth, and physical science concepts. Students were given 45 minutes to complete the test, which incorporated 25 multiple-choice and 5 open-ended questions. Two versions of the Grade 8 science test were created (Booklets A and B). A third booklet was created when Booklet A was linguistically modified. The questions in this test come from the 1988, 1990, and 1996 NAEP tests and TIMSS.

Reading proficiency tests. English reading efficiency / proficiency tests were built from one intact block of the 1994 NAEP standardized reading assessment, the Fluency section of the LAS, and an experimental word recognition tool. One class period was allocated to the reading assessment. The Grade 4 reading passage was followed by 5 multiple-choice and 5 open-ended items. The Grade 8 passage was followed by 3 multiple-choice and 6 open-ended items.

Student background questionnaire. The study included a student background questionnaire, used to determine whether a student's background affected his or her performance on the tests. The questionnaire included items pertaining to students' language background, such as country of origin, length of time in the United States, and language other than English spoken in the home. It also asked students to self-assess their English and home language proficiency. The questionnaire included items selected from both the 1996 NAEP assessment and an earlier CRESST language background study.

Accommodation follow-up questionnaire. Students were asked to respond to an accommodation questionnaire in order to determine whether the accommodation (if any) helped them during the test and how the language in the test could have been made easier to understand.

Teacher and school questionnaires. The teacher questionnaire included items regarding the teachers' educational background and experience, as well as a section on when and how accommodation is used in their classroom(s). The school questionnaire contained items about the school population and its science and ESL resources.

Accommodation Instruments

Instruments used in our accommodations were a Customized English Dictionary, an English-to-English Glossary, an English-to-Spanish Glossary, a test version with Linguistic Modification of items, a Word List to accompany the

Linguistic Modification version, and a non-accommodated or Standard condition test version.

The Glossary and Customized Dictionary accommodations consisted of several sheets of paper stapled together and printed with columns of glossary or dictionary entries. The Word Lists had a similar look, but had no definitions or glosses. We created nondistinctive looking accommodations in order to reduce the stigma of using language aids during the test (a self-consciousness we inferred from the pilot observations of students who had received accommodations in book form).

To compile the Glossaries, Customized Dictionaries, and Word Lists, non-science test words were selected for their potential difficulty and arranged in alphabetical order. They were often the same words that had been modified in the Linguistic Modification version of the test. In addition to the non-science lexicon, the Word Lists contained science words used in the tests.

There was a bubble to the left of each entry. Students were asked to fill in the bubble next to any word that they looked up or—in the case of the word list—that they did not understand. To introduce them to the accommodations and the bubbles, all students were asked to look up a dummy entry (the word *nucleus*) and fill in the bubble next to it.

Following are more detailed descriptions of each booklet.

Customized English Dictionary. The Customized English Dictionary was created by compiling actual entries from the *Merriam-Webster Intermediate Dictionary* (1998 edition). All parts of an entry were included. Science terms were not included. First, this accommodation was tested by a focus group against a customized English dictionary created from a dictionary written for ELL students, the *Longman Dictionary of American English* (1983 paper edition). However, the Webster edition was the accommodation preferred by ELL students, possibly because the Longman entries included examples and thus were longer to read. Students were asked to fill in the bubble next to any word that they looked up. The Grade 4 version contained 61 entries, and the Grade 8 version contained 122 entries.

English-to-English Glossary. This glossary was compiled by assembling the non-science content terms in the test and the words that replaced them in the Linguistic Modification version of the test (see below). Science terms were not glossed. Students were asked to fill in the bubble next to any word that they looked

up. The Grade 4 version contained 63 entries, and the Grade 8 version contained 125 entries.

English-to-Spanish Glossary. This bilingual glossary was compiled by translating non-science words contained in the test into the appropriate Spanish form. Two bilingual (Spanish) educators examined the translations against the context of the test items. Science terms were not glossed. (Glossaries were not created for the Asian ELL students because in our pilot study, we found that significant numbers of Asian ELL students were not literate enough in their home language.) Students were asked to fill in the bubble next to any word that they looked up. The Grade 4 version contained 63 entries, and the Grade 8 version contained 125 entries.

When the two glossary accommodations are discussed, they will be referred to as the Bilingual/English Glossary.

Linguistic Modification of science items. A linguistically modified (Linguistic Modification) version of each science test was 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.

In order to do this, we first reviewed prior research on the effect of linguistic complexity on ELL student performance in content area assessment. 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 test items. As a result, the potentially challenging linguistic features were removed, reduced, or recast. Scientific vocabulary and concepts were preserved; only nontechnical vocabulary was changed.

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

An example of an original Grade 8 science item and its modified version is presented below.

Original Version

If the locations of earthquakes over the past ten years were plotted on a world map, which of the following would be observed?

- Earthquakes occur with the same frequency everywhere on the Earth.
- Earthquakes generally occur along the edges of tectonic plates.
- Earthquakes most frequently occur near the middle of continents.
- Earthquakes do not seem to occur in any consistent pattern.

Modified Version

Sue drew the locations of earthquakes in the past ten years on a world map. She saw that:

- Earthquakes happen just as often everywhere on the Earth.
- Earthquakes usually happen along the edges of tectonic plates.
- Earthquakes most often happen near the middle of continents.
- Earthquakes do not seem to happen in any consistent pattern.

The changes are outlined here:

The Prompt

- A conditional clause with passive voice "If . . . were plotted . . . map," changed to a statement in active voice ("Sue drew . . . on a map.").
- Technical term "plotted" changed to "drew."
- Main clause with prepositional phrase and verb in passive voice ("which of the following would be observed?") changed to short statement plus a relative pronoun ("She saw that:") that the answer choices complete.

The Answer Choices

- "Occur" changed to "happen."
- The phrase "with the same frequency" changed to "just as often."
- Less common adverbs changed to more common ones: "generally" changed to "usually" and "most frequently" to "most often."

See Table A1 in Appendix A for a list of commonly revised linguistic features.

Word List. Students who received the Linguistic Modification version or the Standard condition of the science test also received a Word List of potentially unfamiliar words in the test. The Grade 4 version contained 91 words, and the Grade 8 version contained 198 words. Approximately one third of the words were science words. The words were not glossed or defined. The students were instructed to fill in the bubble next to words they did not understand.

Standard condition. The students who were assigned the Standard condition took the science test for their grade level with the original wording. They also received a Word List, as described above.

Design and Procedure

To investigate our hypotheses concerning the use of accommodation, we tested ELL and non-ELL students in Grades 4 and 8. Science tests containing 20 to 30 NAEP/TIMSS items were administered under four conditions. The Standard condition contained original items with no accommodation. The remaining three conditions included accommodations that focused on the potential challenges of understanding the English language vocabulary and syntax: a Customized English Dictionary, a Bilingual/English Glossary (English-to-Spanish or English-to-English), or a Linguistic Modification version of the test. As the tests under the four conditions were administered at the same time, all students had the same amount of time to complete the test sections. In addition to the science test, a reading assessment and two questionnaires were administered. The study design offered several points of comparison between ELL and non-ELL students, with and without accommodated testing.

Additionally, we created two booklets (A and B) of each test for Grade 8 in order to vary the sequence of test items and discourage cheating. However, there was only one type of booklet for each Linguistic Modification version of the science test. Each had the same item order as booklet A of the non-modified test.

We provided students in the control group with the same amount of time that the groups with accommodations received. However, they were asked to identify any unfamiliar words on the Word List used in the test, an additional task that is only tangential to the assessment. The students who had the Linguistic Modification version of the test also received the Word List so that they would neither feel left out nor be without something to do while waiting for others to finish the test.

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. The schools provided rosters of the participating classes ahead of time. These were examined closely to determine whether the class indeed contained enough ELL students with the specified home language. After the students' information was entered into a

database, students were sorted as ELL and non-ELL participants. The ELL students belonging to the specified home language were noted. When Spanish was the target home language, a specified number of Spanish-speaking ELL students were randomly assigned the English-to-Spanish Glossary condition; then a similar number of non-ELL students were assigned the English-to-English Glossary condition. Other accommodations or no accommodations were assigned randomly among the remaining ELL and non-ELL parts of the class (see Figure 1). Figure 2 is a model for accommodation distribution for classes in which an Asian language was the target home language. For these classes, there was no glossary accommodation, since the pilot study revealed that many of the Chinese and Korean ELL students were not literate enough in their home language.

Administration of tests and questionnaires. There were two testing sessions per class, scheduled in the morning whenever possible, on two consecutive days. Generally, at the beginning of Day 1, the science test was administered, then the accommodation follow-up questionnaire and the background questionnaire. Before the Grade 4 testing, test directions were read aloud. Students at both tested grade levels reviewed sample questions in both multiple-choice and open-ended formats.

The accommodations were distributed randomly among the ELL and non-ELL students. Both ELL and non-ELL groups 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 and used. There were scripts for each grade level and each day of testing. Test administrators were asked to observe the students, answer 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 the reading test instructions and questionnaire instructions were read aloud to Grade 8 students. To ensure accurate responses to the questionnaires, students with questionable or confusing responses were asked to clarify or correct them, often on Day 2.

On Day 2, the reading assessment was administered. Directions were read aloud to each class, but no other accommodation was made.

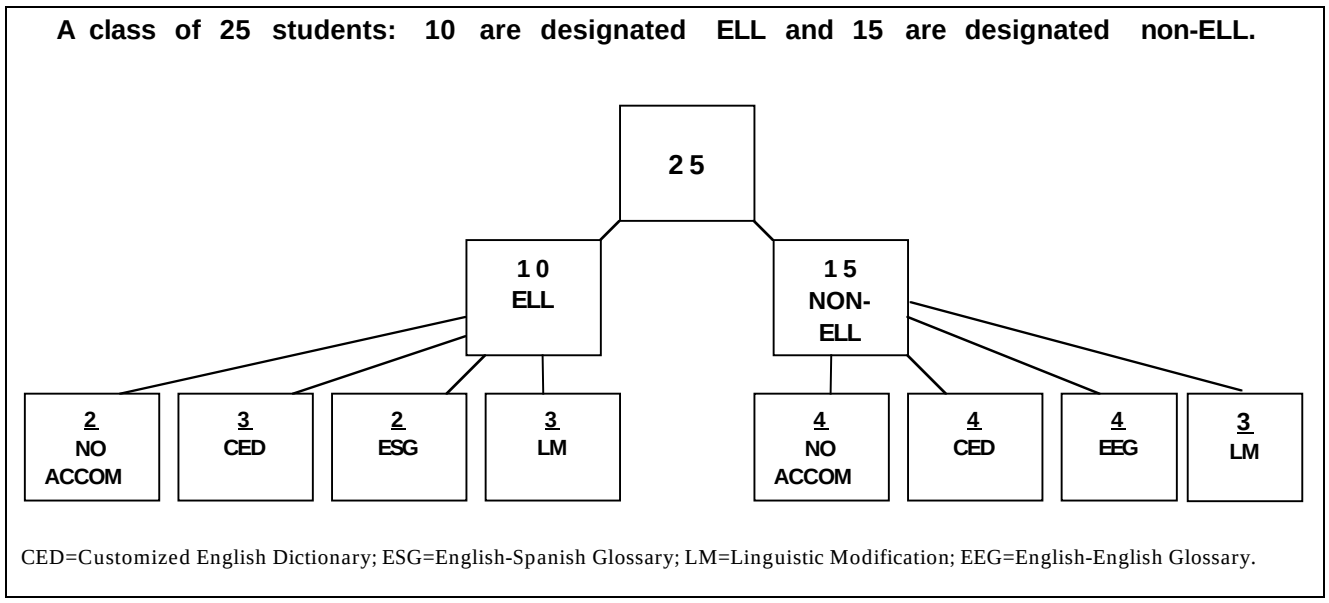

Figure 1. Example of accommodation distribution of three possible accommodations and no accommodation (where Spanish was the target home language).

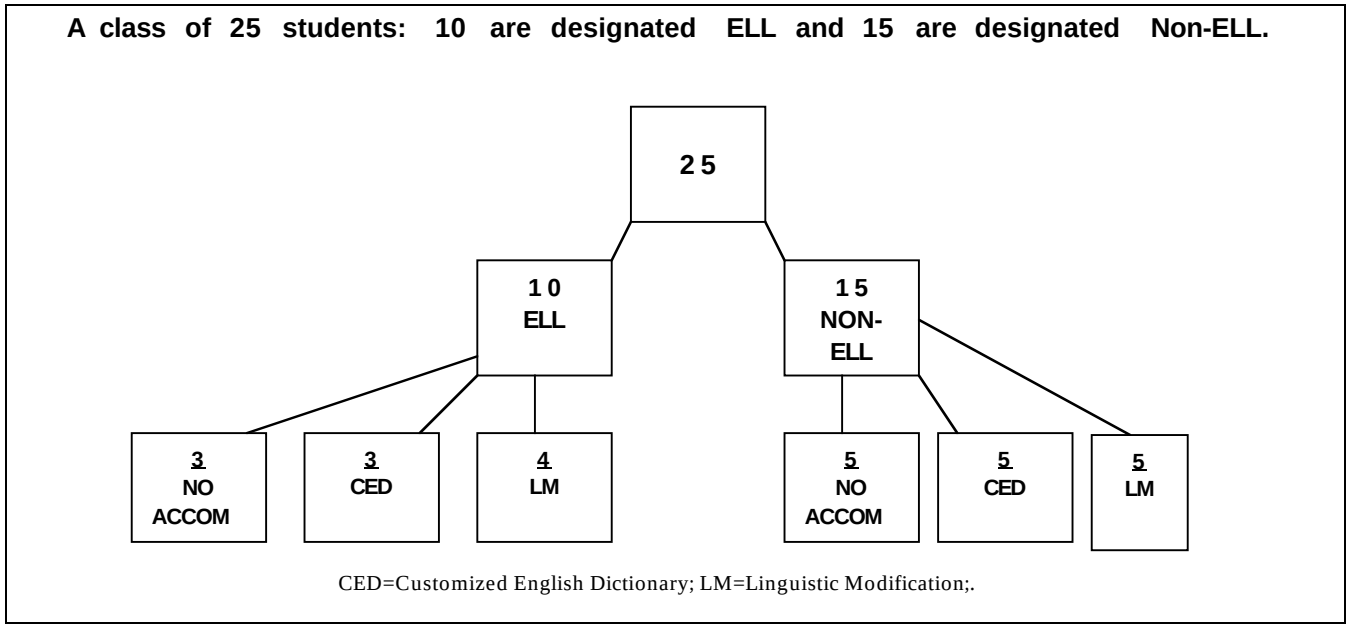

Figure 2. Example of accommodation distribution of two possible accommodations and no accommodation (where an Asian language was the target home language).

Test administration personnel. Test administrators for the study were identified locally and consisted of graduate student researchers and district tutors. Each degreed test administrator was assisted by fellow researcher/tutor or by a degreed test administration contractor. All were trained by the project coordinator 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 traveling to testing sites.

Rating of Open-Ended Test Items

Most of the open-ended test items were scored by two raters who were trained by the project staff.

Initially, open-ended science items were scored by classroom teachers (Grade 4 teachers or middle school language arts teachers, depending on the students' grade level). Scorers were trained in the use of the NAEP rubric and debriefed after each scoring session. However, after having interrater reliability issues, we recruited district curriculum specialists in language arts and science to be third raters.

When this did not help, we conferred with NAEP scoring experts about rating procedures. Based on their advice, we scored one open-ended question at a time rather than scoring a set of questions in a given session. As a trial, two project staff members rated the open-ended item with the lowest interrater reliability coefficient. By focusing on a single item at a time and by discussing the scoring procedure, they increased the percent of agreement from 69% for the first set to almost 100% for the second, different set. The project staff integrated this experience into the training of the re-scoring team.

Due to scheduling difficulties and classroom grading style influences, we decided not to use K-12 classroom teachers. Instead we screened and trained graduate students who had little or no K-12 classroom grading experience, and we found better interrater reliability. Only one team had difficulty agreeing on the interpretation of a complex rubric for a Grade 8 reading item. A third rater also scored that item. These scores were hand-entered and proofread.

Results

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

- Which test accommodations are more effective in reducing the performance gap between ELL and non-ELL students? (Effectiveness)

- Do the accommodations impact the constructs under measurement, i.e., the content of the science test? (Validity)

- Is the outcome of accommodated assessment dependent on student background characteristics? (Differential impact)

Null Hypotheses

The null hypotheses related to the research questions above are these:

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

$\mathbf{H}_{02}$: Accommodation does not impact performance of non-ELL students on science tests. (Validity)

$\mathbf{H}_{03}$: Student background variables do not impact performance on the accommodated science assessments. (Differential impact)

Alternative Hypotheses

The alternative hypotheses corresponding to the null hypotheses above are these:

$\mathbf{H}_{11}$: Some forms of accommodation are more *effective* than others in reducing the science performance gap between ELL and non-ELL students.

$\mathbf{H}_{12}$: Accommodations *do* 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.

$\mathbf{H}_{13}$: Student background variables *do* 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 hypotheses concerning the use of accommodation on students' performance in science, ELL and non-ELL students were tested under four testing conditions. The simplest was the administration of the original NAEP Science items and a few TIMSS multiple-choice items with no language accommodation.[4] Other students received language accommodation in one of three forms: a Customized English Dictionary, an English-to-Spanish Glossary (or English-to-English Glossary[5]), and a Linguistic Modification version of the test items.[6] These testing conditions will be referred to below as the *Standard* condition, the *Customized Dictionary* condition, the *Bilingual/English Glossary*, and the *Linguistic Modification* version. An accommodation was randomly assigned to ELL and non-ELL students within each classroom. Thus, eight comparison groups were possible: 4 levels of accommodation, by 2 levels of ELL status. Table 6 illustrates the design and presents the numbers of students in each cell of the design for Grade 4. Table 7 presents this information for Grade 8.[7]

In this study, we were interested in examining the impact of different types of accommodations and students' ELL status on their performance in science. That is, the two independent variables that were hypothesized to impact the outcome of the

Table 6

Grade 4 Design and Sample Size by Accommodation and ELL Status

	ELL status		
Accommodation	ELL	Non-ELL	Total
Standard condition	*N* = 241	*N* = 268	*N* = 509
Customized Dictionary	*N* = 247	*N* = 270	*N* = 517
Bilingual/English Glossary	*N* = 101	*N* = 135	*N* = 236
Linguistic Modification	*N* = 257	*N* = 284	*N* = 541
Total	*N* = 846	*N* = 957	*N* = 1803

[4]The only difference between the standard condition in NAEP and in our study is that we gave additional time (50%).

[5]Since there was no practical reason to give a bilingual glossary to a portion of non-ELL students, we gave this group an English-to-English glossary.

[6]All students were given the same amount of extra time on the science assessment.

[7]The numbers in each table represent only those participants whose responses were included in the analysis. A small number of students were excluded from the study because they were completely non-English speaking, were enrolled in a different grade, or were administered an inappropriate accommodation.

Table 7

Grade 8 Design and Sample Size by Accommodation and ELL Status

Accommodation	ELL status		Total
	ELL	Non-ELL	
Standard condition	*N* = 241	*N* = 206	*N* = 447
Customized Dictionary	*N* = 129	*N* = 119	*N* = 248
Bilingual/English Glossary	*N* = 241	*N* = 209	*N* = 450
Linguistic Modification	*N* = 245	*N* = 199	*N* = 444
Total	*N* = 856	*N* = 733	*N* = 1589

science assessment were the type of accommodation and student's ELL status. Examining the main effect of the type of accommodation determines whether the accommodation strategies used in this study have any significant impact on the outcome of assessment (science test score). Testing the main effect of student's ELL status provides information on the performance difference between ELL and non-ELL students. Testing the interaction between the type of accommodation and student's ELL status will provide information about two of the main hypotheses of this study (effectiveness and validity).

Reading efficiency/proficiency was used as a covariate in this study. Thus, a two-factor analysis of covariance was deemed suitable for analyzing the results of the study. However, since we were interested in testing particular hypotheses, we conducted a series of *a priori* or planned tests. Instead of using a two-way model to test effectiveness and validity, we used a different one-way ANCOVA for testing each. To test the effectiveness hypothesis, we compared student performance under each accommodation with the Standard condition. For testing the validity hypothesis, we conducted planned tests to compare accommodated and non-accommodated outcomes.

As can be seen from the data in Tables 6 and 7, the numbers of participants in the cells that were created by crossing the type of accommodation with student's ELL status were quite large. In Grade 4, we tested 1,803 students. The cell sizes ranged from a minimum of 101 (ELL students using the English-to-Spanish glossary) to a maximum of 284 (non-ELL students using the Linguistic Modification version). For Grade 8, we tested 1,589 students. The cell sizes ranged from a

minimum of 119 (non-ELL students using the Customized Dictionary) to a maximum of 245 (ELL students taking the Linguistic Modification version). These large cell sizes increase the power of analyses and enabled us to include different background variables in the analyses.

Treating the Missing Data

In large-scale data collection, the problem of missing data is inevitable, particularly when multiple testing sessions of the same group of students occur on different days. This was exactly the case for our study. We collected data from a large number of schools in different locations in the United States, usually testing each class on two consecutive days. The missing data in this study were due mainly to absences on one or the other of the two testing days. As indicated earlier, students responded to the science test, the background questionnaire, and the accommodation follow-up questionnaire in the first session (Day 1) and to the reading subscale items in the second session (Day 2). Some students were absent for the first session (science), the second session (reading), or both. We first tried the simple mean replacement approach by subgroup. Variables that had significant impact on the science and reading scores were used for grouping students. These variables included students' ELL and Title I status, parent education, participation in the free/reduced price lunch program, and type and amount of a language other than English spoken in the home. However, some of these variables did not have enough valid responses (high missing data rates); therefore, they were not used for grouping students. Among the background variables that had significant impact on test scores, students' ELL status and their participation in the free/reduced lunch program had higher response rates, and these were used for grouping students.

Mean scores for the subscales of the science and reading tests were computed and used to replace the missing data. A comparison of some of the analyses performed on the files with and without the missing data replacement revealed major differences in the trend of data between the two kinds of files. These major differences suggested that the missing data replacement procedure using the subgroup means changed the structure of the data to some extent. Therefore, we decided to use a more robust technique for replacing missing data.

We used a regression approach for missing data replacement. A standard residual averaged from five random drawings was added to each imputed value. Since there were three different cases of missing data (i.e., missing science score,

missing reading score, or both), we created three multiple regression equations. We used ELL status and class averages for the reading and science scores as the predictors.

After replacing the missing data through this regression approach, analyses were performed to compare the data for the file in which the missing data had been replaced with the data for the file containing only records of students who had complete data. The results showed consistency in the trend of analyses, which suggested that the regression approach in replacing missing data did not change the overall trends.

Outcome Variables

The focus of this study was the impact of accommodation on performance in science. Therefore, a measure of science content knowledge was used to provide data on the outcome or dependent variable. In addition to a science test, a measure of reading efficiency / proficiency was used as a covariate.

In the pilot phase of this study, the English reading efficiency / proficiency measure consisted of two intact blocks of the 1994 NAEP Reading assessment. These two blocks were used to measure the reading ability of both ELL and non-ELL students. The NAEP Reading items were not intended for measuring ELL students' reading ability, and as the results of pilot study indicated, the test was difficult for these students.

Thus, to present a more valid measure of reading for both ELL and non-ELL students, we decided to use multiple measures of reading proficiency with different levels of difficulty: (a) the Fluency subscale of the LAS and (b) both the multiple-choice reading comprehension items and the open-ended reading comprehension questions from one released block of NAEP.

The NAEP Reading block was selected from the 1994 NAEP main assessment in reading based on the results of our pilot phase of this study. The LAS Fluency subscale was selected, rather than the complete LAS test, since, as the results of our earlier study suggested, it had better discrimination power when compared to the other LAS subscales.

There were technical issues in creating a composite score from the three reading components that were analyzed in this study. The main concern was the possible differences in the construct that the three components measure. Other

problems included the differences in the psychometric characteristics, number of items, and the scale of the three components' scores. To bring the three components into the same scale, component raw scores were transformed to standard scores with a mean of 50 and standard deviation of 10. A mean of the three standard score components was then computed and used as a reading efficiency / proficiency score. However, converting the raw scores into standard scores did not solve all the problems, specifically the problem of the three components measuring different constructs and problems related to the differences in the psychometrics of the three different measures.

Since each of the three components was supposed to measure different aspects of students' reading proficiency, we created a latent composite consisting of the common variance across the three components. A simple structure confirmatory factor analysis was used to create a latent composite of the two parts of the NAEP reading block and the LAS Fluency subscale.

Scoring Science and Reading Tests

Since the science and reading tests in the main study included multiple measures, scoring these tests required a more complex procedure. The reading efficiency / proficiency measure in this study consisted of three components. Even though these three subscales measured reading, each measured different aspects of reading, and their scale scores differed. To capture the common variance among the different subscales in each test, a latent composite of the subscale score for each test was created. Figure 3 shows the structural model for the latent composites for Grade 4, and Figure 4 presents a similar model for Grade 8.

As Figures 3 and 4 show, a science latent composite was created from two components, the multiple-choice items and the open-ended items.

To compare the performance of ELL students with that of non-ELL students, factor scores of the latent composites for reading and science were generated. Factor scores, rather than raw scores were used in computation and also for reporting. The factor scores had a mean of zero and standard deviation of .13 for science and a mean of zero and standard deviation of .16 for reading. Based on these means and standard deviations, factor scores were transformed to a T scale with a mean of 50 and standard deviation of 10. This transformation was made to facilitate the reporting of scores.

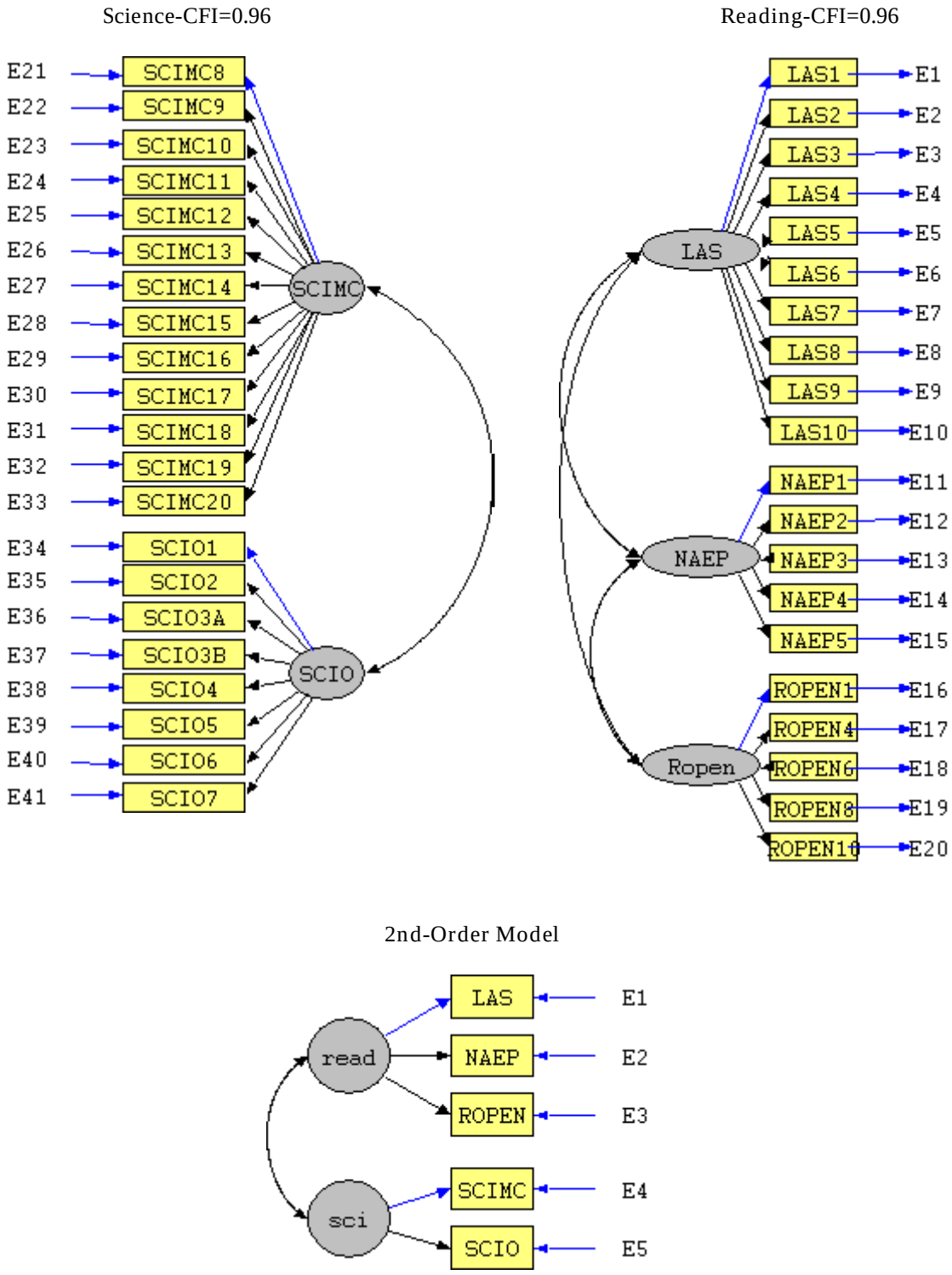

Figure 3. Latent variable models for Grade 4 test results.

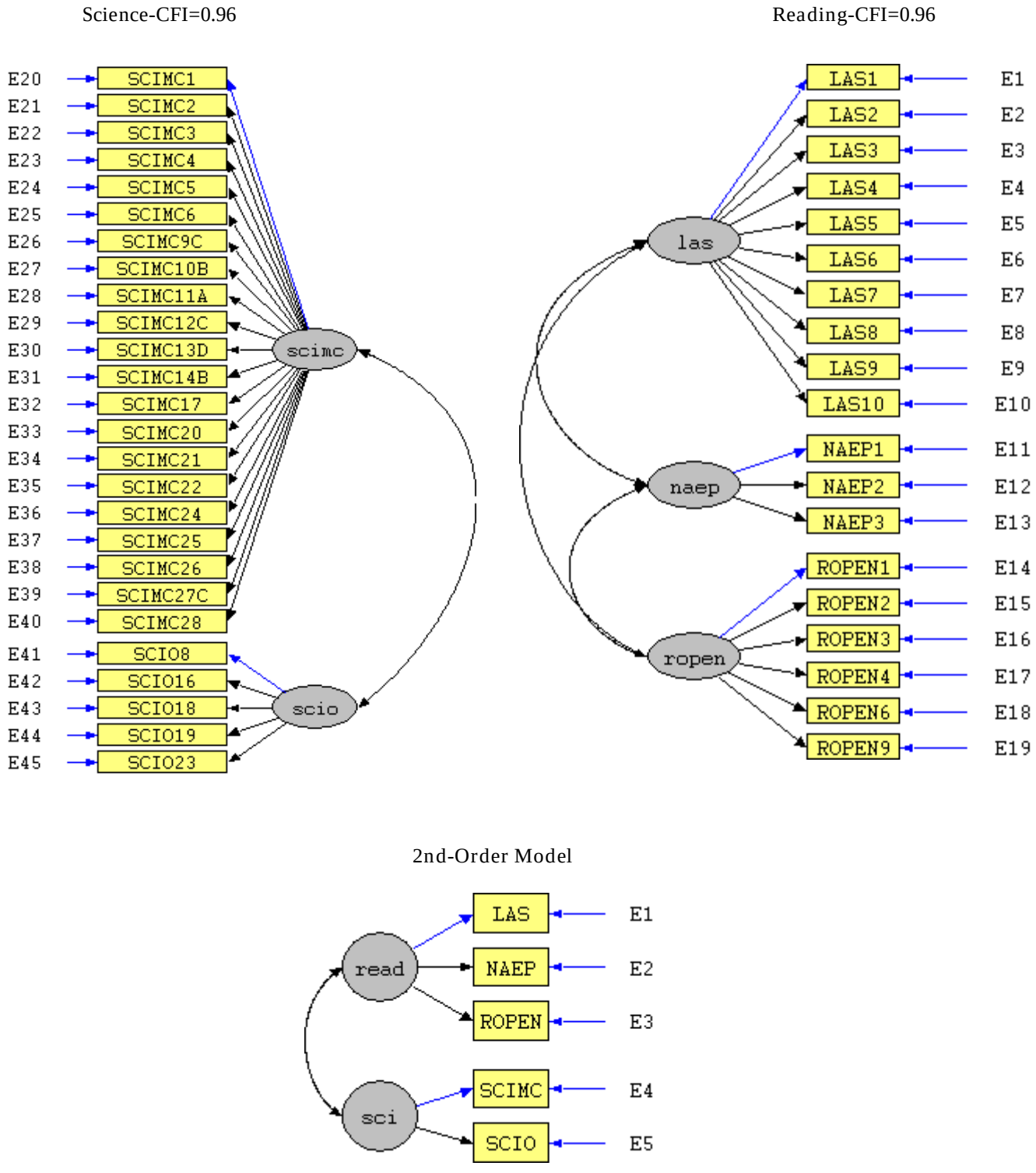

Figure 4. Latent variable models for Grade 8 test results.

Figures 3 and 4 also show a structural model for correlating the two latent composites, the reading latent scores and the science latent scores. By creating latent scores and correlating these scores, the correlation between reading and science was improved. For example, the correlation between the simple composites (between standard scores of the components) of science and reading was .68 for Grade 4 as compared to a correlation of .72 between the latent composites. For Grade 8, the correlation between the simple composites was .61 as compared to a correlation of .68 between the latent composites.

Analyses of Open-Ended Questions

As indicated earlier, most open-ended science and reading items were scored independently by two raters. Interrater reliability indices (percent of exact and within one-point agreement, P.M. 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 by the first (Phase 1) raters.

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

Table 8

Grade 4 Interrater Reliability for Open-Ended Science Items—Phase 1

Item No.	Rater combinations	No. of students	Kappa	Alpha	Agreement
1	1, 2	1212	.94	.98	97%
2	1, 2	1276	.54	.78	73%
3a	1, 2	1197	.71	.85	84%
3b	1, 2	1191	.54	.73	76%
4	1, 2	1178	.79	.92	88%
5	1, 2	1231	.43	.65	69%
6	1, 2	1001	.56	.79	73%
7	1, 2	870	.42	.59	80%

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

For the 8 open-ended science items in Grade 4, the percent of agreement ranged from a low of 69% (for item 5) to a high of 97% (for item 1). Kappa coefficients ranged from a low of .42 (for item 7) to a high of .94 (for item 1). Alpha coefficients ranged from a low of .59 (for item 7) to a high of .98 (for item 1). Looking at a combination of interrater reliability coefficients in Table 8, it is apparent that some of the open-ended items were more difficult to score than others; thus, they suffer from lower interrater reliability. For example, items 2, 3b, 5, 6 and 7 have lower kappa coefficients. These items may impact the overall test reliability and even the overall validity of the science scores.

To increase the interrater reliability of these items, we decided to re-score items with lower-than-average interrater reliability. To put together a re-scoring team, we screened a large group of graduate students who were interested in working for two full weeks, seeking out those with the ability to score a set of sample test responses strictly according to the NAEP rubric and who had few pre-conceived ideas about rating K-12 writing. Their trial scores provided data for computing inter- and intra-rater reliabilities. Applicants with higher inter- and intra-rater reliabilities were selected for the re-scoring team.

Table 9 presents interrater reliability data for science open-ended questions by the re-scoring (Phase 2) team. Comparing the re-scoring session's interrater reliability statistics in Table 9 with the data in Table 8 (original scoring) reveals a major improvement in the interrater reliability indices. For example, the kappa coefficient for item 2 improved from the original .54 to .76. For item 3a, the kappa increased from .71 to .97, and for item 5, the kappa increased from .43 to .92. Similar trends of increase can be seen for all items and for the three interrater statistics.

Table 10 shows interrater reliability indices for reading for Grade 4 students from the original (Phase 1) scoring sessions. Again, some of the questions have poor interrater reliability statistics. For example, only item 6 of this group of items has a relatively high kappa coefficient (.82). The kappa coefficient was .40 for items 1 and 4, .61 for item 8, and .51 for item 10. In our discussion of interrater reliability, we focus on the kappa coefficient because it is a more robust index of interrater reliability (see Abedi, 1996).

Table 9

Grade 4 Interrater Reliability for Open-Ended Science Items—Phase 2

Item No.	Rater combinations	No. of students	Kappa	Alpha	Agreement
1	1, 2	179	1.00	1.00	100%
2	1, 2	1657	.76	.90	85%
3a	1, 2	181	.97	.99	98%
3b	1, 2	1545	.68	.81	97%
4	1, 2	169	1.00	1.00	100%
5	1, 2	1620	.92	.96	96%
6	1, 2	171	1.00	1.00	100%
7	1, 2	165	.97	.99	99%

Table 10

Grade 4 Interrater Reliability for Open-Ended Reading Items—Phase 1

Item No.	Rater combinations	No. of students	Kappa	Alpha	Agreement
1	1, 2	1275	.40	.58	71%
4	1, 2	1225	.40	.75	63%
6	1, 2	1059	.82	.90	96%
8	1, 2	1004	.61	.76	80%
10	1, 2	968	.51	.68	76%

We also re-scored the reading items. Table 11 shows the interrater reliability indices by the re-scoring team. Again, comparison of the interrater reliabilities of the two phases of scoring suggests major improvements with the Phase 2 scoring. For example, for item 1, the kappa coefficient increased from .40 to .96. For item 2, it increased from .40 to .86, and for item 10, the kappa increased from .51 to .92.

Table 12 summarizes the results of interrater reliability analyses for Grade 8 science open-ended items. Similar to the interrater reliability results that were presented earlier, the results for Grade 8 science suggest relatively low interrater indices for some of the science open-ended items. For example, item 19 had a kappa of .46, well below the normally acceptable range. By re-scoring this item, we increased the kappa from .46 to .96 (see Table 13).

Table 11

Grade 4 Interrater Reliability for Open-Ended Reading Items—Phase 2

Item No.	Rater combinations	No. of students	Kappa	Alpha	Agreement
1	1, 2	1647	.96	.98	98%
4	1, 2	903	.86	.95	92%
6	1, 2	160	1.00	1.00	100%
8	1, 2	150	.99	.99	99%
10	1, 2	141	.92	.96	96%

Table 12

Grade 8 Interrater Reliability for Open-Ended Science Items—Phase 1

Item No.	Rater combinations	No. of students	Kappa	Alpha	Agreement
8	1, 2	369	.78	.89	94%
16	1, 2	325	.82	.89	94%
18	1, 2	241	.58	.80	93%
19	1, 2	243	.46	.58	88%
23	1, 2	293	.80	.89	99%

Table 13

Grade 8 Interrater Reliability for Open-Ended Science Items—Phase 2

Item No.	Rater combinations	No. of students	Kappa	Alpha	Agreement
8	1, 2	86	.83	.93	95%
16	1, 2	84	.85	.93	93%
18	1, 2	61	.97	.99	98%
19	1, 2	54	.96	.98	98%
23	1, 2	65	1.00	1.00	100%

Similarly, Table 14 presents interrater reliability coefficients for Grade 8 reading open-ended questions. As the data in Table 14 suggest, most of the open-ended reading items in Grade 8 had very low interrater reliability coefficients. For example, the maximum kappa coefficient for this set of items was .43 (item 1), and one test item had a kappa coefficient of .20 (item 4).

Table 14

Grade 8 Interrater Reliability for Open-Ended Reading Items—Phase 1

Item No.	Rater combinations	No. of students	Kappa	Alpha	Agreement
1	1, 2	301	.22	.48	54%
2	1, 2	293	.37	.61	68%
3	1, 2	263	.43	.76	69%
4	1, 2	245	.20	.46	50%
6	1, 2	236	.41	.76	64%
9	1, 2	184	.40	.72	80%

Table 15 summarizes the results of interrater reliability analyses from the re-scoring team. Comparing the re-scoring interrater reliability indices with the original indices, again there were major improvements for all items. For example, the kappa coefficient for item 1 increased from .22 to .86, and for item 4 from .20 to .80.

Examining the Internal Consistency of Science and Reading Tests

In classical test theory, if all the items on a test measure a single underlying construct, the test is considered to be unidimensional. In such a case, the items should exhibit high internal consistency. We tested the internal consistency of the reading and science tests by computing Cronbach's coefficient alpha separately for the multiple-choice and open-ended items for both tests. The alpha coefficient was also computed for the overall science and reading tests (multiple-choice plus open-ended items).

Table 15

Grade 8 Interrater Reliability for Open-Ended Reading Items—Phase 2

Item No.	Rater combinations	No. of students	Kappa	Alpha	Agreement
1	1, 2	732	.86	.92	93%
2	1, 2	625	.74	.87	87%
3	1, 2	453	.74	.91	83%
4	1, 2	716	.80	.93	90%
6	1, 2	377	.69	.91	80%
9	1, 2	258	.61	.88	78%

Table 16 presents the internal consistency results for the reading and science tests for Grade 8. The internal consistency coefficient for all Grade 8 reading items was .78. The multiple-choice reading test items had a higher alpha coefficient (.73) than the open-ended items (.56). The internal consistency coefficient for the science test (.69) was lower than the coefficient for the reading test (.78). The internal consistency coefficient was higher for the science multiple-choice items (.68) than for the science open-ended items (.45). This low internal consistency coefficient suggests that the science test may be multi-dimensional. Four multiple-choice science items showed low item-to-total correlation.[8] When those items were removed from the analysis, the alpha improved to .75. Therefore, the analysis for the Grade 8 science test was based on 26 science items.

Table 17 presents the internal consistency results for the Grade 4 reading and science tests. The internal consistency coefficients for the Grade 4 science and reading tests were relatively higher than the coefficients for the Grade 8. The overall internal consistency coefficient for reading was .82. As was the case for the Grade 8 reading tests, the coefficient was higher for the multiple-choice (.79) items than for the open-ended items (.60). For the science test, the overall alpha was .71.

Table 16

Grade 8 Internal Consistency Coefficients for Reading and Science Tests

Test / subscale	No. of items	No. of students	Alpha
Reading test			
Multiple-choice	13	1362	.73
Open-ended	6	1374	.56
Total reading	19	1362	.78
Science test			
Multiple-choice	25	1400	.68
Open-ended	5	1395	.45
All items	30	1390	.69
Total science[a]	26	1391	.75

[a]Multiple-choice items 7, 15, 29, and 30 were removed due to poor reliability.

[8]Two of the four multiple-choice science items with low item-to-total correlation have the lowest *p* value of any multiple-choice items in the NAEP 1996 block. Two were at the very end of our test and so were not reached by as many students. Three were physical science items concerning heat and cold. The fourth concerned a lunar eclipse.

Table 17

Grade 4 Internal Consistency Coefficients for Reading and Science Tests

Test / subscale	No. of items	No. of students	Alpha
Reading test			
Multiple-choice	15	1648	.79
Open-ended	5	1675	.60
Total reading	20	1643	.82
Science test			
Multiple-choice	13	1667	.58
Open-ended	8	1667	.64
Total science	21	1667	.71

Testing Hypotheses Concerning Effectiveness and Validity of Accommodation

There were two major research questions in this study:

- How valid are the results of accommodated assessments?
- How effective in reducing the performance gap between ELL and non-ELL students are the accommodation strategies that are used in this study?

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

To test the effectiveness hypothesis, we compared the performance of ELL students who were provided an accommodation in science with the performance of ELL students who were tested under the Standard condition. A significantly higher performance under any accommodation in this study would indicate effectiveness of that particular accommodation.

Results for Grade 4 Students

Table 18 presents descriptive statistics (T-means, standard deviation, and number of students) for each type of accommodation and by ELL subgroups in the Grade 4 classes.

As described earlier, we used latent scores instead of the simple total item scores because we used multiple measures in reading and multiple scales in science.

Table 18

Grade 4 Mean Latent Science Achievement Scores (M = 50, SD = 10)

Accommodation	ELL status		Row total (ELL + non-ELL)
	ELL	Non-ELL	
Standard condition	48.23 (SD = 9.38; n = 268)	52.74 (SD = 9.29; n = 241)	50.36 (SD = 9.60; n = 509)
Customized Dictionary	48.37 (SD = 9.75; n = 270)	52.81 (SD = 10.23; n = 247)	50.49 (SD = 10.22; n = 517)
Bilingual / English Glossary	45.62 (SD = 8.19; n = 135)	52.46 (SD = 9.75; n = 101)	48.55 (SD = 9.50; n = 236)
Linguistic Modification	47.36 (SD = 9.48; n = 284)	52.54 (SD = 10.57; n = 257)	49.82 (SD = 10.33; n = 541)
Column total	47.64 (SD = 9.39; n = 957)	52.67 (SD = 10.00; n = 846)	50.00 (SD = 10.00; n = 1803)

A latent score for reading was computed and used as a covariate. A latent score for science was computed and used as the outcome variable. Latent scores were transformed to scale scores with a mean of 50 and standard deviation of 10.

As the row and column marginals show, ELL students in Grade 4 had lower science test scores (T-scores M = 47.64, SD = 9.39; raw scores M = 7.72, SD = 3.18; n = 957) than non-ELL students (T-scores M = 52.67, SD = 10.00; raw scores M = 9.37, SD = 3.30; n = 846). There were slight differences in the performance of both ELL and non-ELL students under different forms of accommodation. However, as we explain later, these differences did not reach statistical significance.

Comparing the performance of ELL students under accommodation with those under the Standard condition, ELL students scored slightly lower under some of the accommodations. For example, the T-mean science score for ELL students under the Customized Dictionary condition was 48.37 (SD = 9.75; raw scores M = 7.89, SD = 3.31; n = 270); under the Bilingual / English Glossary condition, the T-mean was 45.62 (SD = 8.19; raw scores M = 7.05, SD = 2.74; n = 135); and with the Linguistic Modification version, the T-mean was 47.36 (SD = 9.48; raw scores M = 7.70, SD = 3.19; n = 284) as compared to the T-mean of 48.23 (SD = 9.38; raw scores M = 7.93, SD = 3.21; n = 268) for ELL students under the Standard condition.

For non-ELL students in Grade 4, accommodation did not seem to make any difference. For students tested under the Customized Dictionary condition, the T-mean score was 52.81 (SD = 10.23; for raw scores M = 9.41, SD = 3.28; n = 247). For students under the Bilingual / English Glossary condition, the T-mean was 52.46 (SD

= 9.75; raw scores *M* = 9.17, *SD* = 3.26; *n* = 101) and under the Linguistic Modification condition, the T-mean was 52.54 (*SD* = 10.57; raw scores *M* = 9.46, *SD* = 3.52; *n* = 257), as compared to a T-mean of 52.74 (*SD* = 9.29; raw scores *M* = 9.31, *SD* = 3.10; *n* = 241) under the Standard condition.

Table 19 presents descriptive statistics for the latent reading scores across the ELL and accommodation categories for Grade 4 students. Similar to the science scores, the latent reading scores were transformed on a scale with a mean of 50 and standard deviation of 10. As the data in Table 19 show, consistent with the results of earlier studies, non-ELL students in Grade 4 obtained higher total reading scores (T-scores *M* = 52.50, *SD* = 10.16; raw scores *M* = 11.94, *SD* = 4.20; *n* = 846) than the ELL students (T-scores *M* = 47.79, *SD* = 9.32; raw scores *M* = 9.87, *SD* = 3.82; *n* = 957). The trend of higher reading scores for non-ELL students holds across the categories of accommodations. That is, under all four conditions, Grade 4 non-ELL students had higher mean reading scores than ELL students. However, there were small differences across the accommodation categories for both ELL and non-ELL groups. For example, the average reading score for ELL students who took the science test under all four conditions was 47.79 (*SD* = 9.32, *n* = 957). The mean reading score for ELL students was slightly higher under the Standard (*M* = 48.11, *SD* = 9.67; *n* = 268) and the Customized Dictionary (*M* = 48.11, *SD* = 8.99; *n* = 270) conditions. The mean was slightly lower under the Bilingual / English Glossary condition (*M* = 46.37, *SD* = 8.13; *n* = 135). These differences in students' reading proficiency happened despite randomization of conditions. To control for these initial differences in reading, we

Table 19

Grade 4 Mean Latent Reading Achievement Scores (*M* = 50, *SD* = 10)

Accommodation	ELL status		Row total (ELL + non-ELL)
	ELL	Non-ELL	
Standard condition	48.11 (*SD* = 9.67; *n* = 268)	53.21 (*SD* = 9.89; *n* = 241)	50.53 (*SD* = 10.10; *n* = 509)
Customized Dictionary	48.11 (*SD* = 8.99; *n* = 270)	52.73 (*SD* = 10.65; *n* = 247)	50.32 (*SD* = 10.08; *n* = 517)
Bilingual / English Glossary	46.37 (*SD* = 8.13; *n* = 135)	51.21 (*SD* = 9.59; *n* = 101)	48.44 (*SD* = 9.09; *n* = 236)
Linguistic Modification	47.87 (*SD* = 9.80; *n* = 284)	52.11 (*SD* = 10.13; *n* = 257)	49.88 (*SD* = 10.17; *n* = 541)
Column total	47.79 (*SD* = 9.32; *n* = 957)	52.50 (*SD* = 10.16; *n* = 846)	50.00 (*SD* = 10.00; *n* = 1803)

adjusted the science test scores by the students' reading scores and we compared accommodation outcomes based on the adjusted science scores. We also adjusted for Spanish being the home language (see below).

Effectiveness. To test the hypothesis concerning effectiveness of accommodation, we conducted a series of *a priori* or planned tests. The main reason for conducting *a priori* tests rather than ANOVA or *a posteriori* (post-hoc) tests is that we were not interested in any possible differences. We were only interested in comparisons of accommodated assessments with the Standard condition (no accommodation). The results of our previous studies (see, for example, Abedi, Hofstetter, Lord, et al., 1998) indicated that, in spite of our efforts to eliminate the initial differences in students' level of reading proficiency, students in some of the accommodation conditions had significantly higher reading scores. To control for such initial differences, we adjusted the mean science score by the reading score. The planned tests were conducted on the adjusted science scores.

It should be noted that the English-to-Spanish Glossary accommodated group was restricted to ELL students with Spanish as a home language; thus, it was necessary to control for Spanish being the home language.

Table 20 summarizes the results of our planned comparisons for testing the effectiveness of accommodations. We conducted three planned comparisons, one for each form of accommodation. In the first test, we compared the science mean score (that was adjusted by the reading score and for Spanish being the home language) under the Customized Dictionary condition (48.03) with the mean under the Standard condition (47.91). As the data in Table 20 show, Grade 4 ELL students under the Customized Dictionary accommodation performed slightly better than under the Standard condition, but the difference did not reach the .05 statistical significance (p = .83). In the second test, we found that Grade 4 ELL students performed slightly lower (47.28) with the English-to-Spanish Glossary than under the Standard condition (47.91). This difference did not reach statistical significance (p = .38). Lastly, ELL students performed slightly lower with the Linguistic Modification version of the test (47.20) than under the Standard condition (47.91). However, the difference did not reach statistical significance at the .05 level (p = .21).

Based on the results of the analysis presented above, the three accommodation strategies used in this study did not help Grade 4 ELL students improve their performance. The accommodation strategies used in this study were all language

Table 20

Grade 4 ELL Mean Latent Science Achievement Scores Adjusted for Reading Achievement and Home Language

Accommodation	ELL adjusted means	Contrast with Standard condition
Standard condition	47.91 (*SE* = .41; *n* = 268)	NA
Customized Dictionary	48.03 (*SE* = .40; *n* = 270)	*p* = .83
English-to-Spanish Glossary	47.28 (*SE* = .59; *n* = 135)	*p* = .39
Linguistic Modification	47.20 (*SE* = .39; *n* = 284)	*p* = .21

related and were supposed to help ELL students with their possible language limitations. However, the results of our earlier studies have suggested that in lower grades, language may not be as important a factor as it is in the higher grades (see, for example, Abedi, Leon, et al., 2001). We speculate that the original Grade 4 test version had less language demand than tests for higher grades have, so that language-related accommodation strategies subsequently have less impact.

Validity. Accommodation strategies are valid if they do not impact the construct under measurement; that is, if the accommodations do not change the performance of non-ELL students for whom the accommodations are not intended. To test the validity of the accommodations for our Grade 4 sample, the performance of non-ELL students under accommodation was compared to the performance of non-ELL students under the Standard condition. Similar to the statistical approach that was used for testing the effectiveness hypothesis, we conducted a series of *a priori* or planned comparisons. Table 21 summarizes the results of the planned comparison for testing the validity of accommodation. Descriptive statistics were reported for the science latent test scores. The science test scores were adjusted by students' reading proficiency score, and three planned comparisons were conducted. In these comparisons, student performance under each of the three accommodations was compared to student performance under the Standard condition. As can be seen, none of the comparisons was significant. The probability of a Type I error was above .05 in all three comparisons. These results suggest that the accommodation strategies used in this study did not affect the performance of non-ELL students; therefore, the accommodations did not alter the construct under measurement.

Table 21

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

Accommodation	Non-ELL adjusted means	Contrast with Standard condition
Standard condition	52.26 (*SE* = .48; *n* = 241)	NA
Customized Dictionary	52.65 (*SE* = .47; *n* = 247)	*p* = .57
English-to-English Glossary	53.31 (*SE* = .74; *n* = 101)	*p* = .24
Linguistic Modification	52.80 (*SE* = .46; *n* = 257)	*p* = .42

Results for Grade 8 Students

Similar to the data reported for students in Grade 4, on average, non-ELL students in Grade 8 (T-scores *M* = 53.12, *SD* = 9.59; raw scores *M* = 12.61, *SD* = 3.91; *n* = 856) outperformed ELL students (T-scores *M* = 46.35, *SD* = 9.21; raw scores *M* = 9.44, *SD* = 3.62; *n* = 733) by about 7 points. Table 22 presents the descriptive statistics for the Grade 8 science latent scores.

Among the ELL students in Grade 8, the type of accommodation made a difference in test scoring. ELL students with the Linguistic Modification version of the test scored the highest (T-scores *M* = 47.63, *SD* = 9.53; raw scores *M* = 9.94, *SD* = 3.88; *n* = 209), followed by students under the Customized Dictionary condition (T-scores *M* = 46.68, *SD* = 9.00; raw scores *M* = 9.36, *SD* = 3.51; *n* = 206) and the Standard condition (T-scores *M* = 45.73, *SD* = 9.41; raw scores *M* = 9.30, *SD* = 3.70; *n* = 199). ELL students under the Bilingual/English Glossary condition scored the lowest (T-scores *M* = 44.58, *SD* = 8.38; raw scores *M* = 8.93, *SD* = 3.11; *n* = 119).

Among the non-ELL sample, Grade 8 students under the Bilingual/English Glossary condition performed the lowest (T-scores *M* = 50.73, *SD* = 8.58; raw scores *M* = 11.60, *SD* = 3.87; *n* = 129). Under the Customized Dictionary condition (T-scores *M* = 53.17, *SD* = 9.84; raw scores *M* = 12.73, *SD* = 3.81; *n* = 241) and Linguistic Modification condition (T-scores *M* = 54.00, *SD* = 8.97; raw scores *M* = 12.90, *SD* = 3.76; *n* = 241), non-ELL students in Grade 8 performed about the same as those under the Standard condition (T-scores *M* = 53.48, *SD* = 10.26; raw scores *M* = 12.75, *SD* = 4.10; *n* = 245).

Table 22

Grade 8 Mean Latent Science Achievement Scores (M = 50, SD = 10)

Accommodation	ELL status		Row total (ELL + non-ELL)
	ELL	Non-ELL	
Standard condition	45.73 (SD = 9.41; n = 199)	53.48 (SD = 10.26; n = 245)	50.01 (SD = 10.60; n = 444)
Customized Dictionary	46.68 (SD = 9.00; n = 206)	53.17 (SD = 9.84; n = 241)	50.18 (SD = 9.99; n = 447)
Bilingual / English Glossary	44.58 (SD = 8.38; n = 119)	50.73 (SD = 8.58; n = 129)	47.78 (SD = 9.01; n = 248)
Linguistic Modification	47.63 (SD = 9.53; n = 209)	54.00 (SD = 8.97; n = 241)	51.04 (SD = 9.76; n = 450)
Column total	46.35 (SD = 9.21; n = 733)	53.12 (SD = 9.59; n = 856)	50.00 (SD = 10.00; n = 1589)

Table 23 presents the descriptive statistics for the Grade 8 latent reading scores. The latent reading score was used as a covariate in the model comparing students' science scores under different forms of accommodation. That is, students' science scores were adjusted by reading proficiency scores. We also adjusted for Spanish being the home language.

Consistent with the data presented earlier in this report and also with earlier studies, ELL students performed substantially lower (T-scores M = 45.93, SD = 9.16; raw scores M = 9.37, SD = 2.84; n = 733) than their non-ELL counterparts (T-scores M = 53.49, SD = 9.36; for raw scores M = 12.33, SD = 2.75; n = 856) on the reading test. There were also some differences in the reading test scores within ELL and non-ELL

Table 23

Grade 8 Mean Latent Reading Achievement Scores (M = 50, SD = 10)

Accommodation	ELL status		Row total (ELL + non-ELL)
	ELL	Non-ELL	
Standard condition	45.50 (SD = 9.98; n = 199)	53.59 (SD = 9.53; n = 245)	49.97 (SD = 10.52; n = 444)
Customized Dictionary	48.78 (SD = 9.10; n = 206)	53.48 (SD = 9.86; n = 241)	50.39 (SD = 10.08; n = 447)
Bilingual / English Glossary	45.17 (SD = 8.54; n = 119)	51.71 (SD = 9.07; n = 129)	48.57 (SD = 9.39; n = 248)
Linguistic Modification	45.93 (SD = 8.71; n = 209)	54.34 (SD = 8.74; n = 241)	50.43 (SD = 9.67; n = 450)
Column total	45.93 (SD = 9.16; n = 733	53.49 (SD = 9.36; n = 856)	50.00 (SD = 10.00; n = 1589)

groups across the accommodation categories. For example, non-ELL students who took the Linguistic Modification version of the science test had the highest non-ELL reading scores (T-scores M = 54.34, SD = 8.74; for raw scores M = 12.62, SD = 2.92; n = 241). ELL students who took the science test under the Customized Dictionary condition had the highest ELL reading scores (T-scores M = 48.78, SD = 9.10; for raw scores M = 9.54, SD = 2.70; n = 206). However, since the science test scores were adjusted by the students' reading scores, these initial differences did not affect the outcome of this study.

Effectiveness. To test the effectiveness hypothesis, the performance of ELL students under accommodation was compared to that of ELL students under the Standard condition. A series of *a priori* or planned tests were conducted to see whether any accommodation helped ELL students. Analyses were performed on the science test scores that were adjusted by students' reading proficiency scores and for home language. Table 24 summarizes the results of planned comparisons for testing effectiveness of accommodations used for Grade 8 students. Adjusted mean science scores, the standard error, and the number of students in each group are reported. The only accommodation that significantly impacted the performance of ELL students was the Linguistic Modification version (M = 47.52, SE = .50, n = 209) which was significant at the .05 nominal level (p = .03). This effect was in the expected direction. The other two accommodations did not show any significant impact on the performance of ELL students. For the Customized Dictionary condition, the probability of a Type I error of .91 was obtained. For the Bilingual/English Glossary condition, the p value was .68.

Table 24

Grade 8 ELL Mean Latent Science Achievement Scores Adjusted for Reading Achievement and Home Language

Accommodation	ELL adjusted means	Contrast with Standard condition
Standard condition	45.94 (SE = .51; n = 199)	NA
Customized Dictionary	46.01 (SE = .51; n = 206)	p = .91
Bilingual/English Glossary	45.58 (SE = .68; n = 119)	p = .68
Linguistic Modification	47.52 (SE = .50; n = 209)	p = .03

Validity. As indicated, the Linguistic Modification version was the only effective accommodation for ELL students in Grade 8. That is, the performance of ELL students under this accommodation improved significantly. However, the results on effectiveness of this accommodation may not be sufficient to judge its appropriateness in the assessment of ELL students. If this accommodation also had helped non-ELL students, then one could argue that the construct under measurement had been altered, which would cast doubt on the validity of this accommodation.

To test accommodation validity for the Grade 8 data, we compared the performance of non-ELL students under the different accommodations. In making these comparisons, we conducted a series of *a priori* tests. Table 25 presents the adjusted science mean scores along with the standard error and number of students in each group. It also presents the probability of a Type I error rate for significance of each of the three comparisons. As the data in Table 25 suggest, none of the comparisons was significant. That is, none of the accommodation strategies had any impact on non-ELL students' performance in science. For example, the mean science score for non-ELL students with the Linguistic Modification version (M = 53.42, SE = .47, n = 241) was almost identical with the mean science score for students under the Standard condition (M = 53.38, SE = .47, n = 245). These results suggest that accommodation strategies used in this study did not impact the construct under measurement and can be used for ELL students without adversely affecting the validity of the accommodation.

Table 25

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

Accommodation	Non-ELL adjusted means	Contrast with Standard condition
Standard condition	53.38 (SE = .47; n = 245)	NA
Customized Dictionary	53.17 (SE = .47; n = 241)	p = .74
Bilingual/English Glossary	51.99 (SE = .65; n = 129)	p = .09
Linguistic Modification	53.42 (SE = .47; n = 241)	p = .96

Differential Impact

Using the Grade 8 data, a two-way ANCOVA was run in order to examine whether the impact of accommodation on science performance differed by primary home language. (As stated above, the science test scores were adjusted by the students' reading proficiency scores.) The glossary accommodation was not included in this model, as glossaries were not available for students with a primary language other than English or Spanish. The results of the model indicate that the main effect of primary home language was significant ($p < .001$). The main effect of accommodation was not significant ($p < .18$). The interaction between accommodation and primary home language was significant ($p < .04$), suggesting a differential impact of accommodation on the primary home language. Table 26 and Figure 5 help to explain this interaction. Students with English as the primary home language who took the Linguistic Modification version of the test performed slightly lower in science (adjusted mean = 50.36) than those tested under the Standard condition (adjusted mean = 50.84). On the other hand, students with a non-English, non-Spanish primary home language taking the Linguistic Modification version of the test performed substantially higher in science (adjusted mean = 52.97) than those tested under the Standard condition (adjusted mean = 50.93). In other words, it appears that ELL students with a home language other than English or Spanish (e.g., Chinese, Korean, etc.) benefited from the Linguistic Modification version.

Item-Level Analysis

As discussed earlier, linguistic modification of test items was the only effective accommodation in this study. None of the other accommodations helped to significantly reduce the performance gap between ELL and non-ELL students. To learn more about the performance of students under the Linguistic Modification

Table 26

Grade 8 Adjusted Mean Latent Science Achievement by Primary Home Language

	Primary home language		
Accommodation	English	Spanish	Other non-English
Standard condition	50.84 (*SE* = .64; *n* = 135)	49.36 (*SE* = .53; *n* = 113)	50.93 (*SE* = .69; *n* = 196)
Customized Dictionary	51.75 (*SE* = .65; *n* = 127)	48.93 (*SE* = .52; *n* = 115)	50.33 (*SE* = .69; *n* = 205)
Linguistic Modification	50.36 (*SE* = .64; *n* = 134)	50.18 (*SE* = .52; *n* = 112)	52.97 (*SE* = .69; *n* = 204)

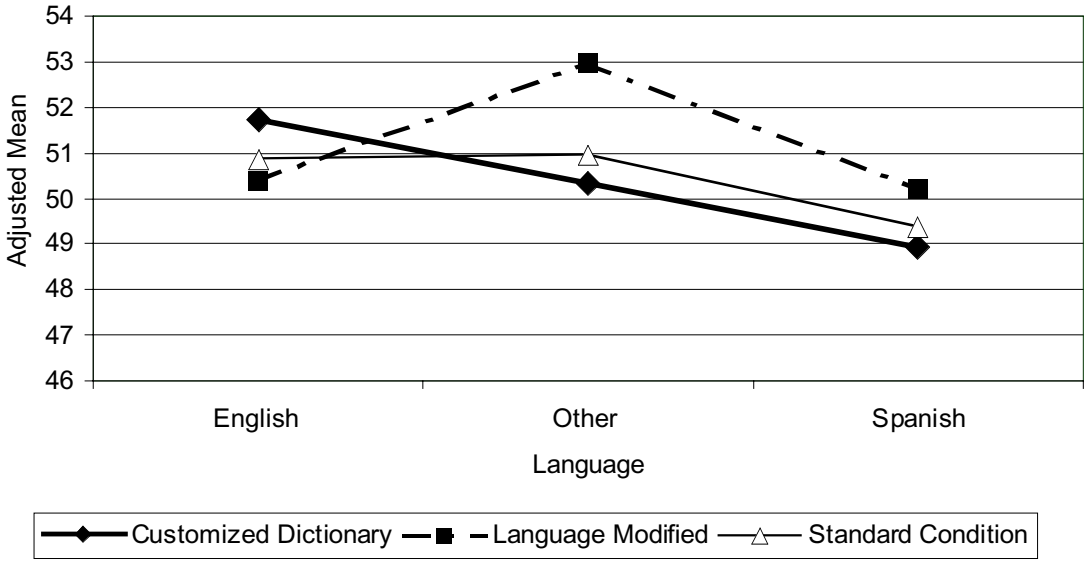

Figure 5. Grade 8 adjusted science means by primary home language by accommodation.

accommodation, we performed analyses at the item level. As reported earlier, in general, ELL students scored lower on the science test than non-ELL students. If this performance difference was mainly due to language factors, one would expect larger performance differences between ELL and non-ELL students on items with a higher level of linguistic complexity. To determine differential performance of ELL students on the science test due to the linguistic complexity of the items, the difference between the performance of ELL and non-ELL students on each science item was computed. As expected, the difference in the *p* value (proportion of correct response) between ELL and non-ELL students varied greatly across the science test items with different levels of linguistic complexity. If larger differences between the performance of ELL and non-ELL students correspond to greater linguistic complexity, one can then attribute the performance difference between ELL and non-ELL students mainly to language factors. The results of our item-level analyses confirmed this assumption and indicated that the higher the level of linguistic complexity, the larger the performance difference between ELL and non-ELL students. Also, the larger the performance difference between ELL and non-ELL students, the more that language modification of test items helped to reduce the performance gap.

Table 27 presents a summary of item-level analyses on the science test items for Grade 8. The information includes the differences in item *p* values for ELL and non-ELL students, as well as for each accommodation group compared to the Standard group. A negative sign in the *p* value indicates higher performance by non-ELL students. We sorted test items based on the size of *p* value differences between ELL and non-ELL students. Out of the 30 science test items, 26 of them had negative *p* value differences between ELL and non-ELL students ranging from –.24 to –.01, indicating that ELL students performed lower on those items (column A of the table).

In Table 27 we also present the *p* value differences between the accommodated groups and the Standard condition. Comparing *p* value differences between ELL and non-ELL students with the *p* value differences of different forms of accommodation reveals interesting trends. Compared with the Standard condition, students tested with the Linguistic Modification version showed more improvement than those tested with other accommodations. For example, students performed better on 22 out of 30 items in the Linguistic Modification version (column D) than students tested under the Customized Dictionary condition (performed better on 14 items; column B) and students tested under the Bilingual/English Glossary condition (performed better on 13 items; column C). There are 13 items on which non-ELL students outperformed ELL students with a *p* value difference of 0.11 or greater (column A). For all 13 items, students who received the Linguistic Modification version of the science test outperformed those who received the Standard condition.

We also computed the correlation between the *p* value differences. A negative correlation between the ELL/non-ELL *p* value differences and the accommodated/Standard condition *p* value differences indicated the level of effectiveness of each accommodation in reducing the performance gap between ELL and non-ELL students. This negative correlation was the greatest for the Linguistic Modification version ($r = –.60$, $p<.01$). The Customized Dictionary condition missed significance at the .05 nominal level ($r = –.35$, $p = .06$). The Bilingual/English Glossary condition did not reach the .05 nominal level of significance ($r = –.17$, $p = .38$).

Table 27

Mean *P* Value Differences for ELL Students, Test Booklet A

Science item #	A ELL minus non-ELL (*p* value difference)	B Customized Dictionary minus Standard condition	C Bilingual/English Glossary minus Standard condition	D Linguistic Modification minus Standard condition
2	–0.24	0.02	–0.02	0.10
5	–0.24	–0.06	–0.06	0.08
10	–0.24	–0.01	–0.02	0.02
28	–0.24	0.21	0.10	0.13
6	–0.19	–0.01	–0.10	0.10
13	–0.19	0.00	0.10	0.18
4	–0.18	0.03	–0.06	0.04
11	–0.17	0.11	0.07	0.11
26	–0.17	0.16	0.13	0.09
22	–0.02	0.03	–0.01	0.05
1	–0.15	0.08	–0.04	0.02
25	–0.15	–0.01	0.00	0.03
12	–0.11	–0.03	0.14	0.05
16	–0.10	–0.02	0.01	–0.06
14	–0.10	0.09	0.09	–0.01
20	–0.09	–0.09	–0.08	0.02
24	–0.09	0.01	0.16	0.03
3	–0.08	–0.05	–0.02	0.04
9	–0.08	–0.02	0.00	–0.04
18	–0.06	–0.01	–0.03	0.00
19	–0.04	0.02	–0.01	0.05
27	–0.04	–0.05	0.06	–0.02
8	–0.03	0.04	–0.02	0.01
17	–0.03	0.05	–0.12	0.07
23	–0.01	–0.01	–0.02	–0.01
21	–0.01	–0.01	–0.02	0.06
30	0.03	–0.04	0.02	0.04
7	0.04	0.03	–0.07	–0.01
15	0.04	–0.08	0.08	–0.05
29	0.10	–0.01	–0.08	–0.02
Correlation with column A	NA NA	–0.35 ($p < .06$)	–0.17 ($p < .38$)	–0.60 ($p < .01$)

Background and Accommodation Questionnaires

In addition to taking science and reading tests, students responded to a set of background and accommodation questions. Students' responses to these questions provided additional information for our research hypotheses. We will first discuss the results from the data on the background questions and then address the results based on the accommodation questions.

Background questions. The same background questionnaires were used for Grades 4 and 8 with only one difference: The background questionnaire for Grade 4 students was linguistically simplified to assure more valid responses. The background questionnaire asked students in about their country of origin, length of time in the United States, initial grade in which they attended school in the United States, and whether a language other than English was spoken in the home. It also asked students to self-assess their proficiency both in English and in their home language, if not English.

Background questions for Grade 4. All Grade 4 students were given the student background questionnaire after they took the science test and completed the accommodation questionnaire. Table 28 summarizes the response frequencies to background questions for students in Grade 4. As the data show, regarding country of birth, a majority of students in Grade 4 (79.5%) indicated that they had been born in the United States, and only 20% stated that they had been born in other countries. This relatively large percentage of students indicating that they were born in the United States may be inconsistent with the number of students who were categorized as ELL: Even though ELL students are not necessarily born outside the United States, a majority of them are expected to be.

In response to the question about how long they have lived in the United States, a majority of Grade 4 students (68.0%) indicated that they had lived in the United States all their life. This response also may be inconsistent with the structure of the sample of Grade 4 students, about 50% of whom are designated as ELL students.

Of the Grade 4 students, 944 or 56.7% indicated that the initial grade they attended school in the United States was preschool; 489 or 29.4% said that their initial grade in the United States was kindergarten. The rest (less than 14%) indicated that they first attended school in the United States after kindergarten.

Table 28

Grade 4 Frequencies and Percentages for Student Background Questionnaire: Birth Country, Time in U.S. , Initial Grade

#	Question	Frequency	%
1	Country of birth:		
	China	30	1.8
	Cuba	1	0.1
	Korea	20	1.2
	Mexico	156	9.3
	Taiwan	5	0.3
	United States	1330	79.5
	Other	131	7.8
2	Time lived in United States:		
	Less than 1 year	26	1.6
	1 year	28	1.7
	2 years	37	2.2
	3 years	53	3.2
	4 years	63	3.8
	More than 4 years	327	19.6
	All my life	1135	68.0
3	Initial grade attended school in United States:		
	Preschool	944	56.7
	Kindergarten	489	29.4
	1st grade	88	5.3
	2nd grade	55	3.3
	3rd grade	42	2.5
	4th grade	43	2.6
	5th grade	4	0.2

In response to the question about language spoken in the home (Table 29), English was the most frequent language spoken in the home (701 or 43.4%), followed by Spanish (554, or 34.3%), followed by Chinese (160 or 9.9%), and other languages (143 or 8.9%). A small number of students indicated that they spoke Korean in the home (3.5%). The frequencies and percentages of the languages spoken in the home correspond more closely with the distribution of students based on their ELL status. Close to half of the students spoke English and half spoke a language other than English in the home.

Table 29 also shows the mean Likert scale scores for self-reported proficiency in English and another language (the student's native language). The Likert scale

Table 29

Grade 4 Frequencies and Percentages for Student Background Questionnaire: Language Fluency

#	Question	Frequency	%
12	What language other than English do you speak at home now?		
	Chinese	160	9.9
	English	701	43.4
	Korean	57	3.5
	Spanish	554	34.3
	Other	143	8.9
		M[a]	*SD*
6	How well can you understand spoken English at school?	3.61	0.67
13	How well do you speak the other language at home?	3.59	0.66
14	How well do you read the other language at home?	3.21	0.98
15	How well do you write the other language at home?	3.11	1.00

[a]Students responded on a Likert scale ranging from 1 (*not very well at all*) to 4 (*very well*).

ranged from 1 (*not very well at all*) to 4 (*very well*). All the Likert scale averages are above 3.0, indicating that the sampled students believed that they were proficient in English and the other language (when applicable).

Table 30 presents the response frequencies for the Grade 4 background questions by ELL status. Some of the background questions show different response patterns across the ELL categories. For example, a higher proportion of non-ELL students indicated that they had been born in the United States (89.7%) than ELL students (70.7%). More non-ELL students indicated that they lived in the United States "All my life" (79.2%) than ELL students (58.3%). A similar trend exists for the questions on the initial grade of school attended in the U.S. and on the language spoken in the home. Of the non-ELL respondents to the question about language spoken in the home, 68.4% indicated that they spoke English at home as compared to 21.7% of ELL students who indicated that they spoke English at home.

Table 30 also reports the mean Likert scale scores for self-reported proficiency in language by student ELL status. As expected, non-ELL students reported a higher level of understanding of spoken English (M = 3.78, SD = 0.50) than the ELL students (M = 3.46, SD = 0.76). Responses to questions concerning a language other than English were not expected from the non-ELL students, the majority of whom were native English speakers.

Table 30

Grade 4 Student Background Questions by ELL Status

#	Question	ELL		Non-ELL	
		Frequency	%	Frequency	%
1	Country of birth:				
	China	26	2.9	4	0.5
	Cuba	1	0.1	0	0.0
	Korea	13	1.4	7	0.9
	Mexico	132	14.7	24	3.1
	Taiwan	3	0.3	2	0.3
	United States	634	70.7	696	89.7
	Other	88	9.8	43	5.5
2	Time lived in the United States:				
	Less than 1 year	24	2.7	2	0.3
	1 year	24	2.7	4	0.5
	2 years	30	3.4	7	0.9
	3 years	40	4.5	13	1.7
	4 years	50	5.6	13	1.7
	More than 4 years	205	22.9	122	15.7
	All my life	521	58.3	614	79.2
3	Initial grade attended school in the United States:				
	Preschool	440	49.3	504	65.3
	Kindergarten	280	31.4	209	27.1
	1st grade	59	6.6	29	3.8
	2nd grade	44	4.9	11	1.4
	3rd grade	31	3.5	11	1.4
	4th grade	36	4.0	7	0.9
	5th grade	3	0.3	1	0.1
12	Now, at home, we speak mostly:				
	Chinese	139	16.1	21	2.8
	English	187	21.7	514	68.4
	Korean	30	3.5	27	3.6
	Spanish	429	49.7	125	16.6
	Other	78	9.0	65	8.6
		M[a]	*SD*	*M*[a]	*SD*
6	How well can you understand spoken English at school?	3.46	0.76	3.78	0.50
13	How well do you speak the other language at home?	3.57	0.69	NA[b]	NA[b]
14	How well do you read the other language at home?	3.19	0.98	NA[b]	NA[b]
15	How well do you write the other language at home?	3.10	0.98	NA[b]	NA[b]

[a]Students responded on a Likert scale from 1 (*not very well at all*) to 4 (*very well*). [b]NA = Not applicable.

Background questions for Grade 8. Similar to the approach used with Grade 4 students, all Grade 8 students were given a student background questionnaire after the science test and accommodation questionnaire. The questions were the same as those in the Grade 4 student background questionnaire. Tables 31 and 32 present frequencies and percentages of responses by Grade 8 students to the background questions. The trend of responses was very similar to those observed for Grade 4 students. A majority of Grade 8 students (65.5%) indicated that they had been born in the United States. Also, a majority of them 63.5%) said that they lived all their life in the United States. Over 70% of the students indicated that they initially attended preschool or kindergarten in the United States, and about half of the students (46.6%) reported that they spoke English in the home.

Similar to the data reported for Grade 4 students, all mean scores (on a Likert scale ranging from 1—*not very well at all*—to 4—*very well*) for language proficiency (English and the other language) were high (greater than 3), suggesting that students represented themselves as proficient in understanding spoken English and in speaking, reading, and writing another language.

Table 32 presents the frequency of responses to the Grade 8 background questionnaire by ELL category. The percentage of non-ELL students who had been born in the United States (85.2%) was substantially higher than the percentage of ELL students who indicated that they had been born in the United States (42.1%). The percentage of non-ELL students who indicated that they had lived in the United States for more than 4 years or always was also substantially larger (97.4%) than the corresponding percentage for ELL students (70.1%). Of the entire non-ELL sample of Grade 8 students, 68% indicated that they spoke English in the home. Non-ELL students also reported a higher level of proficiency in understanding spoken English (M = 3.76, SD = .48) than the ELL students did (M = 3.26, SD = .74).

The results of our analyses of the student background questionnaires indicate that responses to the background questions were not as reliable as responses to content-based items. Language factors may play a role in this case. We found more discrepancies with the responses of Grade 4 students than with those of Grade 8 students. This may be due to language factors. Understanding the content of the questions was more difficult for Grade 4 students than for Grade 8 students, in spite of our post-pilot efforts to make the questionnaire language simpler and more direct.

Table 31

Grade 8 Frequencies and Percentages for Student Background Questionnaire: Birth Country, Time in U.S., Initial Grade, Language Fluency

#	Question	Frequency	%
1	Country of birth:		
	China	31	2.2
	Cuba	2	0.1
	Korea	16	1.1
	Mexico	256	17.9
	Taiwan	1	0.1
	United States	937	65.5
	Other	187	13.1
2	Time lived in the United States:		
	Less than 1 year	34	2.4
	1 year	31	2.2
	2 years	54	3.8
	3 years	52	3.6
	4 years	45	3.1
	More than 4 years	307	21.4
	All my life	909	63.5
3	Initial grade attended school in the United States:		
	Preschool	700	49.1
	Kindergarten	356	24.9
	1st grade	51	3.6
	2nd grade	27	1.9
	3rd grade	44	3.1
	4th grade	46	3.2
	5th grade	47	3.3
	6th grade	62	4.3
	7th grade	55	3.9
	8th grade	39	2.7
11	Now at home, we speak mostly:		
	English	622	46.6
	Chinese	71	5.3
	Korean	24	1.8
	Spanish	512	38.3
	Other	107	8.0
		M[a]	*SD*
7	How well can you understand spoken English at school?	3.53	0.66
12	How well do you speak the other language at home?	3.49	0.65
13	How well do you read the other language at home?	3.19	0.89
14	How well do you write the other language at home?	3.12	0.90

[a]Students responded on a Likert scale from 1 (*not very well at all*) to 4 (*very well*).

Table 32

Grade 8 Student Background Questionnaire by ELL Status

#	Question	ELL		Non-ELL	
		Frequency	%	Frequency	%
1a	Country of birth:				
	United States	275	42.1	662	85.2
	Others	378	57.9	115	14.8
2	Time lived in the United States:				
	1 year or less	63	9.6	2	0.3
	2–4 years	133	20.3	18	2.3
	More than 4 years (all my life)	460	70.1	756	97.4
3	Initial grade attended school in the United States:				
	Preschool	206	31.6	494	63.7
	Kindergarten	153	23.5	203	26.2
	1st grade–4th grade	106	16.3	62	8.0
	5th grade–8th grade	187	28.7	16	2.1
12	Now at home we speak mostly:				
	English	125	20.7	497	68.0
	Asian (Chinese, Korean)	60	9.9	35	4.8
	Spanish	343	56.7	169	23.1
	Other	77	12.7	30	4.1
7	After reading a book at school, which would you be able to do?				
7a	Oral book report	234	32.8	412	48.9
7b	Written book report	400	56.1	571	67.8
7c	Multiple-choice test	254	35.6	507	60.2
		M[a]	*SD*	*M*[a]	*SD*
6	How well can you understand spoken English at school?	3.26	0.74	3.76	0.48
13	How well do you speak the other language at home?	3.41	0.69	3.62	0.55
14	How well do you read the other language at home?	3.16	0.86	3.24	0.93
15	How well do you write the other language at home?	3.08	0.89	3.19	0.91

[a]Students responded on a Likert scale from 1 (*not very well at all*) to 4 (*very well*).

Accommodation follow-up questionnaire results. In any accommodation study, it is extremely helpful to collect data from the accommodation recipients on the applicability and usefulness of the accommodation strategies. It also is informative to know how much the students actually used the accommodations. To

obtain data on accommodation from students' perspectives, we developed and used an accommodation follow-up questionnaire. As indicated in the Design and Procedure section earlier, we used four testing conditions, all with extra time (50%): (1) a Customized Dictionary, (2) a Bilingual / English Glossary (English-to-Spanish or English-to-English), (3) a Linguistic Modification test version, and, to obtain baseline data, (4) a Standard testing condition. The same accommodation questions were used for the four booklets.

Accommodation questionnaire, Grades 4 and 8. All students were given the accommodation follow-up questionnaire after the science test. Table 33 presents the accommodation questions.

To examine the pattern of responses across the ELL categories (comparing responses of ELL and non-ELL students), frequencies of responses to the accommodation questions were obtained separately for each group.

Accommodation question 1 asks students whether there were words that they did not understand on the science test. Response options to this question were "I had no problem understanding the science test," "Some words," and "Many words." The response options used a Likert format. Mean ratings for the questions were computed. To compare the response patterns of ELL and non-ELL students, we present the data and general results for both groups and both grade levels.

Table 33

Accommodation Questionnaire Items

Q1. In the science test, I did not understand: (*Some words; Many words; I had no problem understanding . . .*)
Q2. In the science test, I did not understand: (*Some sentences; Many sentences; I had no problem understanding . . .*)
Q3. Most of these science problems were: (*Very easy; Easy; Difficult; Very difficult*)
Q4. Did you look up words during the test? (*Occasionally; For about half …; Often; For every problem; I didn't have a glossary; I had a glossary, but I didn't use it.*)
Q5. Did the glossary help you? (*No, Yes some, Yes a lot, I didn't have one.*)
Q6. If the glossary was not helpful, why not? (*The words … were not there; The definitions…were hard to understand; I had a glossary but I already understood…; I didn't have a glossary; The glossary was helpful.*)
Q7. To make it easier for me to understand the science questions, please give me: [Please choose all that are true for you.] (Eleven types of accommodation were listed.)

Tables 34 and 35 present data for the first accommodation question for Grade 4 and Grade 8, respectively. This item asked about the difficulty of the test words. The overall mean for non-ELL students was 1.64 (*SD* = .58, *n* = 769) for Grade 4 and 1.59 (*SD* = .58, *n* = 658) for Grade 8. For ELL students, the mean was 1.84 (*SD* = .59, *n* = 883) for Grade 4 and 1.89 (*SD* = .58, *n* = 559) for Grade 8. The mean for ELL students was higher than the mean for non-ELL students, which suggests that ELL students found the test words more difficult to understand. However, the mean difference between ELL and non-ELL students was almost identical across the different

Table 34

Grade 4 Means for "In the science test I did not understand..."

Accommodation	ELL	Non-ELL	Row total
Standard condition	1.82 (*SD* = .57; *n* = 244)	1.61 (*SD* = .54; *n* = 211)	1.72 (*SD* = .57; *n* = 455)
Customized Dictionary	1.82 (*SD* = .59; *n* = 254)	1.63 (*SD* = .59; *n* = 227)	1.73 (*SD* = .60; *n* = 481)
Bilingual/English Glossary	1.90 (*SD* = .65; *n* = 123)	1.70 (*SD* = .60; *n* = 98)	1.81 (*SD* = .63; *n* = 221)
Linguistic Modification	1.86 (*SD* = .58; *n* = 262)	1.65 (*SD* = .60; *n* = 233)	1.76 (*SD* = .60; *n* = 495)
Column total	1.84 (*SD* = .59; *n* = 883)	1.64 (*SD* = .58; *n* = 769)	1.75 (*SD* = .59; *n* = 1652)

Note. Responses: 1 = *I had no problem*, 2 = *some words*, 3 = *many words.*

Table 35

Grade 8 Means for "In the science test I did not understand..."

Accommodation	ELL	Non-ELL	Row total
Standard condition	1.93 (*SD* = .59; *n* = 145)	1.61 (*SD* = .57; *n* = 178)	1.75 (*SD* = .60; *n* = 323)
Customized Dictionary	1.88 (*SD* = .57; *n* = 148)	1.57 (*SD* = .59; *n* = 176)	1.71 (*SD* = .60; *n* = 324)
Bilingual/English Glossary	1.89 (*SD* = .56; *n* = 97)	1.62 (*SD* = .62; *n* = 109)	1.75 (*SD* = .60; *n* = 206)
Linguistic Modification	1.87 (*SD* = .59; *n* = 169)	1.58 (*SD* = .56; *n* = 195)	1.71 (*SD* = .59; *n* = 364)
Column total	1.89 (*SD* = .58; *n* = 559)	1.59 (*SD* = .58; *n* = 658)	1.73 (*SD* = .60; *n* = 1217)

Note. Responses: 1 = *I had no problem*, 2 = *some words*, 3 = *many words.*

accommodation conditions (a mean difference of around .20). This pattern is representative of many of the response results for this questionnaire. (See Appendix B for additional response results for the accommodation questionnaire.)

Because we did not specify any research hypotheses on the background and accommodation follow-up questionnaires, we have supplied descriptive results. Thus, no statistical significance tests were planned to compare response patterns of ELL and non-ELL students. See the Discussion section for a summary of the accommodation questionnaire results.

Discussion

The main goals of this study were to assess whether selected language accommodations are effective and valid in the type of science test used in large-scale assessments. In addition, student background variables were studied to judge their impact on student test performance. Briefly, the accommodations examined in this study included use of a Customized English Dictionary, a Bilingual/English Glossary, and a Linguistic Modification version of the science test.

Taking into account the quantitative and qualitative findings of the study, we will discuss what is immediately apparent about the effectiveness, validity, and feasibility of these accommodations.

Accommodation Justifications

The literature shows that many different forms of accommodation strategies have been used nationwide in the assessment of ELL students, and some of these strategies were discussed in the Literature Review section. We tried to select accommodation strategies based on several criteria: frequency of usage and nationwide recognition, feasibility, and a direct relationship to language decoding. Each of the accommodation strategies that we used in this study (Customized English Dictionary, Bilingual/English Glossary, and Linguistic Modification of test items) can clearly function as an aid to the language needs of ELL students.

Reading Test Justification

Science tests contain blocks of text that must be deciphered accurately in order for mastery of the science material to be demonstrated. To compare the science test results of students in the various accommodation groups, it was necessary to take into account their reading proficiency levels as a measure of their ability to decipher the test items.

ELL designation is useful as an initial discrimination tool for distributing accommodations among ELL students. However, the designation is not determined in the same manner everywhere, nor is it kept up-to-date throughout the school year.

English reading ability is a desirable covariate. However, students' local reading assessments also vary widely among schools and school districts. It is therefore necessary to compare students' science test ability with a well-known and accurate measure of reading ability before making observations about the effect of accommodations designed to aid reading and understanding. A reading test provides a more trustworthy covariate than a student's ELL designation.

Selecting a reading test that can serve both ELL and non-ELL students is a challenge. A test that could be at the optimum level of difficulty for one group may not be so for another group. For example, the results of our earlier studies indicated that some of the test items in the LAS might be too easy for non-ELL students and even, in some cases, for ELL students. On the other hand, our earlier studies also indicated that the NAEP Reading blocks may be too difficult for ELL students. These test items, therefore, may not have enough discrimination power for either group. To provide a more valid instrument for measuring ELL and non-ELL students' levels of reading proficiency, we decided to use a battery of test items, thus providing discrimination power for both ELL and non-ELL students. The LAS portion of our reading instrument provided a distribution of proficiency among ELL students, and the NAEP section of the tool gave a distribution among the non-ELL students. In this way, a more descriptive continuum of reading ability was derived.

Questionnaire Justifications

A student background questionnaire was developed to examine the impact a student's background might have on accommodated assessment. This information allows us to test whether background impacts performance on the science assessment and whether this impact differs under the various accommodated conditions. Informed decisions can then be made with regard to which accommodations should be used for students with different background types.

During the pilot study, questions arose as to whether the students were effectively using the accommodations. As a result, an accommodation questionnaire was developed for student self-reporting of the effectiveness of the various accommodated conditions. This information has confirmed and put into context the

results of the accommodation analysis. Questionnaires were also developed to collect information at both the teacher and school levels.

Design Justifications

In this study, we included multiple forms of accommodation to enable us to compare the effectiveness of accommodation by different approaches. We sampled students from different language and cultural backgrounds to check for any possible cultural and/or linguistic interference that may impact the outcome of accommodated assessment. Possible variables included any effect of home language, approaches to test taking, and attitudes towards glossary and dictionary use. 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 the general student population. Finally, we included a measure of English efficiency/proficiency because we believe neither ELL nor non-ELL groups are homogeneous within themselves. ELL and non-ELL students vary substantially in their English reading capabilities, and the effectiveness of accommodation, to a greater extent, depends on students' English language backgrounds.

There are many different issues concerning the use of accommodation. Among them, the issues concerning effectiveness, validity, and feasibility are especially important. Accommodations that are not effective in reducing the performance gap between ELL and non-ELL students may not have any practical use. Accommodations that are effective but logistically difficult to administer are not feasible, especially in large-scale assessments. More importantly, if accommodation affects the construct, even if it is effective and feasible, its use may not be valid.

Observations on Glossary Accommodations

The types of accommodations we tested seemed efficient, valid, and feasible, based on previous accommodation studies and on our pilot study. After our pilot testing revealed several flaws with published English dictionary and bilingual dictionary use, we created the lexicon accommodation tools for this study to provide an aid to unfamiliar words in the test, excluding science content words. The accommodations either defined, translated, glossed, or replaced unfamiliar (non-science) words in the test.

A Customized English Dictionary and customized glossaries were created for the science tests to study usage of lexicon reference tools that are more feasible than

published dictionaries and glossaries. Because various types of language tools are called "dictionaries" and "glossaries," we will define our terms in detail as we discuss the study's accommodation tools.

There is a significant difference between providing a Customized English Dictionary and an English-to-Spanish Glossary as an accommodation. A glossary 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

A noncompact English dictionary often offers more than just a synonym of the unknown item. Instead, full definitions (in each word's various parts of speech) are provided. If a published dictionary is provided as an accommodation for a science test, a student might use it to look up scientific words in the tests. Look at the definitions for *water* and *fuel* (*emphasis* ours):

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 might assist a student in answering the following NAEP Science questions:

Which of the following is found in every living cell?

- alcohol
- cellulose
- chlorophyll
- hemoglobin
- water

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

- nuclear reactors
- hot springs
- solar batteries
- burning of fuels
- I don't know

The Customized English Dictionary provided the full definitions of the noncontent words in the test, excluding entries for science-related terms, such as the terms above, in order to preserve the validity of the testing condition.

Customized glossaries, such as this study's English-to-English Glossary and English-to-Spanish Glossary, provide the meaning of test words only in the context in which they appear in the test. This is especially useful to those unpracticed in discriminating between dictionary entries. A customized glossary—whether mono- or bilingual—can be used more efficiently than a published English or bilingual dictionary.

These customized lexicons are more practical for national assessments when they accompany the tests and do not have to be separately provided for, so that all students have equal access to them. Therefore, students at schools without sets of dictionaries (English or bilingual) would not be penalized. In addition, customized dictionaries and glossaries look like test booklets rather than published dictionaries, to which some students show an aversion. (It is possible that they do not want to be seen using one.)

Normally, a bilingual glossary, such as this study's English-to-Spanish Glossary, is a cross-lingual list of words that appear in the test. The glossary defines the words that are used to build the context of the item. It does not serve as a reference on the subject being tested.

Monolingual lexicons, such as this study's English-to-English Glossary and Customized English Dictionary, have several advantages over bilingual glossaries and bilingual dictionaries. They serve the needs of students of all home 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. The student may not always be able to infer the word in his/her home language from a definition or synonym in English, whereas the bilingual glossary may immediately provide the student the equivalent word in the home language (Rivera & Stansfield, 1998).

Published bilingual dictionaries vary greatly. Some are word lists with one- or two-word translations and a pronunciation guide, whereas some provide pronunciation, a short translation and examples of the word in sentences. But no definition is given, unless a picture is provided. Rarely are technical specialty words included (see Appendix A). Because no unfair advantage is provided, a bilingual glossary is less of a threat to validity than a monolingual dictionary.

For some students, 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. Students who regularly use a bilingual dictionary may feel that a necessary tool of access has been withdrawn when they are not allowed to use it during an assessment. Also, students in bilingual programs may be more familiar with some words and concepts in their home language.

Observations on Linguistic Modification of Test Items

Linguistic modification of test items provides more than the lexicon decoding that glossaries and dictionaries provide. It also reduces the general linguistic complexity of test items. (See Table A1 in Appendix A for a list of features that were modified for this accommodation.) As a passive rather than an active accommodation, linguistic modification's utility is not dependent on the extra effort of looking up words. As long as the construct being measured is not changed by the process of modifying the test item language (e.g., content words and concepts are retained), the validity of the accommodation is established. Some states, such as Pennsylvania, are incorporating linguistic modification into the composition of state examination items. Thus, all students face less of the linguistic complexity that interferes with demonstrating their content knowledge.

Sampling Challenges

The study was designed with the intention to test Grade 4 and Grade 8 ELL students and their non-ELL classmates in an approximately 50-50 split between ELL

and non-ELL students. However, in some schools, the ELL students were not in the same classrooms as the non-ELL students, which in Grade 4 always contributed an additional teacher variable. We tried to limit testing in middle school science classes to the fewest number of teachers possible while still getting both ELL and non-ELL samples.

Our design faced the challenge of "dilution" of language background groups in middle schools. An elementary school often contained a large number of students from the same language background group, but a middle school tended to draw students from many language groups in the area, creating a science or ESL class that may contain only a few students from the language background group being studied. This was especially true when searching for classes containing significant numbers of Asian ELL students. In these cases, some schools brought together all of a Grade 8 science teacher's ELL students for a single period of testing and then provided us with a class of mostly non-ELL students, often studying science with the same teacher.

Another difficulty in obtaining many Grade 8 ELL participants is the greater likelihood that a student who has been in the U.S. for several years has been re-designated as an English-proficient student. According to several of the study's participating teachers and ESL coordinators, some Grade 8 students who continue to be designated as ELL students may be students with undiagnosed learning disabilities.

With the exception of Spanish speakers, it was difficult to locate Grade 8 classes containing significant numbers of ELL students with the same target language. The sample size of Grade 8 students in the main study was slightly reduced and the Grade 4 sample size slightly increased.

Procedure

The two grade levels (Grades 4 and 8) included three language background groups. To evaluate the impact of accommodation on student performance, two groups of students (ELL and non-ELL) took the assessment with no accommodation. These groups served as comparison groups. Having four accommodation conditions (a Customized English Dictionary, a Bilingual/English Glossary, a Linguistic Modification version, and the Standard condition) for both ELL and non-ELL students generated eight cells for each grade of the Spanish-speaking population. For Asian-language speaking students, there were three accommodations for both

ELL and non-ELL students, create six cells in each grade level. A total of 3,448 students were tested in this study.

As indicated above, three different forms of accommodation were used in the main study. A control or comparison group was included in the study to measure the effectiveness of accommodation strategies. In addition to a comparison group that received no accommodation, non-ELL students were sampled in this study to serve as another control or comparison group. The performance of non-ELL students who received an accommodation determined the impact of accommodation on the construct under measurement.

We provided students in the control group with the same amount of time that the groups with accommodations received, but they were asked to identify any unfamiliar words on a list of words used in the test, an additional task that is only tangential to the assessment. The results of this "survey" will be shared with the participating schools upon request. Our experience using similar testing conditions in our earlier CRESST studies informed the selection of procedures for this study.

Findings

The results of this study show that some of the accommodation strategies used were effective in increasing the performance of ELL students and reducing the performance gap between ELL and non-ELL students. The results suggest that the effectiveness of accommodation may vary across grade level.

In general, accommodations did not have a significant impact on students' performance in Grade 4. Neither ELL nor non-ELL students benefited from any of the three accommodation strategies that were used in this study. We believe the lack of effect on Grade 4 ELL students can be explained. All three accommodation strategies used in this study were language-related, and language factors have less impact on the instruction and assessment of Grade 4 students. In higher grade levels, complex language may interfere with content-based assessment. Language factors affect the assessment of ELL students in lower grades, but other factors such as poverty and parent education may be more powerful predictors of student performance in lower grades. As for non-ELL students in Grade 4, the lack of significant impact on their performance is an encouraging result because it suggests that the accommodation did not alter the construct under measurement.

Results for Grade 8 students were different than for Grade 4 students. The Linguistic Modification version of the science test had a significant impact in the

expected direction on ELL students' performance. That is, the accommodation helped ELL students increase their performance without affecting the performance of non-ELL students. This accommodation appears to be both effective and valid. It is also easy to implement; therefore, it could be provided in large-scale assessments.

Assessment of Reading Ability

An important feature of this research is 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 NAEP, designed for the general population, do not seem to discriminate between different levels of English language learners. In addition, reading assessments designed for ELL students do not seem to discriminate between non-ELL students' abilities. To measure the reading comprehension of all students, instruments for ELL students and for students who are fluent in English were combined in a three-part test. The combination test included a section of the LAS test that can discriminate among various levels of ELL students, and the multiple-choice and open-ended items from one block of the 1994 NAEP Reading assessment. The results provided a useful covariate in analyzing the science tests results.

Student Background Variables and the Accommodation Questionnaire

Background questions. The same background questions were used for Grades 4 and 8. The background questionnaire asked students about their country of origin, length of time in the United States, initial grade in which they had attended school in the United States, and whether a language other than English was spoken in the home. It also asked students to self-assess their proficiency both in English and in their home language, if not English.

The results of our analyses of the student background questionnaires indicated that student responses to the background questions are not as reliable as their responses to content-based items. We found more discrepancies in the responses of Grade 4 students than of Grade 8 students, which may result from language factors.

Accommodation follow-up questionnaire. In any accommodation study, it is extremely helpful to collect data from the accommodation recipients on the applicability and usefulness of the accommodation strategies. It is also useful to learn how much the students actually used the accommodations. To obtain data on accommodation from the students' perspectives, we developed and used accommodation follow-up questionnaires. The results are presented in Tables 34

and 35 (in the Results section), and Tables B1 to B12 (in Appendix B). Tables 34 and 35 are a representative sample of the difference between ELL and non-ELL student responses.

In the accommodation questionnaire responses, there was a trend that supports the inference that there are significant linguistic barriers in science test items. The following are a few summary statements based on questionnaire responses for both Grades 4 and 8.

- ELL students, more than non-ELL students, indicated that there were words and sentences in the science test that they did not understand (see Tables 34, 35, B1, and B2).
- On average, ELL students, more than non-ELL students, indicated that most of the science problems were difficult (see Tables B5 and B6).
- ELL students admitted to looking up words during the test more often than non-ELL students (see Tables B7 and B8).

Some of the response data may indicate that ELL students were less able to distinguish the reasons why an item was difficult. More non-ELL than ELL students felt that questions about the science they had been taught would have improved their performance (see Tables B11 and B12).

Finally, both ELL and non-ELL students strongly suggested that easier words and more time would help them with the test (see Tables B11 and B12).

Differential Impact of Primary Home Language

In order to determine whether student background variables—such as primary home language—affected performance on the accommodated assessments, we examined the students' primary home language and science test results under various accommodation conditions. It appears that students whose primary home language was neither English nor Spanish benefited the most from the Linguistic Modification version of the test.

Item-Level Analysis

We compared the percentage correct for each science test item for ELL and non-ELL students, as well as for each accommodated group compared to the standard group. The more language demand an item had, the more that Grade 8 ELL students benefited from the Linguistic Modification version of the science test.

Reading Findings

Under all four accommodation conditions, non-ELL students had higher mean reading scores than ELL students. However, there were small differences in the mean scores across the accommodation categories for both ELL and non-ELL groups. These differences in students' reading proficiency occurred in spite of our effort to randomize the accommodation conditions to avoid such differences. To control for these initial differences in reading, we adjusted the science test scores by the students' reading scores, and we then compared accommodation outcomes based on the adjusted science scores.

Implications for Policy, Practice, and Research

Following is a list of key findings and recommendations based on the analyses of this study.

Not surprisingly, students designated as ELL by their schools scored significantly lower than non-ELL students on science and reading tests.

ELL students who were better readers, as measured by the latent reading score, performed better than those who were poorer readers on science questions with high language demands.

The accommodation strategies used in this study had different results at the different grade levels. For Grade 4 students, none of the accommodation strategies significantly helped either ELL or non-ELL students. For Grade 8 students, however, the linguistic modification approach reduced the performance gap between ELL and non-ELL students significantly. The results of this study—which are consistent with the results of earlier CRESST studies—suggest that reducing language complexity helps to narrow the performance gap between non-ELL and ELL students. Thus, modifying test questions to reduce unnecessary language complexity should be a priority in the development and improvement of all large-scale assessment programs.

The literature strongly suggests that a language proficiency test must be included in any studies dealing with the assessment and accommodation of ELL students. However, lack of a single valid and reliable instrument for measuring the level of English language proficiency is noticeable. This study used a combination of instruments (multiple measures) to assess the level of students' English language proficiency. A latent composite of the measures, instead of a simple composite, was

used to combine the multiple measures. Results of the study support the use of multiple measures of English language proficiency when used as a latent composite.

This study collected self-reported data including measures of language proficiency for both English and Spanish and self-reported achievement scores. In spite of issues with the validity and reliability of self-reported data, the study found students' self-reported data to be useful and to have a moderate level of reliability and validity. We recommend collecting self-reported data on language proficiency and achievement when there is no plan to collect actual data in these areas.

Results showed that, in addition to language proficiency measures, other background factors influenced ELL performance. We recommend that both state assessments and other large-scale assessments endeavor to collect background information such as the length of time students have been living in the United States, type and amount of language spoken in the home, proficiency level in English and in student's native language, and number of years taught in both languages. This is even more important given the broad range of ELL performance we have observed.

Results also showed some level of inconsistency between school ELL classification and other language proficiency measures. National agreement is needed to create a common definition of ELL students. Nationally consistent criteria for classifying the various levels of English language proficiency are needed. Even more important are criteria for appropriate accommodations for ELL students. Comparisons between states and reporting progress will otherwise be impossible.

Feasibility considerations are important. Some accommodations, such as linguistic modification and customized glossaries and dictionaries, require additional up-front preparation time. Another cost consideration is that glossary and dictionary accommodations require additional test administration time to allow for the time spent looking up unfamiliar words. On the other hand, because national and state assessments involve large numbers of ELL students, cost-benefit analyses of various accommodations would be worthwhile. At minimum, accommodations' costs should be tracked and evaluated.

We believe that the issues concerning the feasibility of any accommodation should be at the forefront of accommodation choice decisions. We have qualitative data (e.g., teacher and student comments, and test administrator observations) that illuminate the importance of this issue. In the pilot phase of this study, we found, for

example, that there are problems and limitations in providing ELL students with a commercial dictionary as an accommodation (see Abedi, Courtney, et al., 2001). Accommodations that are effective and valid may not help if they are not feasible.

Ideally, accommodations will have no effect on non-ELL students, while reducing the language barrier for ELL students (Rivera & Stansfield, 1998; Shepard, Taylor, & Betebenner, 1998). The results of this study showed that none of the accommodations used in the study impacted the performance of non-ELL students. This finding is promising because it suggests that the validity of the assessment was not compromised by using accommodation. However, all accommodation strategies, including the accommodations used in this study, must be further examined for possible impact on the measured construct for the target population.

The results of this study, along with the findings of other CRESST studies, show that some accommodations are effective in improving the performance of ELL students. Our previous studies had indicated that a glossary plus extra time increased the performance of ELL students (Abedi, Lord, Hofstetter, et al., 2000). Other accommodation strategies such as extra time and a customized dictionary also increased the performance of ELL students (Abedi, Lord, Kim, et al., 2000). However, these latter accommodations increased the performance of non-ELL students as well, an unexpected finding that threatens the validity of the assessment. The only accommodation that increased the performance of ELL students without affecting non-ELL student performance was the linguistic modification of the assessment tool.

Though we believe that our findings are promising, ultimately there are still variables and factors that can be examined. As the ELL student population continues to grow, we must constantly find means to meet their needs. It is therefore imperative that we find effective, valid, and feasible ways to make fair assessments before inferences can be drawn from them.

References

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

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

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

Abedi, J., Leon, S., & Mirocha, J. (2001). *Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (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 strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

Abedi, J., Lord, C., Kim, C., & Miyoshi, J. (2000). *The effects of accommodations on the assessment of LEP students in NAEP.* 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., & Plummer, J. (1997). *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, & 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* (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* (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.

California Department of Education. (2000). *California demographics data.* Retrieved from: http://www.cde.ca.gov/demographics/

Chard, D. J., Simmons, D. C., & Kameenui, E. J. (1998). Word recognition: Research bases. In D. C. Simmons & E. J. Kameenui (Eds.), *What reading research tells us about children with diverse learning needs: Bases and basics* (pp. 141-167). Hillsdale, NJ: Erlbaum.

Council of Chief State School Officers. (2001). *1999-2000 State student assessment programs annual survey summary report.* Washington, DC: Author.

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

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

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

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, IL.

Gough, P. B. (1996). How children learn to read and why they fail. *Annals of Dyslexia, 46,* 3-20.

Hafner, A. L. (2001, April). *Evaluating the impact of test accommodations on test scores of LEP students and non-LEP students.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Hakuta, K., & Beatty, A. (2000). (Eds.). *Testing English-language learners in U.S. schools*. Washington, DC: National Academy Press.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development, 54*, 84-90.

Imbens-Bailey, A., & Castellon-Wellington, M. (1999, September). *Linguistic demands of test items used to assess ELL students.* Paper presented at the CRESST National Conference, Los Angeles, CA.

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 & available educational programs and services: 2000-2001 Summary report*. Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

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

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

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

Linn, R. L. (1995). *Assessment-based reform: Challenges to educational measurement*. Princeton, NJ: Educational Testing Service.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and Assessment in Teaching* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.

Liu, K., Thurlow, M., Erickson, R., Spicuzza, R., & Heinze, K. (1997). *A review of the literature on students with limited English proficiency and assessment* (Rep. No. 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, IL.

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 Publication No. 2000-473). Washington, DC: National Center for Education Statistics.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P., & Buxton, B. (1987). An alternative to multiple-choice vocabulary test. *Language Testing, 4,* 142-154.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society* (pp. 80-87). London: Centre for Information on Language Teaching and Research.

Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 200-220). Hillsdale, NJ: Erlbaum.

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

North Central Regional Educational Laboratory. (1996a). *Part 1: Assessment of students with disabilities and LEP students. The status report of the assessment programs in the U.S. State student assessment program database.* Oakbrook, IL: Author & 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: Author & 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 Publication No. 97-482). Washington, DC: National Center for Education Statistics.

O'Sullivan, C. Y., Reese, C. M., & Mazzeo, J. (1997). *NAEP 1996 science report card for the nation and the states* (NCES Publication No. 97497). Washington, DC: National Center for Education Statistics.

Read, J. (2000). *Assessing vocabulary.* New York: Cambridge University Press.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153-196). New York: Academic Press.

Rivera, C., & Stansfield, C. W. (1998). Leveling the playing field for English language learners: Increasing participation in state and local assessments through accommodations. Retrieved 22 September 2003 from: http://ceee.gwu.edu/standards_assessments/researchLEP_accommodcase.htm

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

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.* Washington, DC: Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. EDO-TM-97-02)

Roeber, E., Bond, L., & Connealy, S. (1998). *Annual survey of state student assessment programs. Vol. I, II. Data on 1996-97 statewide student assessment programs, Fall 1997.* Washington, DC: Council of Chief State School Officers.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's grade 4 mathematics performance assessment* (CSE Tech. Rep. No. 486). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Thurlow, M. L. (2001, April). *The effects of a simplified-English dictionary accommodation for LEP students who are not literate in their first language.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

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

Appendix A
Methodology

Post-Pilot Modifications to the Study Design and Instruments

Before finalizing the design of the main study, experts in the field were presented with a revised project plan that included descriptions of the accommodations and the instruments, a chart of the sampling plan, and the accommodation distribution plan. Based on their responses to our questions and our own findings, we made several adaptations of the study design.

Adaptation of sampling. New sampling goals were set to diversify the Spanish-language population and better target Asian-language speakers. These goals were adapted as site participation changed as described below.

1. We confined the language background groups studied to Spanish, Chinese, and Korean. To get a large enough Korean ELL sample, we applied for testing in Site 6.

2. Data collection in Site 6 received district approval and school interest, but was stalled at the subdistrict level, despite the support of board members from the local Korean community. This resulted in our replacing the Korean language target with a broader "other Asian" language target (Site 4).

3. Data collection in Site 7 was cancelled when the district required active consent by parents. This eliminated testing significant numbers of Spanish-speaking students originally from Caribbean cultures. The sample sizes of Mexican and Central American students were subsequently increased.

4. When Site 8 was consulted for access to significant numbers of Chinese-speaking students, the district's continuing legal dispute about the testing of ELL students caused the district to demur, lest they appear to condone the testing of ELL students.

5. When access to specific Asian-language populations seemed all but cut off, Site 4 agreed to participate in both the Grade 4 and Grade 8 portions of the study at the end of the school year, thus providing access to a broad range of Asian ELL students. Note that this meant that some Grade 4 students were tested at the end of the school year.

6. We increased the *n* size for the more accessible student groups, namely Grade 4 and Spanish-speaking students.

Accommodation modifications. Based on the pilot study findings and some suggestions by our panel of experts, we made the following modifications in devising the main study accommodations.

1. We eliminated the bilingual resource for Chinese and Korean students because so many were not literate in their home languages. (During the pilot, when Chinese or Korean bilingual dictionaries were distributed

among Grade 4 ELL students, some students remarked that they could not read their home language.)

2. We replaced the published dictionary and bilingual glossaries with the more feasible glossaries: the Customized English Dictionaries and the English-to-Spanish Glossaries.

3. we devised an English-to-English glossary for non-ELL students being tested in classrooms where some Spanish-ELL students would be using bilingual dictionaries.

Modifications to the tests. To eliminate some internal consistency problems with a few items (on the Grade 8 science test), and to balance out the score value of multiple-choice vs. open-ended items on each test, the science tests were revised. To balance the life science content in the Grade 4 test, three environmental science and four physical science items, all multiple-choice, were added from the TIMSS released science items, many of which test scientific processes and investigation. One life science multiple-choice item from TIMSS replaced a NAEP item with a poor item-total correlation. By balancing the science content, the test more strongly adhered to the NAEP framework. Open-ended items in the non-life-sciences were not added to replace any of the NAEP open-ended items because the pilot study test's thematic life science block of NAEP items begs to be administered intact.

The reading tests were modified in order to obtain a better distribution of scores for both ELL and non-ELL students. This was achieved by combining a LAS Fluency section and one block of the NAEP Reading subscale used in the pilot.

Class roster changes. The main study utilized an expanded class roster which asks the instructor/school to provide more background features about each participant such as gender, socio-economic indicator (as determined by participation in the free/reduced lunch program), years in the ELL program, and parents' level of education. These student background variables were compared to student test results.

Student background questionnaire simplified. Grade 4 ELL students did not answer the student background questionnaire with ease and often had to be queried the next day about answers that did not make sense. For the main study, the entire form was linguistically modified for easier comprehension. In our discussion with accommodation research experts, it was decided that the students should be asked to self-assess their science education, state which areas of science they are learning,

and to describe the type of science class they are enrolled in. Three items were added to the student background questionnaire for this.

Accommodation questionnaire revised. This questionnaire was added to the pilot study when it seemed that students were not taking advantage of the English and bilingual dictionaries. Student responses to this questionnaire informed the study design as well as helped us revise the questionnaire for use in the main study. Open-ended responses to "What else would make it easier for you to understand the questions in the test?" became the possible multiple-choice responses in the main study's questionnaire.

Teacher questionnaire expanded. Receiving language accommodation *only* during standardized test taking is not the best introduction to accommodation. The use of lexicon tools, for example, is more valuable if the tool itself is familiar to the student. The scope of the teacher questionnaire was increased to include a second set of pages asking about the instructor's experience with providing specific accommodations throughout the school year in both instructional and test-taking situations. Just as with the student questionnaire, there were new science lesson questions in the revised teacher questionnaire.

Data entry automated. The main study instruments were reformatted in TELEform software (Cardiff Software, Inc., Vista, CA) so that all instruments, including open-ended item score sheets, could be scanned, processed and entered into the database even more accurately than before. An additional benefit to automated data entry is that test and questionnaire results were available for review on an ongoing basis throughout the main study's data collection. A drawback to using "scannable" tests and questionnaires with schoolchildren is that students sometimes doodled in identification barcodes on the test page corners, thus creating scanning errors that had to be sought out, tracked down and repaired.

Table A1

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 nontechnical meanings
- science keywords misinterpreted
- derivatives of content words

Table A2

Chart of Selected *Content* Words in Pilot Science Tests and Their Appearance in English and Bilingual Dictionaries / Glossaries

Word in test	Where in test	Grade of test	Chinese dictionary	Korean dictionary	Spanish dictionary	English dictionary
Atom	Question	8B	Y	Y	N	Y
Caterpillar	Question	4	Y	N	Y	Y
Cellulose	Answer	8A	N	N	Y	Y
Enzyme	Question	8B	Y	N	N	Y
Fuel	Answer	8B	Y	Y	Y	Y Advantage
Gasoline	Question	4	Y	Y	Y	Y Advantage
Grasshopper	Question	4	Y	Y	N	Y
Half-life	Question	8A	N	N	N	Y Advantage
Hemoglobin	Answer	8A	N	N	N	Y
Insulation	Answer	8A,B	Y	Y Insulate	N	Y Advantage
Mirror	Question	8A,B	Y	Y	Y	Y Advantage
Mitochondria	Question	8B	N	N	N	Y Advantage
Moon	Question	4	Y	Y	Y	Y Advantage
Newborn	Question	4	N	N	Y	Y
Oil	Answer	4	Y	Y	Y	Y Advantage
Protein	Question	8B	Y	N	N	Y
Pupa	Question	4	N	N	N	Y
Salamander	Question	4	N	N	N	Y
Tectonic plate	Answer	8A	N	N	N	Y Advantage

Note. Y = Yes, appears in dictionary. N = No, does not appear. "Advantage" indicates where the dictionary definition might provide an unfair advantage in answering the question.

Table A3

Chart of Selected *Non-Content* Words in Pilot Science Tests and Their Appearance in Accommodation Tools

Word in test	Where in test	Grade of test	Chinese dictionary	Korean dictionary	Spanish dictionary	English dictionary
Arthritis	Question	8A,B	Y	N	N	Y
Batteries	Answer	4	Y	N	Y	Y
Boundary	Question	8A,B	Y	Y	Y	Y Advantage
Clump	Story	4	Y	Y	N	Y
Different	Question	4	Y	Y	N Differ	Y
Dune	Answer	4	Y	Y	N	Y
Equation	Question	8A	Y	N	N Equate	Y
Interaction	Question	8A	Y	Y	N	Y
Resources	Question	8A,B	Y	Y	N	Y
Sand	Answer	8A	Y	Y	Y	Y Advantage
Sediment	Answer	8A,B	Y	Y	N	Y
Smog	Question	4	Y	Y	N	Y Advantage
Solar	Answer	8B	Y	Y	N	Y
Storm window	Answer	8A,B	P Storm	P Storm	N	Y Advantage
System	Answer	8B	Y	Y	Y	Y Advantage

Note. Y = Yes, appears in dictionary. N = No, does not appear.

Appendix B
Results

Table B1

Grade 4 Mean of "In the science test I did not understand..."

Accommodation	ELL	Non-ELL	Row total
Standard condition	1.78 (*SD* = .63; *n* = 245)	1.58 (*SD* = .58; *n* = 211)	1.69 (*SD* = .62; *n* = 456)
Customized Dictionary	1.75 (*SD* = .65; *n* = 253)	1.65 (*SD* = .61; *n* = 227)	1.70 (*SD* = .63; *n* = 480)
Bilingual / English Glossary	1.80 (*SD* = .65; *n* = 123)	1.70 (*SD* = .63; *n* = 97)	1.75 (*SD* = .64; *n* = 220)
Linguistic Modification	1.78 (*SD* = .65; *n* = 262)	1.62 (*SD* = .63; *n* = 233)	1.71 (*SD* = .64; *n* = 495)
Column total	1.77 (*SD* = .64; *n* = 883)	1.63 (*SD* = .61; *n* = 768)	1.71 (*SD* = .63; *n* = 1651)

Note. 1 = *I had no problem*, 2 = *some* sentences, 3 = *many sentences.*

Table B2

Grade 8 Mean of "In the science test I did not understand..."

Accommodation	ELL	Non-ELL	Row total
Standard condition	1.96 (*SD* = .57; *n* = 144)	1.75 (*SD* = .57; *n* = 178)	1.84 (*SD* = .58; *n* = 322)
Customized Dictionary	1.97 (*SD* = .53; *n* = 149)	1.78 (*SD* = .62; *n* = 177)	1.87 (*SD* = .59; *n* = 326)
Bilingual / English Glossary	1.94 (*SD* = .57; *n* = 97)	1.75 (*SD* = .58; *n* = 108)	1.84 (*SD* = .58; *n* = 205)
Linguistic Modification	1.99 (*SD* = .59; *n* = 168)	1.74 (*SD* = .58; *n* = 194)	1.85 (*SD* = .59; *n* = 362)
Column total	1.97 (*SD* = .56; *n* = 558)	1.75 (*SD* = .59; *n* = 657)	1.85 (*SD* = .59; *n* = 1215)

Note. 1 = *I had no problem*, 2 = *some* sentences, 3 = *many sentences.*

Table B3

Grade 4 Mean of "Most science problems were..."

Accommodation	ELL	Non-ELL	Row total
Standard condition	2.06 (*SD* = .91; *n* = 245)	2.01 (*SD* = .78; *n* = 211)	2.04 (*SD* = .85; *n* = 456)
Customized Dictionary	2.21 (*SD* = .87; *n* = 253)	1.97 (*SD* = .78; *n* = 227)	2.10 (*SD* = .84; *n* = 480)
Bilingual/English Glossary	2.16 (*SD* = .91; *n* = 123)	2.05 (*SD* = .83; *n* = 98)	2.11 (*SD* = .87; *n* = 221)
Linguistic Modification	2.11 (*SD* = .90; *n* = 261)	1.98 (*SD* = .83; *n* = 233)	2.05 (*SD* = .87; *n* = 494)
Column total	2.13 (*SD* = .89; *n* = 882)	2.00 (*SD* = .80; *n* = 769)	2.07 (*SD* = .86; *n* = 1651)

Note. 1 = *very easy*, 2 = *easy*, 3 = *difficult*, 4 = *very difficult*.

Table B4

Grade 8 Mean of "Most science problems were..."

Accommodation	ELL	Non-ELL	Row total
Standard condition	2.67 (*SD* = .66; *n* = 146)	2.18 (*SD* = .63; *n* = 178)	2.40 (*SD* = .69; *n* = 324)
Customized Dictionary	2.55 (*SD* = .72; *n* = 148)	2.27 (*SD* = .66; *n* = 175)	2.40 (*SD* = .70; *n* = 323)
Bilingual/English Glossary	2.45 (*SD* = .76; *n* = 97)	2.26 (*SD* = .64; *n* = 107)	2.35 (*SD* = .72; *n* = 204)
Linguistic Modification	2.51 (*SD* = .72; *n* = 169)	2.21 (*SD* = .61; *n* = 194)	2.35 (*SD* = .68; *n* = 363)
Column total	2.55 (*SD* = .72; *n* = 560)	2.23 (*SD* = .64; *n* = 654)	2.38 (*SD* = .69; *n* = 1214)

Note. 1 = *very easy*, 2 = *easy*, 3 = *difficult*, 4 = *very difficult*.

Table B5

Grade 4 Mean of "Did you look up words during the test?"

Accommodation	ELL	Non-ELL	Row total
Standard condition	2.39 (*SD* = 1.27; *n* = 158)	2.20 (*SD* = 1.17; *n* = 112)	2.31 (*SD* = 1.23; *n* = 270)
Customized Dictionary	2.25 (*SD* = 1.27; *n* = 244)	2.00 (*SD* = 1.20; *n* = 220)	2.13 (*SD* = 1.24; *n* = 464)
Bilingual/English Glossary	2.34 (*SD* = 1.29; *n* = 111)	2.34 (*SD* = 1.32; *n* = 89)	2.34 (*SD* = 1.30; *n* = 200)
Linguistic Modification version	2.59 (*SD* = 1.32; *n* = 180)	2.19 (*SD* = 1.20; *n* = 131)	2.42 (*SD* = 1.28; *n* = 311)
Column total	2.39 (*SD* = 1.29; *n* = 693)	2.14 (*SD* = 1.22; *n* = 552)	2.28 (*SD* = 1.26; *n* = 1245)

Note. Only those students who received a glossary should have responded to this question—Results are difficult to interpret. 1 = *No,* 2 = *Occasionally,* 3 = *Half the problems,* 4 = *Often,* 5 = *Every problem.*

Table B6

Grade 8 Mean of "Did you look up words during the test?"

Accommodation	ELL	Non-ELL	Row total
Standard condition	2.23 (*SD* = 1.29; *n* = 97)	1.75 (*SD* = 1.18; *n* = 92)	1.99 (*SD* = 1.26; *n* = 189)
Customized Dictionary	2.01 (*SD* = 1.26; *n* = 138)	1.54 (*SD* = 0.95; *n* = 169)	1.75 (*SD* = 1.12; *n* = 307)
Bilingual/English Glossary	1.96 (*SD* = 1.10; *n* = 111)	1.56 (*SD* = 1.03; *n* = 89)	1.75 (*SD* = 1.08; *n* = 200)
Column total	2.06 (*SD* = 1.22; *n* = 331)	1.60 (*SD* = 1.03; *n* = 363)	1.82 (*SD* = 1.15; *n* = 694)

Note. The Linguistic Modification version did not allow responses to this item. 1 = *No,* 2 = *Occasionally,* 3 = *Half the problems,* 4 = *Often,* 5 = *Every problem.*

Table B7

Grade 4 Mean of "Did the glossary help?"

Accommodation	ELL	Non-ELL	Row total
Standard condition	1.97 (*SD* = .75; *n* = 136)	1.87 (*SD* = .78; *n* = 99)	1.93 (*SD* = .76; *n* = 235)
Customized Dictionary	1.94 (*SD* = .74; *n* = 238)	1.83 (*SD* = .76; *n* = 213)	1.89 (*SD* = .75; *n* = 451)
Bilingual/English Glossary	2.10 (*SD* = .75; *n* = 102)	2.13 (*SD* = .78; *n* = 86)	2.11 (*SD* = .76; *n* = 188)
Linguistic Modification	2.05 (*SD* = .79; *n* = 137)	1.94 (*SD* = .82; *n* = 98)	2.00 (*SD* = .80; *n* = 235)
Column total	2.00 (*SD* = .76; *n* = 613)	1.91 (*SD* = .79; *n* = 496)	1.96 (*SD* = .77; *n* = 1109)

Note. Only those students who received a glossary should have responded to this question. Results are difficult to interpret. 1 = *No,* 2 = *Yes some,* 3 =*Yes a lot*.

Table B8

Grade 8 Mean of "Did the glossary help?"

Accommodation	ELL	Non-ELL	Row total
Standard condition	1.57 (*SD* = .64; *n* = 97)	1.22 (*SD* = .44; *n* = 79)	1.41 (*SD* = .59; *n* = 176)
Customized Dictionary	1.68 (*SD* = .63; *n* = 137)	1.37 (*SD* = .57; *n* = 167)	1.51 (*SD* = .62; *n* = 304)
Bilingual/English Glossary	1.71 (*SD* = .64; *n* = 89)	1.49 (*SD* = .62; *n* = 104)	1.59 (*SD* = .64; *n* = 193)
Column total	1.65 (*SD* = .64; *n* = 323)	1.37 (*SD* = .57; *n* = 350)	1.51 (*SD* = .62; *n* = 673)

Note. The Linguistic Modification version did not allow responses to this item. 1 = *No,* 2 = *Yes some,* 3 =*Yes a lot*.

Table B9

Grade 4 Percent of Selected “If the glossary was not helpful, why not?”

	ELL	Non-ELL	Row total
I had a glossary, but I already understood all the words.	26.7 (*n* = 154)	23.6 (*n* = 152)	25.1 (*n* = 306)
The words I needed help with were not there.	19.8 (*n* = 114)	21.6 (*n* = 139)	20.8 (*n* = 253)
The definitions in the glossary were hard to understand.	7.1 (*n* = 41)	14.8 (*n* = 95)	11.2 (*n* = 136)
I didn't have a glossary.	46.4 (*n* = 267)	40.0 (*n* = 257)	43.0 (*n* = 524)
Column total	100.0 (*n* = 576)	100.0 (*n* = 643)	100.0 (*n* = 1219)

Note. Only those students who received a glossary should have responded to this question. Results are difficult to interpret.

Table B10

Grade 8 Percent of Selected “If the glossary was not helpful, why not?”

	ELL	Non-ELL	Row total
I had a glossary, but I already understood all the words.	18.4 (*n* = 72)	25.9 (*n* = 131)	22.6 (*n* = 203)
The words I needed help with were not there.	23.5 (*n* = 92)	20.9 (*n* = 106)	22.1 (*n* = 198)
The definitions in the glossary were hard to understand.	12.3 (*n* = 48)	4.7 (*n* = 24)	8.0 (*n* = 72)
I didn't have a glossary.	45.8 (*n* = 179)	48.4 (*n* = 245)	47.3 (*n* = 424)
Column total	100.0 (*n* = 391)	100.0 (*n* = 506)	100.0 (*n* = 897)

Table B11

Grade 4 Percent of Selected "To make it easier for me to understand the science questions please give me..."

	ELL	Non-ELL	Row total
The same problems, but easier words	39.0 (*n* = 346)	39.5 (*n* = 307)	39.2 (*n* = 653)
The same problems, but simpler sentences	25.5 (*n* = 226)	25.8 (*n* = 201)	25.6 (*n* = 427)
Some simpler science problems	26.9 (*n* = 239)	29.6 (*n* = 230)	28.2 (*n* = 469)
Questions about science that I was taught	28.2 (*n* = 250)	35.3 (*n* = 275)	31.5 (*n* = 525)
More pictures in the test	31.6 (*n* = 281)	31.2 (*n* = 243)	31.5 (*n* = 524)
An English dictionary or glossary	29.7 (*n* = 264)	23.9 (*n* = 186)	27.0 (*n* = 450)
A translation dictionary for my first language	11.4 (*n* = 101)	6.0 (*n* = 47)	8.9 (*n* = 148)
Some words translated into my first language	11.7 (*n* = 104)	6.6 (*n* = 51)	9.3 (*n* = 155)
A test written in my first language	12.8 (*n* = 114)	6.2 (*n* = 48)	9.7 (*n* = 162)
Questions read aloud	13.9 (*n* = 123)	14.8 (*n* = 115)	14.3 (*n* = 238)
More time	42.7 (*n* = 379)	40.2 (*n* = 313)	41.5 (*n* = 692)

Table B12

Grade 8 Percent of Selected "To make it easier for me to understand the science questions please give me..."

	ELL	Non-ELL	Row total
The same problems, but easier words	32.7 (*n* = 201)	28.5 (*n* = 199)	30.5 (*n* = 400)
The same problems, but simpler sentences	29.6 (*n* = 182)	33.1 (*n* = 231)	31.5 (*n* = 413)
Some simpler science problems	26.8 (*n* = 165)	27.2 (*n* = 190)	27.0 (*n* = 355)
Questions about science that I was taught	31.7 (*n* = 195)	47.6 (*n* = 332)	40.1 (*n* = 527)
More pictures in the test	21.1 (*n* = 130)	21.9 (*n* = 153)	21.6 (*n* = 283)
An English dictionary or glossary	18.0 (*n* = 111)	18.3 (*n* = 128)	18.2 (*n* = 239)
A translation dictionary for my first language	8.5 (*n* = 52)	1.7 (*n* = 12)	4.9 (*n* = 64)
Some words translated into my first language	8.9 (*n* = 55)	2.6 (*n* = 18)	5.6 (*n* = 73)
A test written in my first language	7.2 (*n* = 44)	1.7 (*n* = 12)	4.3 (*n* = 56)
Questions read aloud	7.5 (*n* = 46)	6.0 (*n* = 42)	6.7 (*n* = 88)
More time	19.8 (*n* = 122)	17.0 (*n* = 119)	18.4 (*n* = 241)