

Accommodations for English Language Learner Students: the Effect of Linguistic Modification of Math Test Item Sets

Final Report



Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets

June 2010

Authors:

Edynn Sato, Principal Investigator
WestEd

Stanley Rabinowitz, Principal Investigator
WestEd

Carole Gallagher, Senior Research Associate
WestEd

Chun-Wei Huang, Senior Research Analyst
WestEd

Project Officer:

Ok-Choon Park
Institute of Education Sciences

NCEE 2009-4079
U.S. Department of Education



U.S. Department of Education

Arne Duncan

Secretary

Institute of Education Sciences

John Q. Easton

Director

National Center for Education Evaluation and Regional Assistance

Rebecca A. Maynard

Commissioner

June 2010

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract ED-06C0-0014 with Regional Educational Laboratory West administered by WestEd.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the report.

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. While permission to reprint this publication is not necessary, the citation should read: Sato, E., Rabinowitz, S., Gallagher, C. Huang, C.-W. (2010). *Accommodations for English language learner students: the effect of linguistic modification of math test item sets*. (NCEE 2009-4079). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the Institute of Education Sciences website at <http://ncee.ed.gov> and the Regional Educational Laboratory Program website at <http://edlabs.ed.gov>.

Alternate Formats Upon request, this report is available in alternate formats, such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of potential conflict of interest

The research team for this study was based at Regional Educational Laboratory West administered by WestEd. Neither the authors nor WestEd and its key staff have financial interests that could be affected by the findings of this study. No one on the 11-member Technical Working Group, convened annually by the research team to provide advice and guidance, has financial interests that could be affected by the study findings.*

* Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

Contents

Acknowledgments	vii
Executive summary	1
1. Study overview	5
Study context	5
Description of the accommodation (linguistic modification)	8
Research questions	8
Overview of study design	10
Structure of the report	12
2. Study design, study sample, and item set development	13
Study design	13
Overview of study steps	14
Sample recruitment	15
Participant flow	18
Considerations related to student sample	20
Item set development and administration	22
Item refinement based on cognitive interviews	26
Item refinement based on pilot test data	28
3. Implementation of the accommodation (linguistic modification) and methods for analysis	32
Operational administration of the item sets	32
Scoring and analysis of data	32
Missing data	39
4. Study results	40
Primary analysis: differences in the impact of linguistic modification across student subgroups	40
Secondary analyses	45
Summary of key findings from primary and secondary analyses	50
5. Interpretation of key findings, study challenges, and direction for future research	52
Interpretation of findings from the primary analysis: interaction between student subgroup and item set	52
Interpretation of findings from secondary analyses: impact of linguistic modification on construct assessed	53
Challenges related to the study context and design	54
Challenges related to item selection and item set development	54
Other directions for future research	55
References	57
Appendix A. Power analysis for primary research questions	67
Appendix B. Operational test administration manual	68
Appendix C. Student Language Background Survey	76
Appendix D. Guide for developing a linguistically modified assessment	80

Appendix E. Workgroup training materials	91
Appendix F. Overview and protocol for cognitive interviews	98
Appendix G. Item parameter estimates for IRT models.....	108
Appendix H. Descriptive statistics from four scoring approaches.....	112
Appendix I. ANOVA findings across four scoring approaches.....	116
Appendix J. Cross-approach comparisons.....	119
Appendix K. Results of the classical item-level analyses.....	122
Appendix L. Summary of differential item functioning findings.....	125
Appendix M. Exploratory factor analysis results	127
Appendix N. Operational item set—original.....	132
Appendix O. Operational item set—linguistically modified.....	133

Tables

Table 1. Overview of data collection activities related to item development and refinement	11
Table 2. Overview of data collection activities related to impact analyses	12
Table 3. Timeline for study activities, January 2007–January 2009	15
Table 4. Description of study sample, by school.....	17
Table 5. Overview of item screening process.....	23
Table 6. Mean item set scores and score differences by scoring method, item set, and student subgroup.....	41
Table 7. Post-hoc comparison of interaction effect (based on 1-PL model)	43
Table 8. Mean percent correct (item <i>p</i> -value) and the associated standard deviation across all items, by student subgroup and item set	45
Table 9. Internal consistency reliability coefficient, by student subgroup and item set	46
Table 10. Correlations between item set raw score totals and state standardized math achievement test score, by grade, for non–English language learner students who were proficient in English language arts	50
Table A1. Full study design sample.....	67
Table D1. Linguistic modification guidelines and strategies.....	86
Table E1. Linguistic skills	95
Table E2. Academic language functions	96
Table G1. Item parameter estimates for 1-PL model.....	109
Table G2. Item parameter estimates for 2-PL model.....	110
Table G3. Item parameter estimates from 3-PL model.....	111
Table H1. Mean math raw scores, by grade, student subgroup, and item set.....	112
Table H2. Mean theta estimates from the 1-PL model, by grade, student subgroup, and item set	113
Table H3. Mean theta estimates from the 2-PL model, by grade, student subgroup, and item set	114

Table H4. Mean theta estimates from the 3-PL model, by grade, student subgroup, and item set	115
Table I1. Analysis of variance for linguistic modification effects on student subgroups (based on raw scores).....	116
Table I2. Analysis of variance for linguistic modification effects on student subgroups (based on 1-PL model).....	117
Table I3. Analysis of variance for linguistic modification effects on student subgroups (based on 2-PL model).....	117
Table I4. Analysis of variance for linguistic modification effects on student subgroups (based on 3-PL model).....	118
Table J1. Evaluation of model fit, by item set, for item response theory models	120
Table K1. Item-level statistics for original item set.....	123
Table K2. Item-level statistics for linguistically modified item set.....	124
Table L1. Summary of findings from analysis of differential item functioning, NEP students versus EP students	125
Table L2. Summary of findings from analysis of differential item functioning, EL students versus EP students	126
Table M1. Estimated factor loadings based on one-factor solution, by item set and student subgroup	127

Figures

Figure 1. Study design	13
Figure 2. Consolidated Standards of Reporting Trials flow diagram of school recruitment.....	16
Figure 3. Consolidated Standards of Reporting Trials flow diagram of student participants	20
Figure 4. Profile plot of cell means, by item set and student subgroup (based on 1-PL model).....	44
Figure M1. Scree plot for non-English language learner students who are proficient in English language arts, taking original item set	128
Figure M2. Scree plot for non-English language learner students who are not proficient in English language arts, taking original item set	128
Figure M3. Scree plot for English language learner students taking original item set.....	129
Figure M4. Scree plot for non-English language learner students who are proficient in English language arts, taking linguistically modified item set	130
Figure M5. Scree plot for non-English language learner students who are not proficient in English language arts, taking linguistically modified item set	130
Figure M6. Scree plot for English language learner students taking linguistically modified item set.....	131

Acknowledgments

Throughout the planning and design phases of this study, the authors sought technical advice from members of a technical work group, with whom they consulted on methodological issues (experimental design, sampling frame, power estimates, data collection and analysis, and reporting strategies) and from whom they requested feedback on associated products (protocols and instruments). These consultations were intended to ensure the study's appropriate technical rigor and relevance so that findings would be trustworthy and useful. Education survey experts contributed recommendations for improving the Student Language Background Survey so that items would be clear and unambiguous.

Members of the technical workgroup included Dr. Jamal Abedi, University of California, Davis; Dr. Lloyd Bond, Carnegie Foundation for the Advancement of Teaching; Dr. Geoffrey Borman, University of Wisconsin; Dr. Brian Flay, Oregon State University; Dr. Tom Good, University of Arizona; Dr. Corinne Herlihy, MDRC; Dr. Joan Herman, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles; Dr. Heather Hill, Harvard University; Dr. Roger Levine, American Institutes for Research; Dr. Juliet Shaffer, University of California, Berkeley; and Dr. Jason Snipes, Academy for Educational Development.

The authors also solicited comments from educators who work with general and special student populations, math content experts, state-level test developers, and measurement specialists. These included Patricia Armstrong, WestEd; Dr. Patrick Callahan, WestEd; Joanne DaLuz, WestEd; Dr. Paula Diamanti, Second Language Testing, Inc.; Dr. Cassandra Hawley, University of California, Davis; Dr. Loretta Kelley, Kelley, Petterson, and Associates, Inc.; Dr. Robert Lissitz, University of Maryland-College Park; Dr. Susan Porter, University of California, Davis; Dr. Charlene Rivera, Second Language Testing, Inc.; and Dr. Charles Stanfield, Second Language Testing, Inc.

Senior advisory staff with whom the authors consulted during the study were associated with the following professional organizations and agencies: Assessment and Accountability Comprehensive Center; Council of Chief State School Officers; National Center for Research on Evaluation, Standards, and Student Testing; and National Council of Teachers of Mathematics.

The authors also would like to thank Dr. Andrea Lash for her guidance on research design and Dr. Carol Whang for her support during recruitment and data collection.

Finally, the authors would like to express appreciation to reviewers from Mathematica's Analytic Technical Support team for their preliminary reviews of the manuscript and to anonymous external reviewers who provided helpful feedback during development of this report.

Executive summary

Both theory and research suggest that students, especially English language learner (EL) students, could be constrained in showing what they know and can do in mathematics if the test items used to assess their achievement are measuring factors other than their math-related knowledge and skills (Abedi, Hofstetter, & Lord 2004; Butler & Stevens 2001). This interference from construct-irrelevant factors has been found to be most pronounced for students with limited English proficiency, such as EL students and non-EL students who fail to achieve proficiency on state English language arts (ELA) assessments (Abedi 2001). Research has shown that math test items can be linguistically modified to reduce the complexity of the language used, without altering the construct (for example, math understanding) being assessed, thereby enabling student access to the tested content (Abedi & Lord 2001; Abedi, Courtney, & Leon 2003).

This study examined the effect of linguistic modification on middle school students' ability to show what they know and can do on math assessments. To do so, two item sets with 25 multiple-choice items each were developed, one containing original math items and one containing these items with linguistic modifications. Items were selected from two content strands: (1) measurement and (2) number sense and operations. Efforts were made to ensure that both sets of math test items met stringent guidelines for grade and population appropriateness, content rigor, and standardized administration. In developing the two item sets, researchers solicited input from experts and collected data through cognitive interviews and pilot testing.

The two sets of math items (original and linguistically modified) were administered to three subgroups of students in grades 7 and 8 who differed in their English language proficiency (ELP) and ELA skills: EL students, non-ELA-proficient non-EL (NEP) students, and ELA-proficient non-EL (EP) students. Item sets were assigned to students within each class at random, with approximately half of each subgroup receiving the original set of items and half receiving the linguistically modified set.

Participating districts were asked to provide archived data about each tested student, including students' most recent test scores in ELA and math and EL students' level of ELP. The ELA scores were used to distinguish EP students from NEP students. The math scores were used in correlational analyses as a proxy for math ability for EP students. The ELP scores were used to identify sampled students as EL students or non-EL students and to determine each EL student's level of English proficiency.

Key findings

The primary research question asked if the effect of linguistic modification on students' performance on the two sets of math items (original and linguistically modified) varied across the three subgroups of students (EL, NEP, and EP students). If so, did the linguistic modification improve student math performance for the EL and NEP students relative to the

EP students? If the linguistic modification increased the accessibility of EL and NEP students to the assessed math content, one would expect a significant score gain between the original and linguistically modified item set for EL and NEP students while no or minimal effect for EP students. Student math performance was evaluated using four scoring approaches commonly used by states in analyzing performance data from statewide testing. These include computing raw scores for each item set (that is, summed number correct) as well as estimating student math understanding (theta estimates) from three item response theory (IRT) models: a one-parameter logistic IRT model (1-PL), a two-parameter logistic IRT model (2-PL), and a three-parameter logistic IRT model (3-PL).² The IRT-based scoring approaches used a random groups equating design to equate the original and linguistically modified item sets, whereby the mean and the standard deviation of math understanding was assumed to be the same across item sets for EP students. This constraint, which is necessary to make meaningful comparisons across item sets, assumes that there is no effect of linguistic modification for the EP group. Then, for each approach and for each student subgroup, the mean score difference (that is, the difference between the mean score for the original item set and the mean score for the linguistically modified item set) was calculated.

- Differences across EL, NEP, and EP students in the effects of linguistic modification on students' math performance depended on the scoring approach used (that is, how scores for each student were calculated or estimated). When scores were constructed based on the 1-PL IRT model, a significant difference in theta scores on the two sets of items (original and linguistically modified) was detected across student subgroups (between EL students and EP students, in particular). This small but significant effect was not detected in the analyses based on raw scores or theta estimates from the 2-PL or 3-PL models.
- Despite inconsistent significance test results across the four approaches, the mean differences in performance on the two item sets for each student subgroup showed a consistent trend—the mean difference in performance on the two items sets was greatest for EL students, followed by NEP students. For EP students, the difference in raw scores on the original and linguistically modified item set was very close to zero (less than 0.01 standard deviation units).³
- The effect size, or magnitude of the difference in mean scores between the original item set and the linguistically modified item set for EL students, was 0.16 standard deviation units using a raw score metric and 0.17, 0.12, and 0.09 standard deviation units when the scores were derived using the 1-, 2-, and 3-PL models, respectively.⁴ Regardless of

² In a 3-PL model, each item is described by three types of parameters: item difficulty, item discrimination, and a pseudo-guessing parameter. If guessing is assumed to be minimal (or none), then the 3-PL model becomes the 2-PL model. If the item discriminating power is assumed to be equal across test items under the 2-PL model, then it becomes the 1-PL model.

³ This difference was set to zero in the analyses based on the IRT models. See chapter 3 for a detailed discussion.

⁴ This standardized mean difference is derived by subtracting the mean ability estimate for the original item set from the mean ability estimate for the linguistically modified item set and dividing by the standard deviation of the total student group (all three subgroups pooled) for the original item set. The total group standard deviation is

approach, the effect sizes found in this study were greater than the mean effect size for linguistic modification (0.03) reported in a recent meta-analysis studying accommodations for EL students (Kieffer, Lesaux, Rivera, & Francis, 2009).

- Because there is no universal guideline for evaluating the practical importance of a standardized effect size estimate for an educational intervention, it also is useful to compare this estimate to another empirical benchmark that reflects changes in student academic achievement. A standardized difference of 0.17 based on the 1-PL model, for example, is more than half the magnitude of growth in achievement that might be expected from *one full year of schooling* (.32), as measured by a standardized test (Hill et al. 2008).

Additional analyses were conducted to more fully examine potential accommodation effects. These secondary analyses were conducted based on raw scores exclusively—not based on IRT-based estimates. The secondary research questions, analyses, and associated findings are summarized below:

- *Differential item functioning.* To determine whether items were equally appropriate for assessing the targeted math construct across student subgroups, a differential item functioning (DIF) analysis using the Mantel-Haenszel procedure (Holland & Thayer 1988) was conducted. An item showing DIF may require additional review by experts to ensure no construct-irrelevant factor (such as unnecessary English language complexity) was introduced that might advantage one student subgroup over another. The DIF analysis between EL students and EP students indicated that one item exhibited DIF in the original item set and two items exhibited DIF in the linguistically modified item set; the DIF analysis between NEP students and EP students revealed that no item in either item set exhibited DIF. Subsequent review of these items by content, population, and assessment experts did not find evidence of bias in either item set.
- *Factor analyses.* For each item set within each student subgroup, a series of exploratory factor analyses (EFA) examined the number of underlying factors in the two item sets. Regardless of item set (original or linguistically modified) or student subgroup (EL, NEP, or EP), the EFA indicated that there was only one dominant factor (math understanding) underlying the data. The EFA results also served as the foundation for a series of nested confirmatory factor analyses (CFAs), which tested for differences in measurement structure (items and the underlying factor relationship as represented by factor loadings) across student subgroups. Overall, the CFA results suggested that: (1) although one dominant factor (math understanding) appeared to underlie the item sets for each student subgroup, the item sets had a different measurement structure for each of the student subgroups, and this held for both the original and the linguistically modified item sets; (2) the relationship between the underlying (dominant) factor and the items appeared to be weaker for EL students and, to a lesser extent, for NEP students than for EP students; and (3) the functioning of some items was improved after being linguistically modified for both EL and NEP students.

used so that the resulting standardized difference is more comparable to the estimates reported by Hill, Bloom, Black, & Lipsey (2008).

- *Reliability analysis.* To explore interitem relationships on each item set, a Kuder-Richardson reliability (KR-20) coefficient was estimated for each student subgroup. A higher KR-20 value indicates that items assess the underlying construct being measured more homogeneously. As expected, on both item sets (original and linguistically modified), the KR-20 was the highest (0.79 and 0.78, respectively) for the EP students and the lowest (0.61 and 0.68, respectively) for EL students. Also, for both EL and NEP students, the KR-20 associated with the linguistically modified item set was higher (0.68 versus 0.61 for EL students; 0.70 versus 0.67 for NEP students) than the KR-20 associated with the original item set. These findings were consistent with the CFA analyses in that the item-factor relationship varied across student subgroups. Together, these findings suggest that the linguistically modified items were more closely tied to the underlying factor (math understanding) than the original items for both EL and NEP students. This does not appear to be the case for EP students.
- *Correlations.* To test whether the linguistic modification altered the targeted math construct, a simple Pearson correlation between the state's standardized math test score and the item set raw number-correct score was computed for each item set and Fisher's z-transformation tests then were used to test the equality of two correlations. For each grade, the correlations between item set raw score total and state test score did not differ significantly across the two item sets, indicating that linguistic modification did not alter the construct (math understanding) being measured by the items.

In summary, findings from this study suggest that: (1) EL, NEP, and EP differences in the effect of linguistic modification across 25 items measuring math understanding varied, depending on the scoring approach; (2) for each student subgroup, the mean difference in performance on the two item sets was greatest for EL students, followed by NEP students; (3) as implemented in the current study, linguistic modification did not alter the targeted math construct assessed; and (4) for all three student subgroups, one dominant factor (math understanding) was found to underlie both item sets; however, the measurement structure between the underlying factor and the items differed across student subgroups.

These findings support future research focused on in-depth item-level analyses. Targeted research in that direction may lead to better understanding of the ways in which item and content characteristics interact with linguistic modification strategies and of possible explanations for the inconsistencies in findings across approaches that emerged in this study.

Increasing understanding of the effectiveness of accommodations is critical for policymakers in this region as decisions about the appropriateness of an accommodation for statewide testing must be based in part on empirical evidence. Because it remains unclear how effective current practices are for accommodating EL students during testing, this study sought to systematically examine the degree to which changes to test items that are research- and theory-based increased access to tested content for EL and NEP students. Though a number of questions remain unanswered, this study contributes to the body of knowledge informing appropriate accommodations guidelines for EL students so that we can develop more valid and reliable measures of what these students know and can do.

1. Study overview

When students take a state achievement test in mathematics, test directions and test items typically are presented in English. Students with low English proficiency might not understand the test directions or the math problem they are expected to solve. As a result, their test scores may be a measure of their limited English skills or other factors rather than an accurate measure of only their math knowledge and skills.

Both theory and research suggest that students, especially English language learner (EL) students, may be constrained in showing what they know and can do during standardized testing if they encounter barriers to accessing the content of the test; the test therefore measures factors other than students' content-related knowledge and skills (Abedi, Hofstetter, & Lord 2004; Butler & Stevens 2001). The complexity of the language used in a test item in terms of vocabulary, grammar, functions, or register—its language load—may interfere with EL students' ability to demonstrate their understanding of the assessed content (Rivera & Stansfield 2001).

Research has shown that math test items can be linguistically modified to reduce language load without altering the construct being assessed (Abedi 2008; Abedi, Courtney, Leon, Kao, & Azzam 2006; Abedi, Hofstetter, & Lord 2004; Abedi & Lord 2001; Sato 2008).⁵ This study was designed to examine whether one type of accommodation, linguistic modification, when applied to math test items, improves the accessibility of assessed math content and increases the validity of items measuring math understanding, particularly for EL students with limited English proficiency and non-EL students who do not reach a level of proficiency on state and federally mandated English language arts (ELA) assessments.⁶

Study context

States are trying to determine whether their assessment practices for EL student populations are consistent with the expectations of the No Child Left Behind Act of 2001. Particularly problematic is access to tested content—conditions that support meaningful engagement with the academic content and constructs on which students are being tested (Sato, Moughamian, Lagunoff, Rayyes, Rivera, & Francis, in press). Students are said to have access to tested content when they can demonstrate what they know and can do in a content area during standardized achievement testing. Access to tested content is constrained when conditions such as aspects of presentation or format of test information or response requirements, sociocultural contexts, or culture-specific references interfere with students' ability to demonstrate their content knowledge and skills. Such conditions disadvantage certain groups of students by

⁵ See chapter 3 for further description of the measurement model that underlies the authors' conception of how accommodations work to increase students' access to tested content.

⁶ Limited proficiency in English is a source of challenge that may affect access to tested content of non-English language learner students with limited English language skills and knowledge (Abedi & Lord 2001; Abedi et al. 2006). Therefore, this study also examined the effect of linguistic modification on non-ELA-proficient non-EL students.

introducing sources of construct irrelevant variance not related to the content or construct intended to be measured (Messick 1989).

When student access to test content is constrained, tests might measure student abilities, skills, and knowledge that are unrelated to the intended test constructs (sources of construct-irrelevant variance). Limited access can allow construct-irrelevant factors to interfere with a student's ability to understand and respond to a test item, so that test results underestimate the student's level of achievement in the target content area. Limited access also can affect the intended construct in that the assessment no longer sufficiently measures the targeted domain (construct underrepresentation).

Strategies to facilitate student access to tested content seek to address particular challenges faced by students during testing. In the case of EL students and math tests, access strategies would address the linguistic challenges these students face comprehending the language of the test so that they are better positioned to demonstrate their math skills and knowledge (Sato, Rabinowitz, & Gallagher, in press). EL student access to tested content is a concern for educators because limited access affects the accuracy of academic assessments for this population, compromises the validity of interpretations drawn from the test results, and raises questions about the comparability of EL students' test scores with those of their English language-proficient peers.

Two factors in particular may constrain accurate measurement of EL students' knowledge and skills and underscore the need for test accommodations. First, EL students are more likely to lack fluency skills in the language of the tests (Solano-Flores & Li 2006). Such students frequently are more fluent in conversational skills than in academic language skills (Cummins 1981; National Research Council 2002).⁷ Thus, during assessment, EL students may need to direct their cognitive resources to interpreting linguistic structures, phonological features, and other aspects of the language of the test, rather than to selecting or developing a response based on the item's targeted content. This notion of cognitive resource allocation and the effect of language load in testing is supported by evidence that the performance gap between EL and other students narrows on math items with low language load, such as math computation problems (Abedi 2001).

Second, EL students' education histories and other experiences may affect how they interpret the language in test items (Abedi 2004; Abedi & Dietel 2004; Garcia 2000; Goh 2004; Kopriva 2000; LaCelle-Peterson & Rivera 1994; Scribner 2002; Solano-Flores & Trumbull 2003; Solano-Flores & Li 2006). EL students' level of language literacy and interpretation of the

⁷ Academic language, broadly defined, includes the language students need to meaningfully engage with academic content. Academic language is not limited to academic vocabulary (such as *perimeter*, *hyperbole*, and *evaporation*) and does not necessarily require separate content-specific language lists. Rather, academic language includes the language demands—the words, grammatical structures, and language functions related to describing, sequencing, summarizing, and evaluating, for example—needed to facilitate student engagement with and achievement of grade-level academic content (Assessment and Accountability Comprehensive Center 2009; Sato & Worth in press). Research suggests that, while related to academic content knowledge, academic language skills require explicit instruction and opportunities for students to strategically apply these skills (Cummins 1980, 2005; Schleppegrell 2001, 2006; Snow, Cancini, Gonzalez, & Shriberg 1989).

context of test items are affected by the dialect they speak, the amount of formal elementary and secondary schooling they completed in the home country and in the United States, the depth and breadth of their academic knowledge base, family mobility, and the consistency and continuity of English language instruction (Albus, Bielinski, Thurlow, & Liu 2001; Liu, Anderson, Swierzbin, & Thurlow 1999; Solano-Flores & Li 2006).

In response to concerns about EL student access to tested content, state education agencies have adopted a variety of policies on test accommodations (Bielinski, Sheinker, & Ysseldyke 2003; Rabinowitz, Ananda, & Bell 2004; Rivera & Collum 2004; Thurlow & Bolt 2001). A test accommodation is deemed reasonable when standardized administration conditions would not provide students with equal opportunity to demonstrate what they know and can do (Abedi & Lord 2001; Butler & Stevens 2001; Holmes & Duron 2000; National Research Council 2004; Tindal & Ketterlin-Geller 2004). The accommodation is intended to minimize the effects on test performance of factors unrelated to the concepts and content-based knowledge and skills being assessed. For EL students, research suggests that one such factor is language (Abedi, Hofstetter, & Lord 2004).

Theoretically, use of a test accommodation should not significantly alter the construct being assessed, so test results for accommodated students are treated as comparable to those for students assessed without accommodations (Baker 2001). However, little empirical data about the actual effects of accommodations on performance are available to inform states seeking to implement fair and appropriate testing practices. This limitation contributes to inconsistency among state policies on allowable accommodations for EL students (Goh 2004; National Research Council 2004; Rivera & Collum 2004; Thurlow, Wiley, & Bielinski 2002).

For these reasons there is a need for methodical and rigorous investigation of the effects of test accommodations on EL students' access to tested content and on that of non-EL students who have not reached proficiency in ELA. This study responds to that need, building on recommendations from research in this field (Abedi 1999, 2001, 2004; Abedi & Lord 2001; Abedi, Courtney, & Leon 2003; Abedi, Lord, Hofstetter, & Baker 2000; Abedi, Courtney, Microcha, Leon, & Goldberg 2005; Rabinowitz & Sato 2005; Rivera & Collum 2004; Thurlow, McGrew, Tindal, Thompson, Ysseldyke, & Elliot 2000).

Statewide testing over the past five years has revealed large differences in achievement in math between EL and non-EL students (Kieffer et al. 2009). Because of the high stakes associated with these assessments, an empirical basis is needed to ensure that tested content is equally accessible to all students, regardless of language background. Findings from this study may advance current understanding of technically sound assessment practices by presenting empirical evidence on how linguistic modification affects access to tested content both for EL and non-EL students. Specifically, this study aims to increase understanding of the effectiveness of one test accommodation for addressing the linguistic access needs of students with limited English proficiency and decreasing the achievement gap between students who are proficient in English—particularly in the language needed to comprehend academic content and demonstrate understanding on assessment of such content—and those who are not.

Description of the accommodation (linguistic modification)

This study investigates whether linguistic modification of assessment items as typically presented on math achievement tests affects student access to math content during standardized testing. Linguistic modification is a theory- and research-based process for changing the language in test items in ways that support clarity without simplifying or significantly altering the targeted construct assessed (Abedi 2008; Abedi et al. 2005; Sato 2008). As previously described, linguistic modification is intended to increase student access to tested content by minimizing the language load associated with the text in a test item that could place certain groups of students at a disadvantage, such as aspects of presentation or format of information, aspects of response requirements, and unfamiliar sociocultural contexts or references. This can be accomplished by, for example, reducing sentence length and complexity and using common or familiar words and concrete language (Abedi 2008; Abedi, Lord, & Plummer 1997; Sato 2008; Sireci, Li, & Scarpati 2002).

To be appropriate, the accommodation should not result in a significant change to the construct being assessed. That is, linguistic modification may remove nonessential language to make an item less linguistically dense or complex, but it should not alter the math knowledge and procedures required to solve the problem. For test results with linguistically modified items to yield valid interpretations, the math content of a linguistically modified item must be comparable to that of the original item.

This study ensured the comparability of the original and linguistically modified items in several ways. First, item-specific data were collected based on expert judgment, cognitive interviews with students as they attempted to solve sample test items, and the results of pilot testing the items in original and linguistically modified formats. Then, using the test equating process described in chapter 3 based on the item response theory (IRT) models, each student's math understanding score (theta estimate) was placed on a common metric to allow for further analyses.

To test the impact of linguistic modification on student access to tested content, two sets of math items (original and linguistically modified) were administered to three subgroups of students that differed in their English language proficiency (ELP) and ELA skills: EL students, non-English-language-arts-proficient non-EL (NEP) students, and English language arts-proficient non-EL (EP) students. Items were selected from two content strands: (1) measurement and (2) number sense and operations. Item sets were assigned to students at random, with approximately half of each subgroup receiving the original set of items and half receiving the linguistically modified set.

Research questions

The research questions for this study emerged from recommendations from previous studies and from the expressed needs of the state education agencies in the West Region. One primary research question and three secondary questions guided this study.

The primary research question focused on the extent to which linguistic modification of mathematics test items improved the accessibility of math content for EL and NEP students. Specifically, the primary research question was:

- Does the effect of linguistic modification on students' math performance (as measured by raw scores or IRT theta estimates) vary across the three subgroups of students (EL, NEP, and EP students)? If so, did the linguistic modification improve the math understanding scores for the EL and NEP students relative to the EP students?

If linguistic modification increased the accessibility of EL and NEP students to the assessed math content, one would expect a significant score gain between the original and linguistically modified item set for EL and NEP students while no or minimal effect for EP students.

The secondary research questions were intended to provide information to support findings from the primary research question and to examine the degree to which linguistic modification retains the integrity of the targeted math constructs and whether the relationship between the underlying constructs and the associated items differed across the three student subgroups within and between item sets. They were as follows:

- For each item set, do any items exhibit differential item functioning (DIF) between EL students and EP students and between NEP students and EP students? How do the DIF findings differ between the original and linguistically modified item sets? In other words, when comparing both EL and NEP students with EP students with similar math achievement levels, do the probabilities of the students answering individual items correctly differ on the test with linguistically modified items as compared with the test with original items? Does linguistic modification reduce the number of items showing DIF?

Findings from these questions are of interest because an item showing DIF may be measuring something other than the construct of interest (math understanding).

- Does the number of factors that underlie student responses to an item set (original or linguistically modified) differ for EL, NEP, and EP students? Do item-factor relationships differ across the three student subgroups? If more than one factor underlies performance on each item set, do the correlations among factors differ across the three student subgroups? Does linguistic modification reduce the number of factors or affect the item-factor relationship?

Answers to these questions would help to evaluate: (1) the number of underlying factors in each item set by student subgroups; (2) for each item set and student subgroup, the relationship between the underlying factors and the associated items ("measurement structure"); (3) the correlations between the underlying factors ("factor structure") if more than one factor was identified; and (4) the extent to which linguistic modification changed the measurement structure and the factor structure for EL and NEP students relative to EP students.

- For the EP students, do raw scores on the original and the linguistically modified item sets correlate similarly with scores from the state’s standardized tests of math achievement?

This question was intended to examine the degree to which mathematics items can be linguistically modified to reduce language load without altering the construct intended to be assessed. If the correlation of item set raw scores with the standardized scores were similar for the original and linguistically modified item sets, it would support the assumption that the items had been linguistically modified without altering the targeted construct.

Findings from analyses associated with the research questions listed above are intended to inform state education agencies seeking to implement defensible policies on test accommodations for EL students in the West Region and to serve the larger research community by extending findings from previous studies.

Overview of study design

Through strategic planning and design, researchers sought to minimize the burden on district and school staff who supported data collection efforts and on students who participated in the study. To meet these needs, the work was conducted in steps. In the initial step, a group of experts developed the two sets of math test items: one set containing released items from state and national assessments (original items) and one set containing linguistically modified versions of those items.⁸ Item selection and linguistic modification procedures are described in greater depth in chapter 2.

The second step focused on validating the two item sets to ensure that they were appropriate for the target age group and student population and yielded results from which valid inferences could be drawn about students’ understanding of math content. Data were collected through cognitive interviews with EL and non-EL students and pilot testing of the items with a small sample ($n = 100$) of EL and non-EL students. How these data were used is described in chapter 2.

In the third step, operational administration of the two item sets, a large sample ($n = 4,617$) of EL and non-EL students was randomly assigned to be tested on either the original or the linguistically modified set of math items. These students were also asked to respond to five questions, available in English and Spanish, about their language background. District and school staff were asked to provide archived data about each tested student, including the student’s most recent test scores in ELA⁹ and math and EL students’ level of

⁸ Released items are those that have been used previously on scored assessments but have been removed from tests and released to the public by developers so that stakeholders can see types of items that appear on the test.

⁹ This measure was used because it provided the most recent report on each student's level of knowledge and skills in reading, writing, and written language conventions. The state's blueprint for its ELA assessments at grades 7 and 8 calls for 24 percent of the test to measure reading comprehension, 23 percent to measure writing strategies, 21 percent to measure language conventions, 17 percent and 20 percent to measure literary response at grade 7

ELP. Analyses of these data are described in chapter 3. Findings from all analyses are presented in chapter 4.

Tables 1 and 2 summarize the data collection activities, the purpose for each activity, and the associated analyses.

Table 1. Overview of data collection activities related to item development and refinement

Data collection method	Intended use of data	Analyses
<i>Expert judgment.</i> Experts in mathematics, applied linguistics, language development, measurement, curriculum and instruction, and EL students reviewed, selected, and linguistically modified items for the cognitive interviews, pilot test, and operational item set administrations (84 items).	Initial screening to reduce a pool of eligible items from approximately 350 to 50, to identify items most appropriate for linguistic modification, and to apply theory- and research-based strategies to linguistically modify items.	Item-level analyses
<i>Cognitive interviews.</i> Nine students (five EL and four non-EL students) were engaged in a think-aloud protocol around a set of math test items (63 items across nine forms).	To better understand how students access tested content and to check assumptions about the effectiveness and appropriateness of linguistic modification strategies (see <i>Item refinement based on cognitive interviews</i> in chapter 2 for the ways in which findings informed refinements).	Item-level analyses
<i>Pilot testing.</i> One of two matched sets of 30 math items (one with linguistically modified items and one with original items) was administered to more than 100 middle school students under experimental design conditions. The test booklet included eight questions about language background (30 matched pairs of items across two forms, $n = 60$).	To refine items and test assumptions about linguistic modification strategies and to improve the clarity of the language background questionnaire and test administration protocol (see <i>Item refinement based on pilot test data</i> in chapter 2 for the ways in which findings informed refinements).	Item-level analyses

and 8 respectively, and 15 percent and 12 percent to measure word analysis, fluency, and systematic vocabulary development at grade 7 and 8 respectively.

Table 2. Overview of data collection activities related to impact analyses

Data collection method	Intended use of data	Analyses
<i>Operational test administration.</i> One of two sets of 25 math items (one set with linguistically modified items and one with original items) was administered to 4,617 middle school students (to meet a target sample of 3,600 students) ^a under experimental design conditions.	To examine the effects of linguistically modified items on the performance of three student subgroups: EL, NEP, and EP students.	Item-level analysis, analysis of variance, differential item functioning analysis, factor analysis
<i>Student language background survey (English and Spanish versions).</i> Students were asked to answer five questions about their language background in their test booklets after completing the math items.	To provide additional types of information about the language background of student participants for subgroup classification and to help identify and control factors that affect test performance.	Qualitative analyses
<i>Student-level data from school records or district database.</i> Schools or districts submitted archived data for all tested students (recent state test scores in ELA and math for EL and non-EL students; ELP score for EL students).	To provide additional types of information about student participants to help identify and control factors that affect test performance.	Correlations

a. Target sample number based on power analyses; see appendix A for details.

Structure of the report

The report is structured as follows. Chapter 2 describes the study design, sample selection and recruitment, item set development processes, and standardized administration procedures. Chapter 3 describes the implementation of the accommodation (linguistic modification), including discussion of considerations and methods for data analysis. Chapter 4 presents findings from data analyses. Chapter 5 summarizes and interprets key findings, describes study challenges, comments on implications of the findings, and offers recommendations for future research.

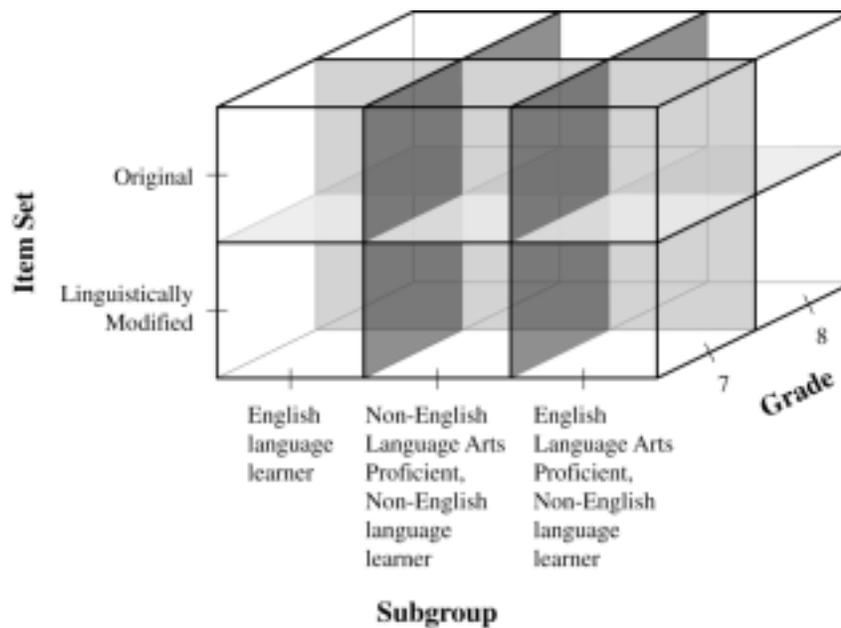
2. Study design, study sample, and item set development

This chapter describes the study design, sample selection and recruitment, item set development processes, and standardized administration procedures.

Study design

This study followed a 2 x 3 x 2 fully crossed factorial design. As shown in figure 1, the factors were item set (original or linguistically modified) student subgroup (English language learner [EL] students, non-English-language-arts-proficient non-EL [NEP] students, and English language arts–proficient non-EL [EP] students), and student grade level (grades 7 and 8).¹⁰

Figure 1. Study design



Item sets were randomly assigned to students in grades 7 or 8, regardless of apparent language status (EL or non-EL student) or other subgroup membership (ELA proficient or non-ELA proficient). ELP status and ELA proficiency status were determined after item set administration, during the data analysis phase, based on the archived student-level data collected from districts. Item set booklets (original and linguistically) were distributed randomly within each classroom at the time of testing to students seated

¹⁰ Although formal subgroup assignment was not confirmed until after testing, using data about ELP status for tested students collected from districts, this design allows for the study's intent to be evident from the outset. From a traditional experimental design framework, this design is more appropriately described as a 2 x 2 design in which two item sets were randomly distributed to students within two grade levels.

in their normal classroom seat assignments. Completion frequencies were examined during analyses to confirm equal distribution of the two item sets across all subgroups.¹¹

Overview of study steps

Prior to data collection, two sets of math items were developed and refined, one with original items and one with linguistically modified items, and students were recruited for the study. Preliminary exploration of accommodation (linguistic modification) effectiveness was conducted through a small sample of students in March 2008 and through pilot testing in April 2008. The final item sets used during operational testing were developed between January and December 2007 (see table 3 on the study timeline). Recruitment for the operational administration extended from February through May 2008. Eligible non-EL students included all general and special education students, with the exception of special education students whose Individualized Education Plan (IEP) called for an accommodation during testing other than extended time or small group administration. The target EL sample was students whose first language was Spanish and who demonstrated early intermediate to advanced levels of ELP on the California English Language Development Test (CELDT). The two final sets of test items were administered between March and July 2008.

Randomization of item sets was at the student level. Test booklets with original items (Form O) and those with linguistically modified items (Form M) were placed in alternating order (M-O-M-O . . .) prior to test administration. Regardless of subgroup membership, students in each school and grade were randomly assigned to receive either a test booklet with original items or a test booklet with linguistically modified items. Onsite coordinators were trained to monitor this process and to provide written documentation of any deviations.

Archived student-level achievement data were collected from schools and districts for each tested student. Data included state standardized test scores in ELA and math and, for EL students only, an ELP score. These data were matched to performance data on the item sets using student identifiers. Final determination of subgroup membership was made at this time. The achievement data also were used to verify (a posteriori) appropriate randomized distribution of the original and linguistically modified item sets across the three subgroups. Findings from this a posteriori analysis did not suggest a need for modification to the sample.

Student responses to the math items were entered and verified between May and September 2008. The district-archived data were received from June through November 2008. Data analyses were then conducted starting December 2009.

¹¹ As shown later in figure 3, the number of English language learner students completing each item set was nearly identical (819 for the original item set and 818 for the linguistically modified item set). The number of non-English language learner students completing each item set also was comparable (1,874 for the original set and 1,869 for the linguistically modified set).

Table 3. Timeline for study activities, January 2007–January 2009

Date	Develop instrument				Validate instrument (implementation phase)				
	Item selection	Item modification (linguistic modification)	Cognitive interviews	Item pilot test	School recruitment	Student sample selection	Student testing	Archived student data retrieval	Data analysis
<i>2007</i>									
January									
February									
March	█								
April	█								
May	█								
June	█	█							
July		█							
August		█							
<i>2008</i>									
February					█				
March		█	█		█				
April		█		█	█	█			
May					█	█	█		
June						█	█		
July							█	█	
August								█	
September									█
October									█
November									█
December									█
<i>2009</i>									
January									█

Sample recruitment

To control for cross-state differences in the math content standards on which state assessments are based, data collection was restricted to one state in the West Region. California was selected because of its large Spanish-speaking EL student population and because the state uses a consistent measure of ELP (the CELDT). Spanish was selected as the language for the study because 75 percent of EL students in the West Region states (Arizona, California, Nevada, and Utah) identify Spanish as their primary or secondary language. Studying only native speakers of Spanish removed sources of variability related to native language.

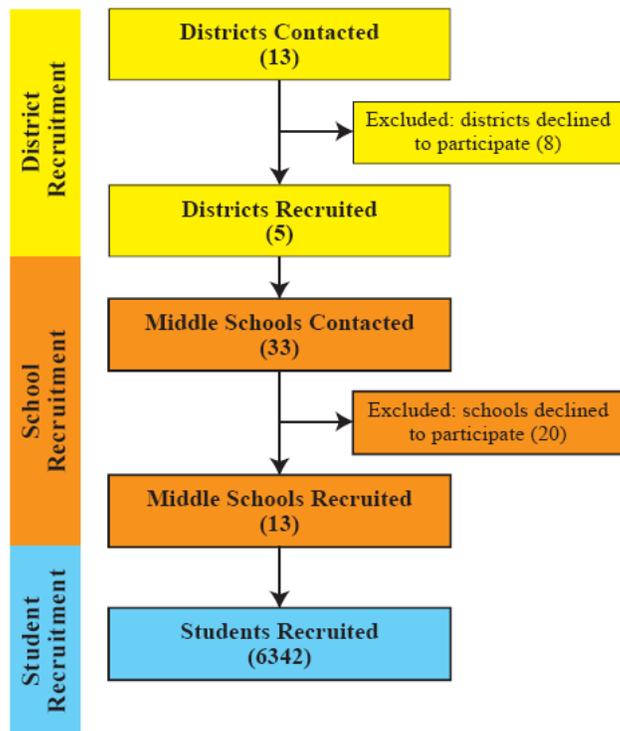
Demographic data from the California Department of Education were used to identify 13 school districts with high percentages (25 percent or greater) of middle school EL students whose native language was Spanish. Superintendents in these 13 districts were contacted to explore their willingness to participate in the study. Several of these districts had participated in previous Regional Educational Laboratory (REL) West research studies and so were familiar with the expectations. Others had expressed an interest in the assessment needs of EL students

and were thus agreeable to considering participation in the study. However, due to previous commitments during the study’s proposed testing window (January–June 2008) and concerns about overburdening busy schools, only 5 of the 13 superintendents agreed to participate. Researchers followed up with the superintendents in these five districts to provide additional information about the study.

In those five districts, 33 middle schools enrolled students in grades 7 and 8. Of those 33 schools, 20 schools declined to participate, despite superintendent approval. For the 13 middle schools that agreed to participate, memoranda of understanding were signed and the participating principals agreed to identify an onsite school coordinator for the duration of the study.

In keeping with guidelines in the Consolidated Standards of Reporting Trials (CONSORT) statement (Moher, Schulz, Altman, & CONSORT Group 2001), figure 2 presents a flowchart showing the number of district, school, and student participants during recruitment for the operational administration. Figure 3 in the next section presents a detailed flow chart focused on the changes in the numbers of student participants.

Figure 2. Consolidated Standards of Reporting Trials flow diagram of school recruitment



In each of the 13 middle schools, an onsite coordinator (generally a teacher, administrator, or support staff; in one district, this was a district-level staff person) was asked to inform grade 7 and 8 math teachers about the study and encourage them to participate. Participation was voluntary. For those schools that agreed to participate, the onsite coordinator in each school provided class enrollment data for participating classrooms in that school and forwarded that

information to the REL West recruiting team. Across all 13 schools, 6,342 students were identified as eligible for testing by participating schools.

Using information about individual classroom enrollment numbers sent by the onsite coordinators, testing materials were shipped to the 13 schools. These materials were packaged so that individual teachers would receive separate sets of materials for each class period. Each individual package included one test booklet for the participating teacher, a sufficient number of test booklets for each student (based on the estimate provided by the onsite coordinator) plus two extras, and parent information and permission letters (in English and Spanish) to be sent to parents or guardians of all students in participating classrooms. Following testing, teachers returned all test materials to the onsite school or district coordinator, who had been asked to return all test booklets (completed and unused) to REL West.

Descriptive information about the 13 schools that participated in the study, including the number of eligible students in each school and number of test booklets actually completed and returned to REL West from each school, is presented in table 4.

Table 4. Description of study sample, by school

School	Total enrollment (all grades) ^a	Percent EL students (all grades) ^b	Percent eligible for free/reduced-price lunch (all grades) ^b	Locale	Number of test booklets distributed (grades 7 and 8)	Number of completed test booklets returned (grades 7 and 8)	Number of students with matched data (grades 7 and 8)	Number of EL students with matched data (grades 7 and 8)
1	1,060	10	70	Urban	111	99 (89%)	94	49
2	890	10	30	Small city	75	68 (91%)	68	7
3	260	35	45	Rural	175	157 (90%)	153	51
4	1,000	60	95	Rural	1,019	835 (82%)	822	425
5	1,240	45	90	Rural	1,472	1,090 (74%)	937	461
6	1,100	20	60	Urban	762	715 (94%)	689	118
7	1,050	15	50	Urban	329	309 (94%)	300	62
8	1,310	10	45	Urban	763	664 (87%)	650	66
9	730	15	70	Urban	53	47 (89%)	46	13
10	360	50	90	Urban	28	26 (93%)	21	17
11	620	45	85	Urban	268	233 (87%)	225	125
12	1,320	45	100	Urban	817	728 (89%)	705	233
13	770	10	80	Urban	470	409 (87%)	387	29
<i>Total</i>	<i>11,710</i>				<i>6,342</i>	<i>5,380 (85%)</i>	<i>5,097</i>	<i>1,656</i>

a. Rounded to the nearest 10.

b. Rounded to the nearest five.

Source: school demographic and locale data, California Department of Education (2008a, 2008b, 2008c) and Sable (2009); all other data, authors' primary analyses.

As shown in table 4 (columns 2–5), participating schools varied in size (total enrollment), demographic composition, and locale. Overall, across all 13 schools, 85% of the 6,342 test booklets distributed were returned completed (column 6 and 7). In 11 schools, the return rate

was greater than 85%, with five of those schools showing a return rate of 90% or greater. Noncompleted (blank) test booklets in all 13 schools can be attributed to one of three sources: (1) onsite coordinators overestimated the number of eligible students so a surplus of booklets resulted; (2) students were absent and onsite coordinators did not report absentees; or (3) across all schools, 16 teachers declined to participate despite initial agreement (classroom packages returned unopened), resulting in the return of approximately 480 unused test booklets (7% of all eligible students).

This suggests that the overall return rate of 85% represents the lower-bound estimate because it is likely that more test booklets were shipped than could be accommodated by school enrollment. With at least an 85% response rate, we believe that the study results can be generalized to students in the participating classrooms.

To minimize the burden on schools and to promote the benefits of study participation as an instructional activity, principals were assured that all students in participating classrooms would be tested (except special education students whose IEP called for a test accommodation other than extended time or small group administration). No students were reported by teachers or onsite coordinators as having been excluded. However, as described in more detail in chapter 3, study analyses included only those students who took the test and who could be matched with district databases of state-level achievement data. Moreover, study analyses included only those EL students who were Spanish-speaking students with “early intermediate” or higher proficiency levels (levels 2–5) on the CELDT test. Tradeoffs associated with these decisions were carefully weighed, as described in greater detail below (see section on considerations related to sample attrition and subset exclusion) and in chapter 5.

Participant flow

The final analytic sample for the study consisted of 4,617 grade 7 and 8 students from the 13 participating schools (figure 3). A multitiered process was used to screen students for eligibility for study analyses and to categorize students into three subgroups: EL, NEP, and EP students. The criteria used to determine sample eligibility and subgroup membership are described below.

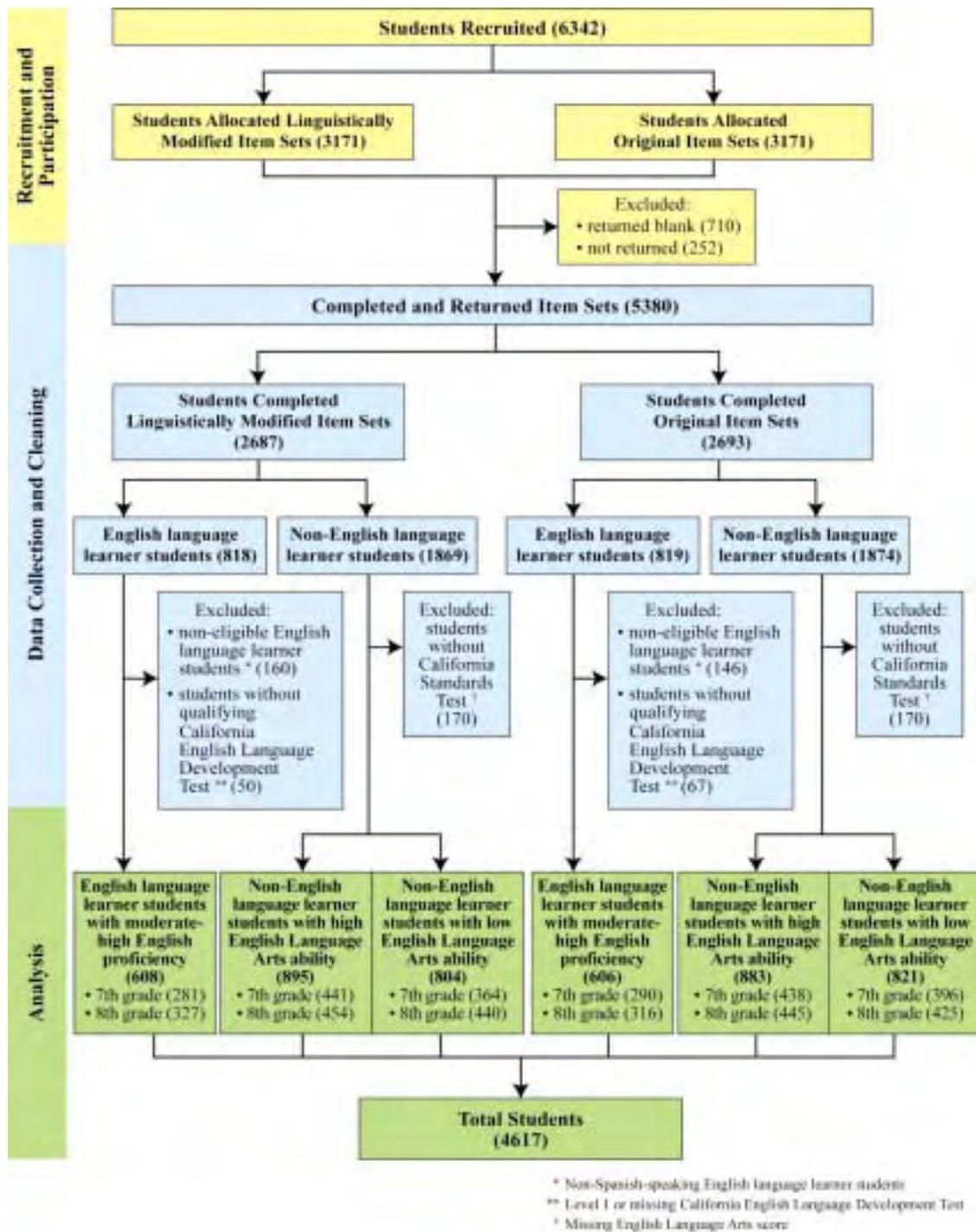
- As shown in figure 3, 5,380 of 6,342 eligible students participated in testing – 962 booklets (15%) were not used.
- Of the 5,380 students assessed in the 13 schools, only students for whom matching state standardized test score data were available were included in the analyses. State standardized test score data were requested from districts for all tested students. Matching involved linking unique student identifiers to available scores on the state’s standardized assessments in ELA and math and, for EL students only, to a score on the CELDT.

- The CELDT score was used to verify EL status; any student whose unique student identifier was matched to a CELDT score and proficiency level was included in the EL student subgroup. All other students were classified as non-EL students.
- Students were asked to answer five questions about their language background after completing the items in their item booklet. This self-reported information was used to identify Spanish-speaking EL students. Of the students identified as EL students based on a matching CELDT score, only those who listed Spanish as their first or home language were included in the analytic sample.
- The proficiency level reported on the CELDT also was used for secondary screening. Only EL students who demonstrated at least an “early intermediate” level of ELP on the CELDT (a proficiency level of 2–5) were included in the analytic sample. (Students at the “beginning” level of ELP typically cannot yet read enough English to be likely to benefit from linguistic modification as a test accommodation and were therefore excluded from the analytic sample.)
- Non-EL students (those with no matching CELDT score) whose unique identifiers were matched with state achievement test scores in ELA were separated into two subgroups: those who were proficient in ELA (scored at or above the state-established proficiency cut score) and those who were not proficient (scored below the cut score).¹²

Overall, of the 5,380 students who were administered either the original or the linguistically modified item set, 86 percent (4,617 students) were included in the final analytic sample (see figure 3).

¹² This measure was used because it provided the most recent report on each student’s level of knowledge and skills in reading, writing, and written language conventions.

Figure 3. Consolidated Standards of Reporting Trials flow diagram of student participants



Considerations related to student sample

Steps were taken throughout the study to monitor threats to validity through loss or exclusion of potential student participants. During the design phase the research team carefully weighed

tradeoffs while making decisions about the final sample of students included in the analyses. During data collection the team was vigilant in tracking every item booklet shipped to schools so the status of each booklet could be accounted for at any time. Unique challenges arose at a number of key points in this process.

- *Recruitment and participation.* Before the testing date each school provided the estimated number of student participants. These estimates guided the shipment of item booklets to schools; extra materials were routinely shipped to ensure an adequate supply of each item set. As a result, some schools returned blank item booklets (as directed), accounting for 710 of the 962 unused item booklets (see figure 3). Schools with the remaining 252 unused booklets did not return them. Follow-up conversations with school staff confirmed that the extra booklets had not been used and had been discarded.
- *Data collection and cleaning.* The focus during this phase was on verifying student subgroup status by matching completed item booklets with archived test records.
 - Across the two item sets 306 EL students who identified their primary language as other than Spanish on the Language Background Survey were excluded. Because research suggests that students with different language backgrounds may experience different language-based challenges during testing (Abedi 2004; Garcia 2000; Goh 2004; Kopriva 2000; Scribner 2002; Solano-Flores & Li 2006), researchers carefully weighed tradeoffs and elected to exclude from analysis a subset of students who would introduce different language background factors that could confound interpretation of findings.
 - Across the two item sets 117 EL students who identified themselves as EL students on the Language Background Survey were excluded from analysis because they could not be linked to a CELDT score or because their CELDT score placed them at the lowest level of English proficiency. While it is possible that EL students in this subset varied systematically from other EL students (for example, newly arrived or highly transient), researchers could not risk including students for whom neither English proficiency nor ELA proficiency could be verified. Researchers also elected to exclude EL students whose CELDT scores classified them at the lowest level of ELP because they typically are less likely to benefit from linguistic modification as a test accommodation in math (Abedi 2004).
 - Across the two item sets 340 non-EL students were excluded from analysis because they could not be linked to an ELA score. While it is possible that the non-EL students in this subset varied systematically from other non-EL students, researchers elected not to risk including non-EL students for whom a baseline ELA proficiency level could not be verified.

Item set development and administration

To develop the items sets used in this study, a work group was convened that included the core study team (senior researchers) and experts in mathematics, linguistics, measurement, curriculum and instruction, and the EL student population. Each invited expert brought particular strengths to the work group, in terms of both academic training and experience.¹³ Work group members received training materials developed by the principal investigators describing the target student population and introducing the theoretical and research-supported guidelines for linguistic modification.¹⁴ The full work group was convened for two days for training and to conduct the screening and development processes (steps 2–5 in table 5) described below. The full group then was divided into subgroup teams, each assigned specific responsibilities (for example, identifying items for cognitive interviews), that continued to meet throughout the course of the study, completing steps 6–8 in table 5. Subgroup teams communicated on a regular basis with one another and with the study’s principal investigators. All steps in this process used procedures described in this study and were guided by recommendations from the study’s Technical Advisory Committee.

Overview and summary of steps

Table 5 summarizes the eight steps in the item selection, development, and administration process. The next section presents a detailed description of each step.

¹³ Two members of the work group had advanced degrees in mathematics, one in applied linguistics, one in English as a second language, one in language development, one in curriculum and instruction, and two in educational measurement/psychometrics. Two had experience teaching mathematics at the middle school level, three worked full-time with EL students (two specifically with Spanish-speaking EL students in California), and four had worked on state test development projects for EL students. Half were senior-level staff with experience conducting research studies.

¹⁴ Appendixes D (guidelines for linguistic modification) and E (training materials) provide in-depth information about the linguistic modification process.

Table 5. Overview of item screening process

Step	Item pool	Screening, selection, or linguistic modification criteria	Outcome
1	All released items for grade 8 National Assessment of Educational Progress and grade 7 California state test	Consistent with type on state test (multiple-choice, with four response choices); of high technical quality; may be aligned to state content standards	256 original items
Which groups of items meet specific screening criteria for this step?			
2	Items with sufficient language to linguistically modify (number sense/operations and measurement content strands)	Sufficient language to linguistically modify (items use language as well as symbols to assess math construct and content); may be story problems; items assess important content for this age group (the two strands that met these criteria were number sense/operations and measurement)	115 original items
Which of these items meet specific screening criteria for this step?			
3	Diverse pool of items aligned to state standards	Aligned to state standards at appropriate grade level, measure different content, represent a range of complexity levels and item types	81 original items
What linguistic modification strategies can be applied to each item?			
4	Items to which specific linguistic modification strategies could be applied	Words typically unfamiliar to or infrequently used by EL students could be changed or removed; complex grammatical structures or sentences could be simplified; past or future tense verb forms could be changed to the present tense and passive verb forms to active forms; item format could be modified; graphic or text could be added for clarity, and the like.	51 matched pairs (51 original items and 51 linguistically modified versions of those items)
Which of these linguistically modified items were linguistically modified without changing the construct intended to be assessed?			
5	Linguistically modified items endorsed by content specialists	Review by content specialists verifies construct intended to be assessed has not been changed	42 matched pairs of items (42 original and 42 linguistically modified items)

Step	Item pool	Screening, selection, or linguistic modification criteria	Outcome
Which of these items should be examined further through cognitive interviews?			
6	Set of original and linguistically modified items (7 different items on each of the 9 forms included 1 original item, 2 linguistically modified items, and 2 matched pairs of items)	Range of item types, complexity levels, and content assessed, making sure that each linguistic modification type was used on at least one form	Feedback on 63 original and linguistically modified items
Which of these matched pairs should comprise the original and linguistically modified item sets used during pilot testing?			
7	Matched pairs of original and linguistically modified items	Comments from students during cognitive interviews suggested these were effective and appropriate items for further testing	Feedback on 30 matched pairs of items (30 original and 30 linguistically modified items)
Which of these matched pairs should comprise the original and linguistically modified item sets used during operational testing?			
8	Matched pairs of original and linguistically modified items	Data from pilot testing suggested these were effective and appropriate items for final item sets	25 matched pairs of items (25 original and 25 linguistically modified items)

Step 1. Work group members collected all available released multiple-choice achievement test items from public web sites of the National Assessment of Educational Progress (NAEP) and the California Standardized Testing and Reporting system (the California Standards Test). This pool included 191 released grade 8 NAEP items and 65 released grade 7 California Standards Test items (total of 256 items). As NAEP does not test students in grade 7, all state test items reviewed were grade 7 items to ensure equal representation of items for both grades 7 and 8. These test items had undergone extensive review by measurement and content specialists at the national and state levels and therefore were considered psychometrically sound, aligned to content standards in mathematics, and developmentally appropriate for grade 7 and 8 students.¹⁵ Items from these two sources

¹⁵ NAEP and state test items have undergone sensitivity and bias reviews. A diverse panel of content, population, and assessment experts, teachers, parents, and other stakeholders review items for possible sources of bias. For example, a test developer seeking to provide real-world relevance to an item may have unintentionally included a graphic or text reference that actually requires specific background knowledge or experience that may not be universally understood by all students (such as a cellular phone or video game).

represent the broad content strands identified by the National Council of Teachers of Mathematics that extend across all states' standards (Kenney 2000).¹⁶

Step 2. The work group conducted a preliminary review of the items in this pool, searching for groups of items with content or format that might be made more accessible to an EL student. Based on the following two criteria, they organized the items into two sets, those amenable to research- and theory-based linguistic modification and those that were not amenable to linguistic modification:

- *Sufficient language to linguistically modify.* The items had sufficient construct-irrelevant language to linguistically modify. Strictly computational items or those that included only numbers or graphs are examples of the types of items that were not amenable to linguistic modification.
- *Item content.* The items sufficiently assessed a grade-level appropriate fundamental skill or central idea or concept, as explicitly stated in the state's content standards.

At the end of this process, the work group determined that those items most amenable to linguistic modification came primarily from two content strands: measurement, and number sense and operations. Items from these two strands often were “story problems” that included text that could be reviewed more closely for possible linguistic modification. In 2007 items from those two strands comprised more than half (54 percent) of the state's test in mathematics at grades 7 and 8, so the work group decided it would be appropriate to focus its efforts on linguistically modifying the 115 multiple-choice items measuring content in those two strands.

Step 3. The work group then considered item-specific characteristics, ensuring diversity in content assessed, item type, and item complexity. If an item's alignment to a state standard could not be verified by a content specialist, it was removed from the pool. This process resulted in a pool of 81 items.

Step 4. The work group then began applying theory- and research-supported linguistic modification strategies intended to minimize construct-irrelevant variance associated with language complexity. The specific linguistic modification strategies applied depended on the item. Strategies included changing or removing words typically unfamiliar to or infrequently used by EL students (such as unfamiliar sociocultural references and idioms), unless the word was determined to be a construct-relevant technical or content-specific word. Other strategies included simplifying complex grammatical structures, changing past or future tense verb forms to the present tense and passive verb forms to active forms, simplifying phrase or sentence structure, and changing item format. As the group worked, it documented all changes to original items and rationales for applying specific linguistic modification strategies. During this process, work group members were careful to maintain the original item's content-related characteristics (target construct assessed, item type, item complexity), as verified by expert judgment. They linguistically modified items in intentional, defensible, and focused ways. All items were required to have four response choices—one correct response and three distractors.

¹⁶ These include number sense and operations, measurement, algebra, geometry, and data analysis and probability.

If an item was found not to be amenable to linguistic modification despite the work group's best efforts, it was removed from the pool. Approximately 30 items were removed from the eligible item pool at that time, resulting in an item pool of 51 matched pairs of items (51 original items and 51 linguistically modified versions of those items).

Step 5. Mathematics content experts and test development specialists then reviewed all linguistically modified items to ensure that changes did not alter the tested construct or violate accepted practice for math assessment. Another nine items were removed from the pool when the work group could not linguistically modify the items in ways that, as judged by the content experts, did not alter the construct intended to be assessed. The final outcome from this process was 42 pairs of matched items, each pair including one original item and one linguistically modified item.¹⁷

Step 6. From this set of 84 items (42 matched pairs of original and linguistically modified items), the work group selected a diverse pool of 63 items for cognitive interviews; each of the nine students then viewed seven different items: one original item, two linguistically modified items, and two matched item pairs. Detailed information about the cognitive interview process is provided below, in the next section.

Step 7. Based on feedback from interviewed students and research staff who conducted the interviews, of the 63 items used during cognitive interviews, 30 matched pairs were selected for pilot testing. Detailed information about the pilot testing process is provided in the section *Item refinement based on pilot test data*.

Step 8. Based on data from pilot testing, 25 of the 30 matched pairs used during pilot testing were selected for the final item sets to be used during operational testing.

Item refinement based on cognitive interviews

To explore initial assumptions about the effectiveness and appropriateness of the linguistic modification strategies, researchers conducted cognitive interviews with nine students in grades 7 and 8.¹⁸ Five of these students were EL students and four were non-EL students. The

¹⁷ Across states, it is standard practice to develop a large pool of items, review them for appropriateness, pilot test them, use them on operational tests for one or more years, and then replace them with newer items to avoid risk of exposure effect (that is, a student remembers the item from the previous year). During the development process, elements of good test development practice, including application of universal design principles, are applied to the items. During the review process, all items are examined for possible sources of bias. At any point in this process, an item may be removed if developers do not think it meets technical adequacy requirements for items on high stakes tests or if they think it may introduce bias. Released items are items that were used for at least one full cycle of operational testing and met all technical adequacy and bias/sensitivity requirements, but have reached "retirement" age. For this reason, the work group was challenged in finding items in this pool that were amenable to linguistic modification, that is, items with extraneous language or unnecessary text complexity; only 42 out of the pool of 115 from two strands (37 percent) were found to have the characteristics desired for this study. Implications of the decision to use psychometrically sound NAEP and state test items is discussed in greater detail in chapter 5.

¹⁸ To minimize burden to students and schools and because cognitive interviews are resource intensive, only nine students were interviewed at this time. In accommodation research or research of special student subgroups, a

student sample was drawn from a mid-size middle school in north-central California whose principal had expressed interest in participating in a research study about test accommodations for EL students. These data were intended to provide initial feedback on linguistic modification strategies and descriptive information about the ways in which students access math items during standardized testing.

Cognitive interviewing strategies are drawn from the family of verbal protocol models that can be used to confirm hypotheses about linguistic access (Solano-Flores & Trumbull 2003). During the cognitive interview a trained researcher guides students individually through a think-aloud protocol as they work on a set of items (concurrent data collection). Once each student has responded to all test questions, the researcher asks the student a set of follow-up questions to clarify or verify comments made earlier or to probe more deeply the student's thinking processes about individual items (retrospective data collection). This scripted, multistep process is intended to reveal the types of knowledge and skills that might support students' ability to correctly respond to the item (Kopriva 2000).

For the cognitive interviews in this study, the principal from the participating school arranged for the interviews to be conducted over the course of three days, during school hours, in a quiet setting in the school. School staff identified five EL and four non-EL students who met the eligibility criteria¹⁹ and whose parents or guardians might consider allowing them to participate. Parents or guardians of identified students received information letters about the interviews and were asked to return a signed permission form. An active consent protocol was used because participating students would be engaged in conversation with a researcher rather than a school staff member. Recruitment was ongoing until the target sample of nine students was reached.

Based on verbal protocol research and experience, researchers determined that each student could respond to cognitive interview questions on approximately seven test items before becoming fatigued. Researchers selected 63 items from the total of 84 (42 matched pairs)—some original and some linguistically modified—that represented the range of linguistic modification strategies implemented by the workgroup. Each of the nine students was assessed using a test booklet with a distinct combination of seven original and linguistically modified items. The process that researchers followed to arrange the items into nine distinct booklets and the think-aloud interview protocol are provided in appendix F.

Researchers administered the think-aloud protocol to each student individually. Students were given the opportunity to practice the think-aloud process before administration of the actual test items. Once the student was trained and the researcher was confident that the student understood the task, the researcher guided the student through the think-aloud process for the

sample size of 5–10 students is considered adequate for cognitive interviews to answer preliminary research questions about the effectiveness of an intervention or accommodation (Almond, Cameto, Johnstone, Laitusis, Lazarus, Nagle, Parker, Roach, & Sato 2009; Van Summeren, Barnard, & Sandberg 1994).

¹⁹ All non-EL students were eligible except those whose IEPs called for an accommodation other than extended time or small group administration. The target English language learner sample was students whose first language was Spanish and who demonstrated early intermediate to advanced levels of English language proficiency on the CELDT.

multiple-choice test items in the test booklet. On average, each student's total interview time was 30 minutes. Each interview was audio recorded, and the interviewers wrote comments on a data collection rubric adapted for this context from Paulsen & Levine (1999) and van Someren, Barnard, & Sandberg (1994).

Researchers and workgroup members reviewed the descriptive data from the cognitive interviews to better understand the strategies that students used to access test content and potential barriers to access. Workgroup members then considered these findings in selecting the set of 30 matched pairs of original and linguistically modified items (from the original 42 pairs) that would be used during pilot testing. The 12 matched pairs that were removed before pilot testing were near duplicates of other items in either content standard tested or linguistic modification strategy applied. The cognitive interviews, which provided information about students' understanding of the items, helped inform the removal of some of the near duplicate items. The goal during this step was to ensure that the test item set represented a range of knowledge and skills, levels of cognitive complexity, and linguistic modification strategies. In addition, based on student responses in the cognitive interviews, the language in two of the eight Student Language Background Survey questions was clarified. The outcome of this process was a set of 30 matched pairs of multiple-choice items and eight language background questions.

Item refinement based on pilot test data

To further examine the effectiveness and appropriateness of the linguistic modification strategies, researchers conducted a pilot test of the items with a convenience sample of grade 7 and 8 EL and non-EL students. The pilot item sets were administered to students in five grade 7 and 8 classrooms in a large middle school in California with a large population of Spanish-speaking students. This school had participated in a previous REL West study, and the principal had expressed interest in research about test accommodations for EL students. The onsite facilitator arranged for the pilot tests to be administered during school hours in intact classrooms by REL West researchers over the course of two days. The final sample of 112 participating students included 64 EL and 48 non-EL students.

Researchers used the pairs of items selected by the work group to develop two matched sets of items: one with the 30 original items and one with the 30 linguistically modified items. Initially, the items were arranged to ensure variability in the content spread across items—items were ordered so that number sense and operations items alternated with measurement items. Researchers then balanced the following considerations in arranging items in the test booklet, using an iterative process to determine optimal sequence: the content standard assessed, use of a graphic or table in the item, use of a proper name in the item, item difficulty, and the item's correct response letter (A, B, C, or D). Because each item presented a different combination of considerations (for example, one item assessed multiplication of a fraction and included a graphic while another item assessed computation skills using percentages and a proper name), item placement was carefully considered before the arrangement was finalized.

The test booklets were placed in alternating order: original item booklet (Form O), then linguistically modified item booklet (Form M), then original, then linguistically modified, and so on. Tests were administered by trained researchers, who gave one test booklet to each student, with students sitting at their desks in their normal seat assignments, alternating between Form O and Form M. Researchers were asked to document any deviation from this test distribution method.

Item- and subgroup-level statistics were generated from the pilot test administration. For each item, an item p -value (proportion of students answering the item correctly),²⁰ point biserial correlation (item to total correlation), and omission rate were examined for each subgroup. Mean and standard deviation of total score also were examined for each subgroup. Together with observations from test administrators, data from pilot testing were used to inform decisionmaking during final selection for the item sets. A research team with expertise in assessment, applied linguistics, math content, and the EL population met to discuss the findings of the pilot test, which included observations from test administrators, and to make recommendations about possible item refinements or deletions. The team considered item format, item content, and performance data (such as item p -values and point biserial correlations) during this discussion.

Two recommendations emerged from this discussion. One was based on rates of omission; the other concerned the Student Language Background Survey. Nearly 40 percent of the student sample did not attempt the last four items. The research team discussed possible reasons with the pilot test administrators and concluded that most students could not answer 30 math items within the time allowed (one class period, or 50 minutes). Team members concurred that five items should be removed to ensure that students had adequate time to answer both the math questions and the Student Language Background Survey questions and that this would not have a significant effect on the reliability of the item sets.

In eliminating the five items, researchers considered responses to the following questions:

- Which items were not the strongest measures of the intended content?
- Which items were not the strongest examples of linguistic modification?
- Which items were similar to others in content assessed or linguistic modification strategy used?
- Which items, across students in this small convenience sample, appeared most and least challenging?

²⁰ The term “ p -value” has different meanings in different fields. More typically, “ p -value” is used in statistical hypothesis testing to indicate the probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true. As used here, however, an item p -value corresponds to the proportion of students who answered a test item correctly, which has been used as an indicator of the difficulty level of a test item in classical test theory. This use of the term “item p -value” is standard in educational measurement.

After discussing responses to these questions, five items were selected for elimination. In each case the rationale for elimination was documented.

The second recommendation, based on student responses and test administrator observations, was to remove three questions from the Student Language Background Survey. These questions were frequently misunderstood or misinterpreted by EL students and were not integral to addressing the study's research questions. Researchers also reasoned that middle school students were more likely to complete a shorter survey.

Each final operational booklet thus contained 25 multiple-choice items (original or linguistically modified; see appendixes N and O) and five language background questions (see appendix C). Technical advisors agreed that a test of this length was appropriate for students in grades 7 and 8 to complete in one class period without undue burden and was adequate for testing the study's hypotheses. Tests of similar length have been used in recent studies examining the effects of accommodations on student performance (Abedi 2001; Abedi et al. 2005; Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch 2000; Castellon-Wellington 2000; Rivera & Stansfield 2001; Hofstetter 2003).

Item booklets

Each test booklet included 25 multiple-choice math items. Test directions were provided in English and Spanish. After completing the math items, students were asked to answer five questions (available in English and Spanish) about their language background. The Student Language Background Survey is provided in appendix C.

Developers of assessments for EL students face particular challenges when items or supporting information (for example, test directions) are translated from English into a second language; they must take steps to ensure that the "adaptations" make the information more accessible to the EL students in their target population (Huempfer 2004; Rabinowitz, Ananda, & Bell 2004; Stansfield 2003). For development of the translated test directions in this study, researchers relied on a native Spanish speaker who is bilingual, works regularly with students in California schools, and has experience in translating documents from English to Spanish in those schools.²¹

During pilot testing, two Spanish-speaking instructional support staff from the participating school were asked to review the translations and provide comments if they had questions or concerns about the appropriateness of the translated directions for their middle school Spanish-speaking EL students. Based on feedback and recommendations from these staff members—

²¹ Despite these efforts, the translated directions may not have been equally accessible to all Spanish-speaking EL students in this study. Because Spanish-speaking students do not comprise a homogeneous population, the translated directions may have been more helpful to those EL students who speak a particular dialect or have a language background comparable to that of the translator. Therefore, additional review and input on the translations were sought during pilot testing. Accommodations and adaptations for this population of students were applied with particular sensitivity to factors such as differences in cultural and linguistic background and experiences.

who work full-time with EL students in this school at this grade level—language in the test directions and in the Student Language Background Survey was refined slightly.

Student achievement history data

Each district was asked to provide archived achievement history data for all tested students (state test scores in ELA and math for all; ELP score and proficiency level for EL students). Four of the five participating districts submitted achievement history data from school records or the district database and researchers matched the archived data for each student to the unique student identifiers on the test booklets. Three districts posted these data on a secure file transfer server and one district submitted data on a CD. The fifth district conducted the matching at the district office and then transmitted data for all tested students on a secure file transfer server.

3. Implementation of the accommodation (linguistic modification) and methods for analysis

This chapter describes the implementation of the accommodation (linguistic modification), including discussion of considerations and methods for data analysis.

Operational administration of the item sets

Grade 7 and 8 students, regardless of subgroup membership, were randomly administered either the control group item set (the one with the original items) or the accommodated item set (the one with linguistically modified items) or. Onsite coordinators were trained by Regional Educational Laboratory (REL) West researchers to strictly follow the prescribed item booklet distribution protocol. The coordinators coached test administrators, who in most cases were classroom teachers, to walk directly from the front of the room to the back, row by row, handing out test booklets in the alternating order (linguistically modified, original) in which they had been placed, to students seated at their desks in their normal seating assignments. Coordinators were asked to monitor this process across all classrooms and to provide written documentation of any deviations. No deviations were reported.

Test administration protocol

The names and contact information for the onsite coordinator were obtained from school principals. Each coordinator submitted class size information, which was used to sort and prepare materials for shipment to participating schools. Coordinators also contacted teachers of sampled classrooms to inform them about study responsibilities. Classroom teachers were identified as the most appropriate test administrators because this mimics current state practice in relation to the administration of standardized achievement tests. Teachers were asked to carefully read the test administration manual (see appendix B) and to contact their onsite coordinator or designated project researcher with any questions. Because the test administration protocol and script were similar to those used for state-administered assessments, teachers were familiar with the procedures and had few questions. In collaboration with the onsite coordinator, REL West researchers were available to respond to questions from teachers at any time before, during, or after test administration. Both the onsite coordinator and test administrators had explicit instructions for noting any irregularities that arose during testing and for returning all test materials, used or unused, following testing.

Scoring and analysis of data

Several different types of analyses were conducted to address the study's primary and secondary research questions. Findings from each analysis were intended to be used in combination to provide evidence about the effectiveness of linguistic modification in increasing student access to tested content.

The primary research question was examined using a series of analyses of variance (ANOVAs). The outcome variable was student math performance as measured by four scoring approaches commonly used by states in analyzing performance data: (1) raw scores (number correct total), (2) one-parameter logistic item response theory (IRT) model (1-PL), (3) two-parameter logistic IRT model (2-PL), and (4) three-parameter logistic IRT model (3-PL). Each approach has particular strengths and limitations and yields different types of information about student performance. In particular, the 1-PL, 2-PL, and 3-PL IRT models are widely used by states for scoring tests comprised of multiple-choice items (and the students' responses are coded either correct or incorrect). The authors of this study included these three IRT models so that strategies for estimating students' underlying abilities through a test would be consistent with states' practices. These four approaches are briefly discussed below, with a focus on their differences.²²

Four scoring approaches

The raw score approach used summed raw scores (number correct) across items in each item set. In this approach, each item, regardless of its difficulty level, is weighted equally. For example, a raw score of 5 from Student A and raw score of 5 from Student B indicate that they have the same level of math understanding even if Student A answered five relatively easy items correctly and Student B answered five relatively difficult items correctly. One of the benefits of reporting raw scores is that they are more easily interpretable than IRT theta scores, as the raw score represents the number of items answered correctly by each student. Moreover, raw scores are commonly used by states for reporting strand- or subdomain-level subscores, such as subscores on number sense and operations.

In contrast, the IRT models allow each item to have its own difficulty level as well as other characteristics. Following the example above, this means that the math understanding level of Student A likely would not be the same as that of Student B. Using the 3-PL model as an example, it takes the following mathematical form (Hambleton & Swaminathan 1985):

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (1)$$

where $P_i(\theta)$ is the probability that a student with ability level θ (theta) answers item i correctly, b_i is the item difficulty parameter for item i , a_i is the item discrimination parameter for item i , c_i is the item pseudo guessing parameter for item i , and D is 1.7 (a scaling factor).

From equation 1, each item under the 3-PL model is characterized by a difficulty parameter, discrimination parameter, and the pseudo guessing parameter. If guessing is assumed to be minimal (or none), the 3-PL model is equivalent to the 2-PL model. If the item discriminating power is assumed to be equal across items, the 2-PL model becomes the 1-PL model.

²² Detailed discussion can be found in Thissen & Wainer (2001).

Equating process in item response theory estimates

Equation (1) also indicates that, for example, adding a constant to both the theta and difficulty parameter would yield the same probability (since they need to be estimated on the same metric). This is a well known identification problem in IRT. To remove this indeterminacy, it is necessary to impose constraints on the parameters, either on the theta distribution or on the item parameter estimates. This is true regardless of the number of test forms administered to the students.

In the current study, it was necessary to equate the two item sets (one with the original items and one with the linguistically modified version of those items) so that comparisons across item sets would be meaningful.²³ To compare item parameter estimates for other noncommon items or to compare student math understanding estimates across forms, it is necessary to perform test equating so that the parameter estimates for both items and persons will be placed on a common metric.

Three equating designs are commonly used in the educational measurement field: random group equating design, single group design with counterbalancing, and common-item nonequivalent groups equating design (Kolen & Brennan 2004). The current study relied on a random group equating design to equate the original and linguistically modified item sets. Because in the current study the two item sets were administered randomly within each participating classroom, the mean and the standard deviation of the latent math understanding distribution for students who completed the original item set would be expected to be the same as that for students who took the linguistically modified version. While in this study the two item sets were administered randomly to the students within each English language learner (EL), non-ELA-proficient non-EL (NEP), and ELA-proficient non-EL (EP) student subgroup, the EP students were selected as the reference group because it was assumed that there would be no effects of linguistic modification on the math performance of EP students.

The following two-step procedure was used for each IRT model:

- *Step 1.* For each item set a separate estimation was conducted using EP students. The constraint was imposed that in each of these two estimations (one subset of EP students took the original item set while another subset of EP students took the linguistically modified item set), the mean and the standard deviation of latent (theta) distribution were the same (mean of 0 and standard deviation of 1). As indicated above, this constraint is consistent with the assumption that there will be no observed effects from

²³ For example, assume a particular item (Item A) was included in two different test forms (Form A and Form B). Also assume that these two forms comprising some common items were then administered to two different student samples (X and Y) from a larger student population. In IRT, the parameter estimate associated with Item A should be the same from these two separate estimations (one was based on X and another was based on Y), regardless of the IRT model used. The reason why parameter estimates for Item A may be different across these two samples/test forms is that the two estimations are not based on the same metric. To compare item parameter estimates for other noncommon items or to compare student math understanding estimates across samples X and Y, it is necessary to equate tests so that the parameter estimates for both items and persons are placed on a common metric.

linguistic modification on the performance of this particular student subgroup. The imposed constraint also removed the indeterminacy of IRT scales as noted above.

- *Step 2.* The derived item parameters from step 1 (one set for the original items and one for the linguistically modified items) were then used to estimate student theta scores (levels of math understanding) for the other two student subgroups (EL and NEP students). In other words, the item parameter estimates derived from EP students were treated as fixed and used to estimate theta scores for EL or NEP students taking either item set. The theta estimates for EP students were obtained directly from step 1.

After equating was completed, the resulting theta scores for the three student subgroups and item parameter estimates (derived with EP students as the reference group) were placed and reported on a common scale on which the origin of the scale was set to the mean (zero) of EP students. Therefore, the resulting theta scores reflect each student's level of math understanding relative to the average level of math understanding of EP students in the sample. Equating in this manner does not affect analyses associated with the primary research question about differential impacts of linguistic modification across student subgroups. The comparison of subgroup differences is not altered by linear transformations of the scale.

The ConQuest software (Wu, Adams, Wilson, & Haldane 2007), developed primarily for Rasch models, was used to estimate the 1-PL model parameters. The PARSCALE software (Muraki & Bock 2003) was used for the 2-PL and 3-PL models. These two proprietary computer programs are commonly used in the field of educational measurement. The resulting item parameter estimates for each model are reported in appendix G.

Primary analysis: analysis of variance (differences in linguistic modification impact across three student subgroups)

As noted above, a series of ANOVAs based on four common scoring approaches was conducted to address the primary research question about the impact of linguistic modification on student performance. Each of the four ANOVAs was applied to the full 2 x 3 x 2 crossed factorial design, where item set (original or linguistically modified), student subgroup (EL, NEP, or EP students), and grade level (7 or 8) were the three factors. The dependent variable in each case was student performance, as measured by raw score total or theta estimates of math understanding from the 1-, 2-, and 3-PL IRT models.

The interaction between student subgroup and item set was of particular interest in these ANOVAs because it addressed the study's primary research hypothesis. If a significant interaction between item set and student subgroup emerged in any of the four ANOVAs, three post hoc comparisons of the six cell means (2 item sets by 3 student subgroups) were planned: EL and NEP, EL and EP students, and NEP and EP students. The purpose of conducting these post hoc comparisons was to further examine where the significant differences occurred. If, for example, the average score from the linguistically modified item set was higher than the average score from the original item set for EL students, this difference was positive and more pronounced for EL students than EP students, and the subsequent post hoc comparison between EL and EP students was statistically significant, it would suggest that linguistic

modification had a larger impact on the math performance of EL students than EP students. A Bonferroni method was used to adjust the significance level because of the multiple comparisons.²⁴ Prior to the post hoc analyses, cell means were adjusted to remove main effects so that the interaction hypothesis could be tested accurately (Marascuilo & Levin 1970; Graham 2000). A profile plot of the original cell means also was produced to depict both the main and interaction effects of item set by student subgroup.

Summary findings from all four analyses are presented in chapter 4 to provide a complete picture of the effect of linguistic modification on student math performance. A detailed ANOVA summary table (along with the post hoc analyses and the profile plot when the interaction effect was statistically significant) for each approach is provided in appendix I. Challenges associated with interpreting accommodation effect across four approaches are discussed in chapter 5.

Secondary analyses

Five types of secondary analyses based on the raw score approach were performed. The purpose of conducting these secondary analyses was to provide additional information about linguistic modification and its effect using different sources of information.²⁵

Classical item-level descriptive analyses. Item-level statistics for each item set were generated to describe the item-level performance of each student subgroup as well as to examine the differences of two item sets. For each of the three student subgroups, reported statistics included percent correct (item p -value) and point biserial correlations (item-total correlations). The omission rates also were examined.

Reliability analyses. To explore interitem relationships on each item set, a Kuder-Richardson reliability (KR-20) coefficient (appropriate for the dichotomously scored items) was estimated for each of the three subgroups. These analyses were conducted to ascertain the extent to which linguistic modification was associated with changes in internal consistency reliability estimates among three student subgroups.

Analysis of differential item functioning. For each item set an analysis of differential item functioning (DIF) was conducted to detect subgroup differences in performance on any item that could not be explained by subgroup differences in the targeted math construct (math understanding). An item exhibiting significant DIF could indicate that the set of items in the study was assessing a construct other than the targeted one for a particular student subgroup. In that case, a source of difference associated with group membership other than the targeted construct may be contributing to differential subgroup performance on an item. If this additional construct is not relevant to the targeted construct assessed, the item could be biased.

²⁴ The Bonferroni correction is used in multiple comparison procedures to calculate an adjusted probability of comparison-wise type I error from the desired probability of family-wise type I error (Myers & Well 2003). The calculation ensures that the adjusted probability is equal to or less than the desired level (for example, 0.05). This method is considered to be conservative but is relatively easy to compute.

²⁵ The authors intended to compare findings from the four scoring approaches only in the primary analyses.

DIF analyses were conducted using the Mantel-Haenszel (MH) procedure (Holland & Thayer 1988). In these analyses, the EP student subgroup served as a reference group while the EL and NEP subgroups were the focal groups. For each item, findings presented include the MH chi-square statistic, a test of significance of MH chi-square value, the odds ratio, and the classification class. A significance level of .01 was used to flag items with significant DIF.²⁶

The DIF classification system used in this study was the delta metric, developed by the Educational Testing Service and commonly used to study DIF in the field of educational testing (Zwick, Thayer, & Lewis 1999). This delta scale was computed and transformed from the odds ratio, with typical values ranging from -3 to $+3$. It can be viewed as the average difference in difficulty of the studied item between comparable members (that is, with similar achievement levels) of the reference and the focal groups on the delta scale. A value of 0 means no differential difference in difficulty between the student groups; a positive value means that the item was differentially more difficult for the reference group (EP students) and therefore favored the focal group (NEP or EL students); and a negative value means that the item was differentially more difficult for the focal group (NEP or EL students) and therefore favored the reference group (EP students).

The ETS further developed a method to classify items based on their delta scores (Zieky 1993). If the delta score is not significantly greater than 0, or if the absolute value of the delta score is less than 1, the item is said to exhibit no or negligible DIF (class A); if the delta score is significantly different from 0 and its absolute value is between 1 and 1.5, DIF is considered to be moderate (class B); and if the delta score is significantly greater than 1 and its absolute value is at least 1.5, DIF is classified as moderate to large (class C). This classification—the magnitude of DIF—was used as a measure of effect size to judge whether an item exhibited enough DIF to warrant further examination.

In keeping with standard practice in states in reviewing performance data, the authors determined that any item demonstrating moderate to large DIF (that is, an item with a classification level C) would be subjected to further review by the study work group for potential sources of bias, such as a graphic that requires knowledge that cannot be assumed to be a part of every student's background knowledge. As is conventional for state testing programs, content (math) and population (EL students) experts would be asked to review all items flagged for moderate to strong DIF (class C) to determine if, in their judgment, any potential sources of bias were present. In typical state test development practice, any item judged to be biased (that is, any item that provides unfair advantage to a particular subgroup over another) would be removed from the final operational test. In this study, however, all primary and secondary analyses were conducted using the full item sets (all 25 items in each set) regardless of DIF findings; that is, the full item sets were treated as the intervention itself. Excluding any item after data were collected and analyzed was believed to potentially threaten the integrity of the intervention.

²⁶ Findings based on the conventional .05 level tend to flag items that do not exhibit sources of differences of primary concern to test developers (Hambleton, Swaminathan, and Rogers 1991).

Factor structure of the item sets. Factor structure analyses were intended to examine whether the underlying math understanding dimensions measured by the original and linguistically modified items were the same for EL, NEP, and EP students. That is, if multiple dimensions (such as math understanding and reading ability) are found, do the correlations among latent factors differ across the three student subgroups? In addition (and regardless of the number of underlying dimensions), do the correlations between latent factors and the associated items differ across the three student subgroups? Finally, do these findings differ between the two item sets?

For each item set, exploratory factor analysis (EFA) was conducted for each student subgroup to estimate the number of constructs assessed by the item set and to describe the underlying measurement structure of the unobservable (latent) factors. The results served as the foundation for a series of nested confirmatory factor analyses.

Confirmatory factor analyses (CFA) were performed to test for differences in both factor and measurement structure across the three student subgroups. The factor structure was examined to better understand how the underlying factors related to one another if more than one dominant factor was identified in EFA; the measurement structure was examined to explore correlations between the associated items and the underlying factors. Using the one-factor solution as an example, for each item set, three nested models were used to examine which measurement structure would best describe the data:

- *Model 1.* Fully unconstrained model in which factor loadings were allowed to vary across student subgroups (the baseline model).
- *Model 2.* Partially constrained/partially unconstrained model in which factor loadings were the same for NEP and EP students but were allowed to differ for EL students (were free to be estimated).
- *Model 3.* Fully constrained model in which factor loadings were the same across the three student subgroups.

Using the standard chi-square difference test, the resulting model-fit statistics were used to test which model best fit the data. If the EFA indicated that there was more than one dominant factor, then multiple-factor models would be tested using CFA. This multiple-factor CFA allows for examination of differences in the factor structure as well as in the measurement structure.

Mplus software (Muthén & Muthén 2006) was used for all factor analyses.

Analyses of test correlations. To examine whether linguistic modification alters the underlying construct being measured (math understanding), the raw score totals of EP students were correlated with their scores on the state standardized math test. A similar correlation between performance on each item set and the score on the standardized math assessment would be expected if the underlying constructs measured by each were comparable. In other words,

correlations of similar magnitude would suggest that the types of linguistic modification used in this study did not alter the validity of math assessment. These analyses focused on the subgroup of EP students because state test scores were expected to be most valid and reliable for this student subgroup and because linguistic modification was expected to have limited, if any, effect on this subgroup's performance on the item sets.

These analyses were conducted separately for students in grades 7 and 8 because they take different state math achievement tests. Simple Pearson correlation coefficients were estimated. To statistically compare differences in correlations between the original and linguistically modified items sets and associated standardized math achievement scores, Fisher's z -transformation tests were used.

Missing data

The analyses relied on two sources of data: data collected from students during operational testing (administration of the two item sets) and archived test scores for all tested students provided by district data managers. Consistent with practice in standardized test administration with dichotomously scored data, items with no response on the 25-item test were treated as incorrect and were coded as zero in the analyses. To be included in the analytic sample, students had to have attempted at least one item.²⁷ Because archived district data were used to define the analytic sample and the subgroups, students with missing archived data were not included in the analytic sample.

²⁷ This is in keeping with standard practice across states.

4. Study results

A combination of analyses was planned and implemented to address the study's primary and secondary research questions. These are described below. Results are presented for the analysis of the study's primary research question, which asks about differences in linguistic modification impacts across student subgroups. As discussed in chapter 3, findings that address the study's primary research question are reported from a series of analyses of variance (ANOVAs) using performance data derived from four different approaches: raw score total and estimates from a one-parameter logistic item response theory (IRT) model (1-PL), a two-parameter logistic IRT model (2-PL), and a three-parameter logistic IRT model (3-PL). Next, results are presented for the analyses of the secondary research questions, including classical item-level analyses, test reliability analyses, differential item functioning (DIF) analyses, factor analyses, and correlations of original and linguistically modified item set scores with state standardized achievement scores.

Primary analysis: differences in the impact of linguistic modification across student subgroups

- Does the effect of linguistic modification on students' math performance (as measured by raw scores or IRT theta estimates) vary across the three subgroups of students (English language learner [EL] students, non-English-language-arts-proficient non-EL [NEP] students, and English language arts-proficient non-EL [EP] students)? If so, did the linguistic modification improve the math understanding scores for the EL and NEP students relative to the EP students?

To test the hypothesis that the linguistic modification impact on student math performance differs across student subgroups, four three-way ANOVAs (item set, subgroup membership, and grade level) using raw score totals or theta estimates from the 1-PL, 2-PL, and 3-PL IRT models were conducted. Each analysis included three main effects, three two-way interaction effects, and one three-way interaction effect. The main research question is addressed by the interaction of item set and student subgroup. This interaction captures the extent to which the differences in student performance on the linguistically modified and original item sets vary across the three student subgroups.

Table 6 presents mean scores on each item set and the original/linguistically modified score differences, by scoring method (raw score or IRT 1-, 2-, or 3-PL model) and student subgroup. The data in the first two columns describe mean scores (with the corresponding standard deviations) based on the original and linguistically modified item sets, respectively. The data in the third and fourth columns describe the average differences in scores on the two item sets in the observed units and in standard deviation units, respectively. These differences represent estimates of the linguistic modification impacts for each student subgroup. The data in the fifth and sixth columns show significance test results for the interaction effect of item set by student subgroup. Detailed descriptive statistics from each scoring approach for each item set, student

subgroup, and grade level are presented in appendix H. ANOVA summary tables for all four approaches are provided in appendix I.

Table 6. Mean item set scores and score differences by scoring method, item set, and student subgroup

	Original item set	Linguistically modified item set	Difference	Effect size ^a	Global significance (item set by student subgroup)	<i>p</i> -value ^b
Raw score						
EL	8.40 (3.52)	9.16 (3.91)	0.76	0.15		
NEP	10.23 (3.85)	10.69 (4.05)	0.46	0.09	No	0.057
EP	15.59 (4.66)	15.63 (4.58)	0.04	<0.01		
1-PL model						
EL	-1.53 (0.49)	-1.37 (0.57)	0.16	0.17		
NEP	-1.14 (0.54)	-1.05 (0.59)	0.09	0.10	Yes	<0.01
EP	0.00 ^c (0.93)	0.00 ^c (0.90)	0.00 ^c	—		
2-PL Model						
EL	-1.24 (0.61)	-1.13 (0.68)	0.11	0.12		
NEP	-0.94 (0.66)	-0.88 (0.70)	0.06	0.07	No	0.15
EP	0.00 ^c (0.88)	0.00 ^c (0.87)	0.00 ^c	—		
3-PL model						
EL	-1.33 (0.65)	-1.24 (0.74)	0.09	0.09		
NEP	-0.99 (0.71)	-0.92 (0.75)	0.07	0.07	No	0.24
EP	0.00 ^c (0.89)	0.00 ^c (0.88)	0.00 ^c	—		

— is not applicable. EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

Note: Numbers in parentheses are standard deviations.

a. This standardized mean difference is derived by subtracting the mean raw score total or IRT theta estimate for the original item set from the mean raw score total or IRT theta estimate for the linguistically modified item set (the resulting difference is listed in column 3) and dividing by the standard deviation of the total student group (all three subgroups pooled) for the original item set. The total group standard deviation is used so that the resulting standardized difference is more comparable to the estimates reported by Hill et al. (2008).

b. Probability that linguistic modification impact difference across EL, NEP, and EP student subgroups are due to chance factors.

c. EP scores are constrained to be 0.00 and equal across items sets (that is, the linguistic modification impact is assumed to be zero).

Source: Authors' analyses of primary data.

Note that the raw score entries in the first two columns of table 6 represent the average number of items in each item set answered correctly. The IRT entries, however, represent subgroup *differences* in estimated thetas (math understanding) between the EL or NEP subgroups and the EP group. Because the EP group was used as the reference group to equate the original and linguistically modified items sets, the IRT-based estimates (columns 1–3) in the table are all expressed relative to the EP group. Moreover, the models in which the estimates were obtained assume that there is no effect of linguistic modification for the EP group.

As shown in table 6, the statistical significance of estimated differences across EL, NEP, and EP subgroups in the effects of linguistic modification on students' math performance depended on the scoring approach used (that is, how the scores for each student were calculated or estimated). For the 1-PL model only, the mean difference between the original and the linguistically modified item sets differed significantly across student subgroups ($p = .005$). For the other three approaches (raw scores, 2-PL model, and 3-PL model), the mean difference between the original and the linguistically modified item sets did not vary significantly across student subgroups.

Table 6 also presents effect sizes that describe the magnitude of the difference—or degree of practical importance—between the original item set and the linguistically modified item set. For EL students, the effect size was 0.15 standard deviation units based on the raw score approach and 0.17, 0.12, and 0.09 standard deviation units based on the 1-, 2-, and 3-PL models, respectively. In all cases, this magnitude of effect was larger than expected, given the 0.03 mean effect size for linguistic modification cited in a recent meta-analysis of accommodations for EL students (Kieffer et al. 2009).

There are several potential reasons why the statistical significance of estimated linguistic modification impact differences across EL, NEP, and EP students depended on the scoring approach used. Although the linguistic modification effect sizes are similar for the raw score estimates and the 1-PL model estimates—the raw score item-set-by-subgroup interaction is not statistically significant at the 0.05 level of significance ($p = 0.06$). This difference in statistical significance between the raw score and 1-PL model may have arisen because scores based on 1-PL IRT are more reliable and accurate than raw scores, as the former take into account item difficulty (Allen & Yen 1979). Because of this, the student math understanding estimates obtained from the 1-PL model may be less adversely affected by random error, thereby increasing statistical power to detect group differences in the effect of linguistic modification.

It is unclear exactly why the linguistic modification effect sizes were smaller when scores were based on 2-PL and 3-PL models than when they were based on the 1-PL model. If the data were not consistent with the assumptions under the 1-PL model, the 1-PL model might be expected to produce less accurate math understanding estimates than more complicated IRT models because it assumes that guessing is irrelevant and that items do not discriminate differentially between students with low- and high math understanding. However, the sample size in the current study may have been too small to obtain reliable 2-PL and 3-PL estimates (Crocker & Algina 1986; Embretson & Reise 2000; Reckase 1979; Hambleton & Jones 1993; Harris 1989; Thissen & Wainer 2001). Relatedly, 2-PL and 3-PL models require estimation of parameters based on the information provided by the student responses on items. This could be

intractable for particularly “easy” and “hard” items because limited information about those students who did not answer the item correctly (or incorrectly) was available during the estimation process. The limited sample size and dependencies in the data for specific items may have reduced the precision of the 2-PL and 3-PL math understanding estimates, reducing the statistical power to detect group differences in the effect of linguistic modification. Appendix J provides a more complete discussion of these issues.

Based on the 1-PL model math understanding estimates, a post hoc comparison of the interaction effect of item set by student subgroup was conducted to further examine how the effect of linguistic modification varied across the three student subgroups. The results in table 7 show that the item-set-by-student-subgroup interaction resulted mainly from the difference in linguistic modification effects between EL students and EP students (adjusted $p = .003$).

Table 7. Post-hoc comparison of interaction effect (based on 1-PL model)

Comparisons	Value of contrast	Standard error	<i>t</i> -statistic	Degrees of freedom	<i>p</i> -value ^a	Adjusted <i>p</i> -value ^b
EL vs. NEP	.07	.051	1.373	4605	.170	.510
NEP vs. EP	.09	.046	1.951	4605	.051	.153
EL vs. EP	.16	.050	3.198	4605	.001	.003**

** Significant at $\alpha = .01$ level. EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts–proficient students who are not English language learners.

Note: Tests were performed assuming equal variances. EL = English language learner students; NEP = non-English language arts proficient students who are not English language learners; EP = English language arts proficient students who are not English language learners.

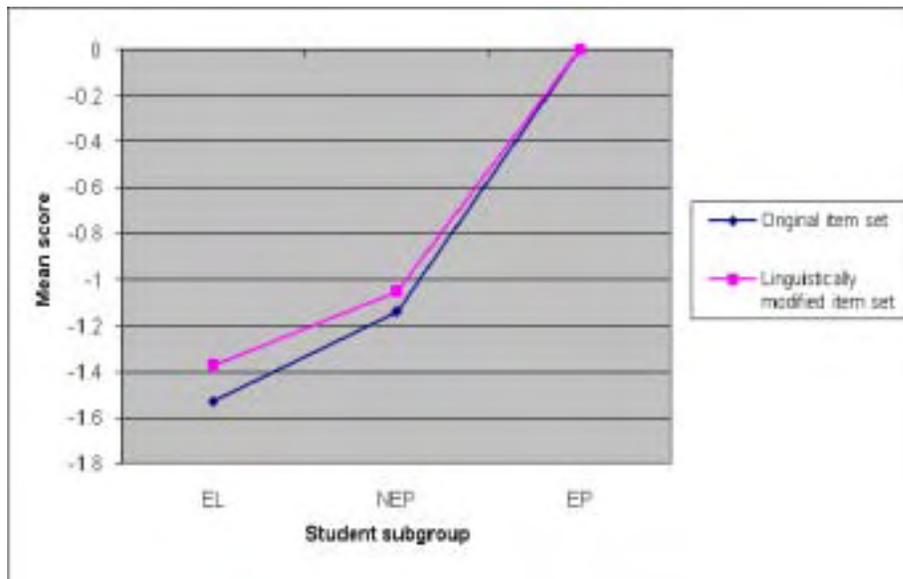
a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true

b. Adjusted using Bonferroni method for multiple comparison adjustment (three comparisons).

Source: Authors’ analyses of primary data.

Figure 4 shows the mean scores for each subgroup on each item set. The means are constrained to be zero for the EP students on both tests, according to the equating process described in chapter 3. The effect of linguistic modification is seen in the difference between the original and the linguistically modified items that occur for the other two subgroups (their scores are reported on a scale relative to the highest scoring group). As hypothesized, the magnitude of the difference between original and linguistically modified item sets was greater for the EL student subgroup than for the EP subgroup. The small difference between EL and NEP students in the effect was not statistically significant, nor was the difference between the two non-EL groups.

Figure 4. Profile plot of cell means, by item set and student subgroup (based on 1-PL model)



Significant at $\alpha = .01$ level. EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

Source: Authors' analyses of primary data.

Because there is no universal guideline for evaluating the practical importance of a standardized effect size estimate for an educational intervention, it also is useful to compare this estimate to another empirical benchmark that reflects changes in student academic achievement. A standardized difference of 0.17 for EL students based on the 1-PL model, for example, is more than half the magnitude of growth in math achievement that might be expected from one full year of schooling (.32), as measured by a standardized test (Hill et al. 2008).²⁸

Finally, as shown in table 6, despite inconsistent significance test results across the four scoring approaches, the mean differences in student performance on the two item sets for each student subgroup exhibited a consistent trend. The mean difference in performance on the two items sets was greatest for EL students, followed by NEP students. For EP students, the difference in scores on the linguistically modified and original item set was very close to zero (less than 0.01 standard deviation units) based on the raw score approach.²⁹

²⁸ This comparison is provided because it provides perspective on the finding. However, it should be noted that the linguistic modification strategies used in the current study were applied to a set of items that were judged to be amenable to linguistic modification. Because a typical standardized math test is likely to include some portion of items that may not be amenable to linguistic modification, this finding should be interpreted with caution.

²⁹ As discussed in chapter 3, the difference in theta scores between two item sets for EP students was set to zero for the item set equating purpose.

Secondary analyses

This section presents the results of classical item-level descriptive analyses, reliability analyses, analysis of differential item functioning (DIF), factor analyses, and correlation analyses.

Classical item-level descriptive analyses

Item descriptive information (percent correct and point biserial correlations) for the original and the linguistically modified item sets are presented in appendix K. Table 8 presents the mean percent correct across all items for the three student subgroups for the original and the linguistically modified item sets.

Table 8. Mean percent correct (item p -value) and the associated standard deviation across all items, by student subgroup and item set

Student subgroup	Original item set mean (standard deviation)	Linguistically modified item set mean (standard deviation)
EL	.34 (.14)	.37 (.14)
NEP	.41 (.17)	.43 (.17)
EP	.62 (.17)	.63 (.16)

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners. *Source:* Authors' analyses of primary data.

Omission rates also were examined. On both item sets, the omission rates ranged from 0 percent to 6 percent, with the highest percentage occurring on item 12 for both item sets. This item appeared to be a relatively difficult item on both item sets for all student subgroups. Omission rates suggest that participating students were able to complete the item set within the testing window (50-minute class period).

Reliability analyses

Table 9 summarizes internal consistency reliability (estimated by KR-20) by each item set and student subgroup.

Table 9. Internal consistency reliability coefficient, by student subgroup and item set

Student subgroup	Original item set	Linguistically modified item set
EL	.61	.68
NEP	.67	.70
EP	.79	.78

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners. *Source:* Authors' analysis of primary data.

Regardless of item set, internal consistency appeared to be highest for the EP students. Also, the reliability estimates on the original item set and the linguistically modified item set were very close for the EP students (.79 and .78, respectively). Since the item sets were believed to be more reliable (that is, items were more tied to the underlying math construct) for the EP students and there would be no (or minimal) effect of linguistic modification on this student subgroup, this finding was expected. Furthermore, for both EL students and NEP students and for EL students in particular, the internal consistency measures tended to be higher for the linguistically modified item set (.68 and .70, respectively) than for the original item set (.61 and .67, respectively). These findings suggest that linguistic modification may have affected the functioning of items among students in these two student subgroups.

Differential item functioning

- For each item set, do any items exhibit DIF between EL students and EP students and between NEP students and EP students? How do the DIF findings differ between the original and linguistically modified item sets? In other words, when comparing both EL and NEP students with EP students with similar math achievement levels, do the probabilities of the students answering individual items correctly differ on the test with linguistically modified items as compared with the test with original items? Does linguistic modification reduce the number of items showing DIF? Findings from these questions are of interest because an item showing DIF may be measuring something other than the construct of interest (math understanding).

DIF analyses were conducted using the Mantel-Haenszel (MH) procedure (Holland & Thayer 1988). In these analyses, the EP student subgroup served as a reference group while the EL and NEP subgroups were the focal groups. For each item, findings presented include the MH chi-square statistic, a test of significance of MH chi-square value, the odds ratio, and the classification class. As discussed in chapter 3, the authors flagged items that fell into class C (moderate to large DIF) for further review.

The detailed DIF findings are summarized and presented in appendix L. As shown in table L1 in appendix L, when comparing the NEP and EP subgroups, items performed similarly across both item sets and no items demonstrated moderate to high DIF.

In contrast, when comparing EL students with EP students (table L2 in appendix L), one item in the original set (Item 8) and two items in the linguistically modified set (Items 16 and 23) demonstrated moderate to high DIF (class C).

In keeping with standard practice for state test development, those items demonstrating moderate to strong DIF (Items 8, 16, and 23) were flagged for further review by content and language experts (study work group) for potential sources of bias, as described in the section on analysis of DIF in chapter 3. Using expert judgment, this panel of experts reviewed the items for potential sources of bias, including the type of challenge associated with the target mathematics content assessed in the item (math load), the degree of complexity of the language in the item (language load), the source of the item (National Assessment of Educational Progress or California Standards Test), and the linguistic modification strategy used in the item. No evidence of item bias was detected.

Factor structure of the item sets

- Does the number of factors that underlie student responses to an item set (original or linguistically modified) differ for EL, NEP, and EP students? Do item-factor relationships differ across the three student subgroups? If more than one factor underlies performance on each item set, do the correlations among factors differ across the three student subgroups? Does linguistic modification reduce the number of factors or affect the item-factor relationship? Answers to these questions would help to evaluate: (1) the number of underlying factors in each item set by student subgroups; (2) for each item set and student subgroup, the relationship between the underlying factors and the associated items (“measurement structure”); (3) the correlations between the underlying factors (“factor structure”) if more than one factor was identified; and (4) the extent to which linguistic modification changed the measurement structure and/or the factor structure for EL and NEP students relative to EP students.

For each item set within each student subgroup, a series of exploratory factor analyses (EFAs) were conducted to examine the number of underlying factors and a series of nested confirmatory factor analyses (CFAs) were conducted. If more than one dominant factor was identified in EFA, a series of nested CFAs would have then been conducted to test for differences in measurement and factor structure across student subgroups. The EFA results for both the original and the linguistically modified item sets suggest that there was one dominant factor in each student subgroup. For the original item set, responses of EP students resulted in a first factor with an eigenvalue of 6.9 (explaining 27 percent of the total variance), while responses for NEP students resulted in a first factor with an eigenvalue of 4.4 (explaining 18 percent of the total variance). Responses for EL students resulted in a first factor with an eigenvalue of 3.9 (explaining 16 percent of the total variance). Scree plots and factor loadings for all items are provided in appendix M.

The results were similar for the linguistically modified item set, although the eigenvalue for the first factor for EL students increased to 4.7 (explaining 19 percent of the total variance). Overall, for all student subgroups and for the original and linguistically modified item sets, the

EFA results suggest that one dominant construct was measured with the 25 item sets, though the strength of that factor may vary across subgroups.

To learn more about the nature of this dominant construct, the EFA results served as the foundation for a series of nested CFAs, which tested for differences in measurement structure (factor loadings) across student subgroups. Since the EFA results identified one dominant factor for each item set per student subgroup, the multiple-group CFAs were based on a one-factor structure.

As indicated in chapter 3, for each item set, three models of interest were compared using the chi-square test for testing difference: model 1, a fully unconstrained model in which all the parameter estimates were allowed to differ across student subgroups; model 2, a partially constrained, partially unconstrained model in which the parameter estimates were allowed to differ for EL students (but the parameter estimates were constrained to be the same for the two non-EL subgroups); and model 3, a fully constrained model in which all the parameter estimates were constrained to be the same across student subgroups.

For the original item set, both models 2 and 3 resulted in a decrease in model fit (chi-square difference = 67.44, degrees of freedom = 21, and $p < .01$ for model 1 compared with model 2; chi-square difference = 113.38, degrees of freedom = 40, and $p < .01$ for model 1 compared with model 3). This suggests that the measurement structure was not the same across student subgroups. For the fully unconstrained model for EP students, almost every item (except item 2) was tied to the underlying factor, with a loading of .32 or more.³⁰ For NEP students, however, 9 of the 25 test items had loadings below .32. Similarly, for EL students, 11 items had loadings below .32.

These results suggest that the relationships between the underlying factor and many of the test items were weaker for EL students and, to a lesser extent, for NEP students than for EP students. This is consistent with the findings from the reliability analysis (see table 9) in that the internal consistency measure for EL students was the lowest among the three student subgroups.

The findings were similar for the linguistically modified item set. A comparison of the three nested models suggests that the correlations between the items and the underlying factor were different for the three student subgroups (chi-square difference = 65.97, degrees of freedom = 20, and $p < .01$ for model 1 compared with model 2; chi-square difference = 150.54, degrees of freedom = 38, and $p < .01$ for model 1 compared with model 3). However, for NEP students, evidence suggests that linguistic modification improved the functioning of most of the 11 items in the original item set with loadings below .32 (for example, the factor loadings for items 3, 12, and 23 were now .32 or above on the linguistically modified item set).

³⁰ See Tabachnick & Fidell (2007; p. 649). Only variables with loadings of .32 and above (that is, those that explain 10 percent overlapping variance by the item to the underlying factor) are interpreted.

For EL students, evidence suggests that linguistic modification improved the functioning of some of the 11 items on the original form with loadings below .32 (for example, the factor loadings for items 16 and 23 were almost zero on the original item set but .32 or above on the linguistically modified item set). For two other items, however, the relationship to the underlying factor appeared to be weaker for the linguistically modified item set than for the original item set for EL students—the loadings for items 21 and 22 decreased to almost zero.

The CFA findings for the linguistically modified item set also were consistent with the reliability analysis; items were tied to the underlying factor more strongly for EP students (as expected given a higher internal consistency measure), with some improvement emerging with the linguistically modified item set. That is, the internal consistency estimate was higher in the linguistically modified item set than in the original item set for EL students and, to a lesser extent, for NEP students.

Overall, the CFA results suggest that the item sets had a different measurement structure for each of the student subgroups, and this result held for both the original and linguistically modified item sets.

Correlations between math raw scores from original and linguistically modified item sets, and standardized tests of math achievement

- For the EP students, do raw scores on the original and the linguistically modified item sets correlate similarly with scores from the state’s standardized tests of math achievement? This question was intended to examine the degree to which mathematics items can be linguistically modified to reduce language load without altering the construct intended to be assessed. If the correlation of item set raw scores with the standardized scores were similar for the original and linguistically modified item sets, it would support the assumption that the items had been linguistically modified without altering the target construct.

Table 10 displays correlations of item set raw score totals for EP students with their scores from the previous year’s California standardized achievement tests in math. Also presented are findings from the statistical test of equality of two correlation coefficients (original compared with linguistically modified item set).

Table 10. Correlations between item set raw score totals and state standardized math achievement test score, by grade, for non-English language learner students who were proficient in English language arts

Grade and measure	Correlation	<i>p</i> -value ^a
<i>Grade 7</i>		
Original item set (<i>N</i> = 434)	0.75	.607
Linguistically modified item set (<i>N</i> = 439)	0.74	
<i>Grade 8</i>		
Original item set (<i>N</i> = 443)	0.63	.945
Linguistically modified item set (<i>N</i> = 453)	0.63	

a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true

Source: Authors' analyses of primary data.

As shown in table 10, the statistical test was not significant at the .05 level for either grade level. These findings suggest that the types of linguistic modification used in this study did not alter the math construct assessed.

Summary of key findings from primary and secondary analyses

Considering the findings from both primary and secondary analyses, this study revealed:

- EL, NEP, and EP differences in the effect of linguistic modification across 25 items measuring math understanding varied, depending on the scoring approach, that is, how the scores for each student were calculated or estimated. When scores were estimated based on the 1-PL model, a significant difference in theta scores on the two item sets was detected across student subgroups (between EL students and EP students, in particular). This small but significant effect was not detected in the analyses based on raw scores or theta estimates from the 2-PL or 3-PL models.
- Despite inconsistent significance test results across the four scoring approaches, the mean differences in performance on the two item sets for each student subgroup showed a consistent trend—the mean difference in performance on the two item sets was greatest for EL students, followed by NEP students. For EP students, the difference in scores on the linguistically modified and original item set was very close to zero (less than 0.01 standard deviation units, based on the raw score approach).
- The effect size for EL students was 0.15 standard deviation units using the raw scores approach and 0.17, 0.12, and 0.09 standard deviation units for the 1-, 2-, and 3-PL models, respectively. Because there is no universal guideline for evaluating the practical importance of a standardized effect size estimate for an educational intervention, it is useful to compare this estimate with another empirical benchmark

that reflects changes in student math achievement. A standardized difference of 0.17 based on the 1-PL model, for example, is more than half the magnitude of growth in math achievement that might be expected from one full year of schooling (.32), as measured by a standardized test (Hill et al. 2008).

- The linguistic modification implemented in the current study did not alter the targeted math construct based on the DIF analysis, EFA, and the correlational analysis.
- Regardless of item set and student subgroups, the finding from EFA indicated that there appeared to be one dominant factor (math understanding) underlying the test.
- Even though there was only one dominant factor, findings from the reliability and CFA analyses suggested that the measurement structure between the underlying factor and the items differed across student subgroups.
- Both reliability and CFA analyses also suggested that the linguistic modification improved the functioning of some items for EL and NEP students. However, for EL students, CFA indicated that two of the 25 linguistically modified items were more weakly associated with the underlying factor than their original counterparts.

5. Interpretation of key findings, study challenges, and direction for future research

The effectiveness of current accommodation practices for making high stakes tests accessible, equitable, and valid for English language learner (EL) students is unclear (Butler & Stevens 2001; Castellon-Wellington 2000; Francis, Rivera, Lesaux, Kieffer, & Rivera 2006; Holmes & Duron 2000; Rivera & Stansfield 2001). Little empirical data are available about the effectiveness of test accommodations for enabling student access to tested content. As a result, policies on allowable accommodations for EL students remain inconsistent across states (Goh 2004; National Research Council 2004; Rivera & Collum 2004; Thurlow, Wiley, & Bielinski 2002). This study contributes to the body of knowledge informing the development of appropriate accommodation guidelines for EL students.

The study examined one accommodation (linguistic modification) as a means of improving EL student access to tested content. Additionally, the study examined the degree to which better access through linguistic modification strategies could reduce the effect of construct-irrelevant variance on test results. Implications of key findings for policymakers in the West Region, lessons learned, and recommendations for future research are discussed in the sections below.

Interpretation of findings from the primary analysis: interaction between student subgroup and item set

The study's primary research question asked whether the effect of linguistic modification on students' scores on the two sets of items (original and linguistically modified) varied across the three student subgroups. Four scoring approaches commonly used by states for scoring tests comprised of multiple-choice items were used in this study to estimate level of math understanding on each item set—one based on the raw score total and three based on item response theory models (a one-parameter logistic IRT model [1-PL], a two-parameter logistic IRT model [2-PL], and a three-parameter logistic IRT model [3-PL]).

Differences across EL, non-English-language-arts-proficient non-EL (NEP) students, and English language arts-proficient non-EL (EP) students in the effects of linguistic modification on student math performance depended on the scoring approach used. For the 1-PL model, the mean difference between the original and the linguistically modified item sets varied significantly across student subgroups; for the other three approaches, the mean difference between the original and the linguistically modified item sets did not vary significantly across student subgroups.

Despite the inconsistency in statistical significance when different scoring approaches were applied, the subgroup differences on the effect of linguistic modification were consistent across all four scoring approaches with the following trend—the mean difference for EL students was the largest, followed by NEP students, and the mean difference for EP students was very small

(based on raw score total). In addition, across the four approaches, the magnitude of the difference in mean scores between the original item set and the linguistically modified item set for EL students ranged from 0.09 (3-PL model) to 0.17 (1-PL model). Using the finding from the work done by Hill et al. (2008), this reflects between one-quarter and one-half the magnitude of growth in math achievement that might be expected from one full year of schooling (.32), as measured by a standardized test.

The magnitude of effect size found in this study (0.09–0.17) appears to be larger than the finding from the meta analysis on studying the effect of linguistic modification for EL students (Kieffer et al. 2009) in which the average effect size was found to be 0.03, regardless of significance test results.

Interpretation of findings from secondary analyses: impact of linguistic modification on construct assessed

Critical to the examination of the effect of this accommodation was the creation of instruments (that is, original and linguistically modified item sets) with sufficient technical rigor to be trustworthy and useful for the intended purposes. To validate the item sets for the purposes of this study, evidence was collected through expert judgment, cognitive interviews, and pilot testing. The final item sets were administered via random assignment to EL and non-EL students in grades 7 and 8 in sampled schools in one state. The steps taken in this study to maintain the integrity of the tested construct provide a useful framework for state policymakers in this region seeking assurances about the validity of results from accommodated assessments: linguistic modification can be applied without significantly altering the math construct intended to be assessed.

Findings from secondary analyses support the conclusion that the types of linguistic modification strategies used in the study did not alter the math constructs assessed. The findings of the EFA suggested that one dominant factor (math understanding) was measured by both item sets, although results from the CFA indicated that the measurement structure (the relationship between the underlying factor and items, represented by the factor loadings) varied across student subgroups for both item sets. Results from the correlational analyses using EP student scores also suggest that the types of linguistic modification strategies used in the study did not alter the math construct assessed.

The item set reliability analysis suggested that linguistic modification may have enhanced, or at least did not reduce, the internal consistency of the item set for EL and NEP students. Specifically, internal consistency appeared to be higher for the linguistically modified set of items than for the original set for both EL and NEP students. As a whole, the items in the linguistically modified set more reliably measured math understanding of students in the EL and NEP student subgroups than the items in the original item set. The findings from CFA also indicated that linguistic modification improved the reliability of that item set for EL and NEP students in that some items, but not all, were more closely tied to the underlying factor after linguistic modification.

Challenges related to the study context and design

There are known challenges in research on human subjects in schools. During the design and implementation phases of this study, researchers carefully weighed tradeoffs to ensure study integrity while minimizing the burden to students, teachers, and support staff at the school and district levels. For example, to lessen the burden on study participants, students were administered either the original or the linguistically modified item set; no student was asked to take both.

A number of factors, including compromises such as those described above, may limit the generalizability of findings beyond schools in the West Region. Specifically, though omission rates were quite low (0–6 percent) and attrition was monitored (see figures 2 and 3 and table 4), systematic differences may have emerged between participants and nonparticipants. Future research could consider a repeated measures design that administers multiple item sets of original and linguistically modified items to the same students over the course of the year, with each set containing items that assess a particular strand of content and apply appropriate linguistic modification strategies. Future research also could better control for students' opportunity to learn the content tested by the study's item sets (for example, calculations of distance, application of the Pythagorean Theorem) so that findings can be attributed with greater confidence to a change in access to the tested content (through linguistic modification) rather than to degree of exposure to the tested content. Additionally, future research could better control for student familiarity with the item types used in the study (for example, word problems) so that findings related to the effect of linguistic modification can be attributed with greater confidence to a change in access rather than to the degree of familiarity with an item type.

Challenges related to item selection and item set development

The validity of this study's findings hinged on demonstrating that the study's item pool included rigorously developed and psychometrically sound math items. Using extensively reviewed National Assessment of Educational Progress (NAEP) and California Standards Test (CST) items meant that the items eligible for use in the study already incorporated key elements of good test development practice, including the application of universal design principles.³¹ Researchers recognized that challenges might emerge since items most amenable to linguistic modification might not have been in the pool of original items considered for this study. The high technical quality of the original items used may help explain why the study found limited evidence of a strong accommodation effect and of differential item functioning (DIF). First, as described previously, all original items had already undergone extensive sensitivity and statistical review (including DIF analyses) before use in the NAEP or CST

³¹ Universal design, a concept that began in architecture and expanded to other fields, including education, supports participation of the widest possible range of students in large-scale assessments in a manner that results in valid inferences about performance. Elements of universally designed assessment include inclusive assessment population; precisely defined constructs; accessible, nonbiased items; amenability to accommodations; simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility (Thompson, Johnstone, and Thurlow 2002).

testing programs. Second, the workgroup that linguistically modified the items was trained to detect potential sources of bias in items and knew to exclude such items from the final original item pool. Finally, data from the cognitive interviews and item piloting were used to confirm that only those original items that demonstrated technical quality would appear in the final item sets.

Future research may want to more closely consider the quality of the original pool of items selected for linguistic modification studies. The items in the original pool used in this study had undergone rigorous reviews by specialists to ensure technical quality (such as content and age group appropriateness and freedom from bias), and these items most likely did not include those item and content characteristics that highly constrain EL students' access to tested content. Thus, to better observe the effects of linguistic modification on student access, future studies should determine a priori the critical language characteristics³² in particular needed in the original items used.

Building on findings from this study, further research is needed to explore in greater depth item-level analyses in relation to those linguistic modification strategies that were most and least effective in increasing student access to tested content, as measured by student performance on each item set. Item and content characteristics may interact with linguistic modification strategies, yielding certain items that perform as expected (for example, Item 5 as evidenced by a higher item *p*-value [percent correct] on the linguistically modified item set for EL students [.27, compared with .18]) but others (for example, Items 2 and 17) that did not show improvement in item *p*-value on the linguistically modified item set. Because of the inconsistencies that emerged in effectiveness of linguistic modification strategies at the item set level, additional research at the item level would help to determine which linguistic modification strategies work best and why. Further study of the conditions under which certain linguistic modification strategies were most effective and research that explains discrepancies, including how certain features of an assessment task might influence the effect of linguistic modification, have practical implications for the development of more valid and reliable measures of what EL students know and can do.

Other directions for future research

State educators and policymakers need empirical evidence on which to base decisions about providing valid, appropriate, and equitable assessments for students within the diverse EL population. Future research can build on the processes established in this study for reviewing and revising items (expert judgment, cognitive interviews) and for implementing the study design (recruitment plan, pilot testing of items) and continue to methodically examine the

³² Critical language characteristics impact language load and include “academic language.” Although there is not just one accepted definition of academic language, there are several resources that address the issue of academic language. See, for example, Aguirre-Munoz, Parks, Benner, Amabisca, & Boscardin 2006; Bailey 2007; Bailey, Butler, & Sato 2007; Butler, Bailey, Stevens, Huang, & Lord 2004; Chamot & O’Malley 1994; Cummins 1980, 2005; Halliday 1994; Sato 2007; Sato & Worth in press; Scarcella & Zimmerman 1998; and Schleppegrell 2001. Not all aspects of academic language in test items lend themselves to linguistic modification without altering the construct being assessed. Therefore, future research should consider implications of including more general versus academic language characteristics in items when examining the effects of linguistic modification.

effects of different linguistic modification strategies on the validity of measures of student academic content understanding.

One logical and important extension of this work would be to focus on those linguistic modification strategies that showed promise through in-depth item-level analyses rather than set-level analyses. As previously discussed, some of the linguistically modified items were more effective than others in improving student performance, as evidenced by findings from the internal consistency analysis and the CFA. But not all items were equally effective in improving student performance. This surfaces the possible limitation of set-level analyses—the impact of effective items could be negated by the impact of less effective items. In-depth item-level analyses, building on the cognitive interview protocols used in this study to test developers’ assumptions about access constraints and enablers, may help ensure that those promising linguistic modification strategies that deserve further study, in terms of how they may interact with item and content characteristics and address student access needs, are not overlooked.

Other directions for future research include examination of the appropriateness and effectiveness of linguistic modification in other math strands (such as algebra) and other content areas (such as science), as well as with students with different language backgrounds (such as Mandarin), literacy and English language proficiency levels, and grade levels. Finally, the degree to which other accommodations (such as use of dictionaries, chunking text) may further influence the effect of linguistic modification on student access to tested content and subsequent performance should be examined.

References

- Abedi, J. (1999, April). *Examining the effectiveness of accommodation on math performance of English language learners*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Abedi, J. (2001). *Assessment and accommodations for English language learners: issues and recommendations* (CRESST Policy Brief 4). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, 33(1), 4–14.
- Abedi, J. (2008). *Linguistic modification: Part I—Language factors in the assessment of English language learners: the theory and principles underlying the linguistic modification approach*. Washington, DC: LEP Partnership.
- Abedi, J., Courtney, M., and Leon, S. (2003). *Research-supported accommodation for English language learners 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., Courtney, M., Leon, S., Kao, J., and Azzam, T. (2006). *English language learners and math achievement: a study of opportunity to learn and language accommodation* (Technical Report 702). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.
- Abedi, J., Courtney, M., Mirocha, J., Leon, S., and Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: bilingual dictionaries and linguistic modification*. Los Angeles: University of California, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., and Dietel, R. (2004). Challenges in the No Child Left Behind Act for English language learners. *Phi Delta Kappan*, 85(10), 782–785.
- Abedi, J., and Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14(3), 219–234.
- Abedi, J., Hofstetter, C., and Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74(1), 1–28.
- Abedi, J., Lord, C., Hofstetter, C., and 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., and Plummer, J. (1995). *Language background as a variable in NAEP mathematics performance: NAEP task 3D: language background study* (CSE Technical Report 429). 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., and Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Technical Report 429). Los Angeles: University of California, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.
- Aguirre-Munoz, Z., Parks, J.E., Benner, A., Amabisca, A., and Boscardin, C.K. (2006). *Consequences and validity of performance assessment for English language learners: conceptualizing and developing teachers' expertise in academic language*. Los Angeles: University of California, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Available at <http://www.cse.ucla.edu/products/reports/R700.pdf>.
- Albus, D., Bielinski, J., Thurlow, M., and Liu, K. (2001). *The effect of a simplified English language dictionary on a reading test* (NCEO LEP Project Report 1). Minneapolis: National Center on Educational Outcomes.
- Allen, M., and Yen, W. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.
- Almond, P., Cameto, R., Johnstone, C., Laitusis, C., Lazarus, S., Nagle, K., Parker, C., Roach, A., and Sato, E. (2009). *White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS)*. Dover, NH: Measured Progress and Menlo Park, CA: SRI International.
- Assessment and Accountability Comprehensive Center. (2009). *Framework for high-quality English language proficiency standards and assessments*. San Francisco: WestEd.
- Bailey, A.L. (2007). *The language demands of school: putting academic English to the test*. New Haven, CT: Yale University Press.
- Bailey, A., Butler, F., and Sato, E. (2007). Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards. *Applied Measurement in Education*, 20(1), 53–78.
- Baker, C. (2001). *Foundations of bilingual education and bilingualism* (3rd ed.). Philadelphia, PA: Multilingual Matters Ltd.
- Bielinski, J., Sheinker, A., and Ysseldyke, J. (2003, April). *Varied opinions on how to report accommodated test scores* (NCEO Synthesis Report 49). Minneapolis: National Center on Educational Outcomes.

- Butler, F.A., Bailey, A.L., Stevens, R., Huang, B., and Lord, C. (2004). *Academic English in fifth-grade mathematics, science, and social studies textbooks* (Final deliverable to IES, Contract No. R305B960002; currently available as CSE Report No. 642). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Butler, F.A., and Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K–12: Current trends and old dilemmas. *Language Testing* 2001, 18(4), 409–427.
- California Department of Education. 2008a. CBEDS School Information Form (SIF). Retrieved February 2009 from <http://dq.cde.ca.gov/DataQuest/downloads/sifenr.asp>.
- California Department of Education. 2008b. English learners by grade and language. Retrieved February 2009 from www.cde.ca.gov/ds/sd/lc/fileselsch.asp.
- California Department of Education. 2008c. Free/Reduced Meals Program & CalWORKS Data Files. Retrieved February 2009 from www.cde.ca.gov/ds/sh/cw/filesafdc.asp.
- Castellon-Wellington, M. (2000). *The impact of preference for accommodations: the performance of ELLs on large-scale academic achievement tests* (CRESST Technical Report 524). Los Angeles: University of California, National Center for the Study of Evaluation, Standards, and Student Testing.
- Chamot, A.U., and O'Malley, J.M. (1994). *The CALLA handbook: implementing the cognitive academic language learning approach*. New York: Longman.
- Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich College Publishers.
- Cummins, J. (1980). The construct of proficiency in bilingual education. In J. E. Alatis (Ed.), *Georgetown University Round Table on Languages and Linguistics: Current issues in bilingual education* (pp. 81–103). Washington, DC: Georgetown University.
- Cummins, J. (1981). Four misconceptions about language proficiency in bilingual education. *NABE Journal*, 5(3), 31–45.
- Cummins, J. (2005). Language proficiency, bilingualism, and academic achievement. In P.A. Richard-Amato and M.A. Snow (Eds.), *Academic success for English language learners* (pp. 76–86). White Plains, NY: Longman.
- Embretson, S., and Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Ericsson, K., and Simon, H. (1980). Verbal reports as data. *Psychological Review*, 87, 215–250.

- Ericsson, K., and Simon, H. (1993). *Protocol analysis: verbal reports as data*. Cambridge, MA: Massachusetts Institute of Technology.
- Francis, D., Rivera, M., Lesaux, N., Kieffer, M., and Rivera, H. (2006). *Research-based recommendations for the use of accommodations in large-scale assessments*. Houston, TX: University of Houston, Texas Institute for Measurement, Evaluation and Statistics, Center for Instruction.
- Fuchs, L.S., Fuchs, D., Eaton, S.B., Hamlett, C.L., Binkley, E., and Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children*, 67, 67–81.
- Garcia, G.N. (2000). Lesson from research: What is the length of time it takes limited English proficient students to acquire English and succeed in an all-English classroom? *National Clearinghouse for Bilingual Education*, 5.
- Goerman, P. (2006). *Adapting cognitive interview techniques for use in pretesting Spanish language survey instruments* (Research Report Series, Survey Methodology 2006-3). Washington, DC: U.S. Census Bureau, Statistical Research Division.
- Goh, D.S. (2004). *Assessment accommodations for diverse learners*. Boston: Pearson.
- Graham, J.M. (2000, January 27–29). *Interaction effects: Their nature and some post hoc exploration strategies*. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.
- Halliday, M.A.K. (1994). *An introduction to functional grammar* (2nd ed.). London: Edward Arnold.
- Hambleton, R., and Jones, R. (1993). *Comparison of classical test theory and item response theory and their applications to test development*. An NCME Instructional Module on Instructional Topics in Educational Measurement. Iowa City, IA: American College Testing Program.
- Hambleton, R., and Swaminathan, H. (1985). *Item response theory: principals and applications*. Boston, MA: Kluwer-Nijhoff Publishing.
- Hambleton, R., Swaminathan, H., and Rogers, J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Harris, D. (1989). *Comparison of 1-, 2-, and 3-parameter IRT models*. An NCME Instructional Module on Instructional Topics in Educational Measurement. Iowa City, IA: American College Testing Program.
- Hill, C., Bloom, H., Black, A., and Lipsey, M. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172–177.

- Hofstetter, C.H. (2003). Contextual and mathematics accommodation test effects for English language learners. *Applied Measurement in Education*, 16(2), 159–188.
- Holland, W. P. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: LEA.
- Holmes, D., and Duron, S. (2000). *LEP students and high stakes assessment*. Washington, DC: U.S. Department of Education, National Clearinghouse for Bilingual Education.
- Huempfer, L. (2004). Can one size fit all? The imperfect assumptions of parallel achievement tests for bilingual students. *Bilingual Research Journal*, 28(3), 379–399.
- Johnstone, J., Bottsford-Miller, N., and Thompson, S. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (NCEO Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Kenney, P.A. (2000). Families of items in the NAEP mathematics assessment. In N.S. Raju, J.W. Pellegrino, M.W. Bertenthal, K.J. Mitchell, and L.R. Jones (Eds.), *Grading the nation's report card: research from the evaluation of NAEP* (pp. 5–42). Washington, DC: National Academy Press.
- Kieffer, M.J., Lesaux, N.K., Rivera, M., and Francis, D.J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79(3), 1168-1201.
- Kolen, M.J., and Brennan, R.L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.
- Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.
- Kopriva, R. (2001). *ELL validity research designs for state academic assessments: an outline of five research designs evaluating the validity of large-scale assessments for English language learners and other test takers*. Paper presented at the CCSSO Annual Conference on Large Scale Assessment, Houston, TX.
- LaCelle-Peterson, M.W., and 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.
- Liu, K., Anderson, M., Swierzbis, B., and Thurlow, M. (1999). *Bilingual accommodations for limited English proficient students on statewide reading tests* (NCEO State Assessment Series, Minnesota Report 20). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

- MacCallum, R.C., Browne, M.S., and Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149.
- Marascuilo, L.A., and Levin, J.R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of type IV errors. *American Educational Research Journal*, *7*, 397–421.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Moher, D., Schulz, K.F., Altman, D.G., & CONSORT Group. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet*, *357*, 1191-4.
- Muraki, E., and Bock, R.D. (2003). PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales (computer software). Lincolnwood, IL: Scientific Software International, Inc.
- Muthén, L.K., and Muthén, B.O. (2006). *Mplus: Statistical analysis with latent variables* (version 4.2). Los Angeles: Muthén & Muthén.
- Myers, J., and Well, A. (2003). *Research design and statistical analysis*. (2nd Ed.). Mahwah, NJ: Erlbaum.
- National Research Council (2002). Reporting test results for students with disabilities and English-language learners, summary of a workshop. In J. Koenig (Ed.), *Reporting test results for students with disabilities and English language learners*. Washington, DC: National Academies.
- National Research Council. (2004). Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments. Committee on participation of English language learners and students with disabilities in NAEP and other large-scale assessments. In J. Koenig and L. Bachman (Eds.), *Keeping score for all: the effects of inclusion and accommodation policies on large-scale educational assessments*. Washington, DC: National Academies.
- Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, *41*(3), 385–97.
- Paulsen, C.A., and Levine, R. (1999). *The applicability of the cognitive laboratory method to the development of achievement test items*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- Rabinowitz, S., Ananda, S., and Bell, A. (2004). *Strategies to access the core academic knowledge of English language learners*. San Francisco: WestEd.

- Rabinowitz, S., and Sato, E. (2005). *The technical adequacy of assessments for alternate student populations*. San Francisco: WestEd.
- Reckase, M. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics*, 4, 207–230.
- Rivera, C., and Collum, E. (2004). *An analysis of state assessment policies addressing the accommodation of English language learners*. Issue paper commissioned for the National Assessment Governing Board Conference on Increasing the Participation of SD and LEP Students in NAEP. Arlington, VA: George Washington University.
- Rivera, C., and Stansfield, C.W. (2001). *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.
- Sable, J. (2009). Documentation to the NCES Common Core of Data Public Elementary/Secondary School Universe Survey: School Year 2006–07 (NCES 2009-302 rev). U.S. Department of Education. Washington, DC: National Center for Education Statistics <http://nces.ed.gov/ccd/pubschuniv.asp>
- Sato, E. (2007). *Rethinking alignment for English language learner assessments and standards: issues and implications for extending current models*. Paper commissioned by the Council of Chief State School Officers. Washington, DC: Council of Chief State School Officers.
- Sato, E. (2008). *Linguistic modification: Part II—A guide to linguistic modification: increasing English language learner access to academic content*. Washington, DC: LEP Partnership.
- Sato, E., Lagunoff, R., Worth, P., Bailey, A.L., and Butler, F.A. (2005). *ELD standards linkage and test alignment under Title III: A pilot study of the CELDT and the California ELD and content standards*. Final report to the California Department of Education. San Francisco: WestEd.
- Sato, E., Moughamian, A., Lagunoff, R., Rayyes, N., Rivera, M., and Francis, D. (in press). Access and opportunity to learn for English language learner students and students with disabilities. San Francisco, CA: WestEd.
- Sato, E., Rabinowitz, S., and Gallagher, C. (in press). *Access and special student populations—the similarities/differences in the needs of English language learners and students with disabilities: implications for standards, assessment, and instruction* [working title]. San Francisco: WestEd.
- Sato, E., and Worth, P. (in press). *Academic language: a taxonomy for supporting English language development in the academic context* [working title]. Paper commissioned by the

- Council of Chief State School Officers. Washington, DC: Council of Chief State School Officers.
- Scarcella, R., and Zimmerman, C. (1998). Academic words and gender: ESL student performance on a test of academic lexicon. *Studies in Second Language Acquisition*, 20, 27–49.
- Schleppegrell, M.J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, 12(4), 431–459.
- Schleppegrell, M.J. (2006). The challenges of academic language in school subjects. In I. Lindberg and K. Sandwall (Eds.), *Språket och kunskapen: att lära på sitt andraspråk i skola och högskola* (pp. 47–69). Göteborg, Sweden: Göteborgs universitet institutet för svenska som andraspråk. Retrieved April 14, 2008, from <http://www.soe.umich.edu/events/als/downloads/schleppegrellp.html>.
- Scribner, A.P. (2002). Best assessment and intervention practices with second language learners. In A. Thomas and J. Grimes (Eds.), *Best practices in school psychology IV*. Washington, DC: National Association of School Psychologists.
- Shaftel, J., Yang, X., Glasnapp, D., and Poggio, J. (2005). Improving assessment validity for students with disabilities in large-scale assessment programs. *Educational Assessment*, 10(4), 357–375.
- Sireci, S.G., Li, S., and Scarpati, S. (2002). *The effects of test accommodations on test performance: a review of the literature* (CEA Research Report 485). Amherst, MA: University of Massachusetts, School of Education.
- Snow, C.E., Cancini, H., Gonzalez, P., and Shriberg, E. (1989). Giving formal definitions: an oral language correlate of school literacy. In D. Bloome (Ed.), *Classrooms and literacy*. Norwood, NJ: Ablex.
- Solano-Flores, G., and Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, 25(1), 13–22.
- Solano-Flores, G., and Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English language learners. *Educational Researcher*, 32(2), 3–13.
- Stansfield, C. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, 20(2), 189–207.
- Tabachnick, B. C., and Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education, Inc.

- Thissen, D., and Wainer, H. (Eds.). (2001). *Test scoring*. Hillsdale, NJ: Erlbaum.
- Thompson, S.J., Johnstone, C.J., and Thurlow, M.L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M., and Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (NCEO Synthesis Report 41). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M.L., McGrew, K.S., Tindal, G., Thompson, S.J., Ysseldyke, J.E., and Elliot, J.L. (2000). *Assessment accommodations research: considerations for design and analysis* (NCEO Technical Report 26). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M.L., Wiley, H.I, and Bielski, J. (2002). *Biennial performance reports: 2000–2001 state assessment data*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Tindal, G., and Ketterlin-Geller, L. (2004). *Research on mathematics test accommodations relevant to NAEP testing*. Washington, DC: National Assessment Governing Board.
- U.S. Department of Education, Office of Elementary and Secondary Education, Standards and Assessment Group and Accountability Group. (1999). *Peer review guidance for evaluating evidence of final assessments under Title I of the Elementary and Secondary Education Act*. Washington, DC: United States Department of Education.
- U.S. Department of Education, Office of Elementary and Secondary Education. (2000). *The provision of an equal educational opportunity to limited English proficient students*. Washington, DC: U.S. Department of Education.
- van Someren, M., Barnard, Y., and Sandberg, J. (1994). *The think-aloud method: A practical guide to modeling cognitive processes*. San Diego, CA: Academic Press.
- Willis, G. (1999). *Cognitive interviewing: A “how to” guide*. Paper presented at the 1999 meeting of the American Statistical Association, Raleigh, NC.
- Wu, M.L., Adams, R., Wilson, M.R., and Haldane, S.A. (2007). *ACER ConQuest: generalized item response modeling software* (version 2). Camberwell, Australia: ACER Press, an imprint of Australian Council for Educational Research Ltd.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zwick, R., Thayer, D.T., and Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36 (1), 1–28.

Appendix A. Power analysis for primary research questions³³

Although the target number of students in this study was 3,600, 4,617 students who met eligibility requirements were recruited.³⁴ The sample included 1,214 English language learner students and 3,403 non-English language learner students. Each item set (original or linguistically modified) was administered to approximately 2,300 students (600 English language learner students and 1,700 non-English language learner students; table A1).

Table A1. Full study design sample

Sample group	Original item set	Linguistically modified item set
EL students	606	608
NEP students	821	804
EP Students	883	895

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

Source: Authors' analyses based on primary data.

The minimum detectable effect size for the main research question, which asks whether the score differences between the original and linguistically modified item sets differ across the three subgroups (EL, NEP, and EP students), was calculated using the following formula:

$$(1) \quad MDES = \sqrt{\frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 (\sum C_i^2) \sigma^2}{n_i}}$$

where *MDES* is the minimum detectable effect size associated with an α level of 0.05 and power of 0.80 ($1 - \beta$), c_i is the contrast coefficient for cell i , σ^2 is the common population (within-cell) variance (assumed to be 1 for this computation), and n_i is the harmonic mean of the sample sizes in the 6 cells .

$$0.25 = \sqrt{\frac{7.846 * 6}{749}}$$

Thus, the study was powered to detect a difference in modification impacts (difference-in-the-difference) of 0.25 standard deviation units across the three subgroups.

³³ For item response theory (IRT) models, researchers generally take into account sample size in determining which IRT model would be appropriate. For the one-parameter logistic (1-PL) model, it is recommended to have at least 200–500 cases; for the two-parameter logistic (2-PL) model, it is recommended to have at least 800–1,000 cases; and for the three-parameter logistic (3-PL) model, it is recommended to have at least 1,200–1,500 cases (Crocker & Algina 1986; Embretson & Reise 2000; Reckase 1979; Hambleton & Jones, 1993; Harris 1989; Thissen & Wainer 2001).

³⁴ As reported in chapters 1 and 2.

Appendix B. Operational test administration manual

Tests may be administered any time between April 28 and July 31, 2008.

This manual contains directions for administering operational tests for the REL-W study, *Assessment Accommodations for English Language Learners*. Please read this manual carefully.

Information about this study

This study is designed to examine the effects of linguistic modification on the validity of assessments for English language learners (ELs). Specifically, it systematically investigates the ways in which linguistic modification, used as a test accommodation, affects students' ability to access mathematics content during testing. This study aims to increase understanding of the effects of a test accommodation that could lead to promising solutions for decreasing the achievement gap between EL and non-EL students. Findings from this study may advance current understanding of technically sound assessment practices by presenting empirical evidence about the ways in which increasing ELL access to tested content may yield more valid measures of what students know and can do.

About the test

As described above, two different versions of a mathematics test, one with original items (Version O) and one with linguistically modified items (Version M), have been developed. One version will be randomly assigned to each student, and each student will be administered only one test. All test items are multiple-choice, with four possible response options. Students will use the same test booklet during the entire test. They will be asked to circle their responses in the test booklet and encouraged to use the white space in the booklet to complete their computations. Tests will be administered in paper-pencil format during school hours to intact math classes. Total testing time is estimated to be approximately 50 minutes. *Calculator use is allowed.*

Students eligible for the test

The test will be administered to students in grades 7 or 8 who are (1) general education students or (2) Spanish-speaking English language learners (ELs). Students with an IEP will be included at the school's discretion, as the only testing accommodations provided will be extra time. This test administration will *not* include students who require modified testing materials such as Braille, large-print, online, or audiotape. However, when possible, all students in a classroom should be tested, and tests will be scored only for the target population.

Informed consent and incentive to participate

Parents of eligible students will receive information about the study (available in Spanish and English). As we have been approved for a passive consent process, if parents/guardians do not want their students to participate, they will be instructed to contact us at REL-W; otherwise, all students are assumed to be eligible to participate. In addition, in some districts, eligible students may be asked to sign an assent form (available in Spanish and English) prior to testing. In all cases, students will be reminded that participation is voluntary but appreciated and that they may refuse to answer any item or question. Each student will receive a pen to thank him/her for participating.

Important information about standardization

It is very important that the standardized procedures described in this manual are followed. Any deviations from standardized testing conditions that emerge during testing should be documented on the Debrief Form and reported to WestEd as soon as possible. If you have any questions concerning these instructions, please contact Edynn Sato at esato@wested.org.

Responsibilities of test administrators

As a test administrator or proctor, you are responsible for

- Reading this manual before the administration of the test to familiarize yourself with its contents.
- Reviewing the scripts that you will read to students during testing.
- Reviewing both forms of the test to familiarize yourself with the format and contents.
- Ensuring that your sets of test materials are ready for administration.
- Administering the test according to the instructions in this manual, including use of scripted language.
- Answering questions that may emerge during testing (see “General protocol for answering questions” below).
- Protecting the security of the tests by carefully following the post-administration guidelines in the manual.
- Returning all test materials to WestEd at the end of the session.

In addition to the responsibilities described above, as a bilingual test administrator or proctor, you also are responsible for

- Reading directions to students in Spanish (scripted language is provided below).
- Attending to the anxiety levels of ELs during testing.

General protocol for answering questions

Because the focus of this study is on linguistic modification of test items, test administrators are required to follow the guidelines below when responding to questions that students ask either before or during the test administration.

- Administrators should feel free to answer questions pertaining to the general administration of the test (e.g., marking instructions for indicating the correct answer, remaining time for testing).
- Administrators should not answer questions pertaining to the content of an item, the specific language or terminology used in an item, or steps for solving an item. Instead, administrators may respond to these types of questions by reminding students that they may skip items that they do not understand. Administrators are encouraged, however, to keep track of questions that come up by writing them down on the provided Debrief Form and submitting this to WestEd.

Test materials

Prior to the scheduled test administration time, please check to make sure you have the following materials:

- Test administration manual (this document).
- Sample test booklet.
- Student test booklets.
- Pencils.
- Debrief Form.
- Pens to hand out to students following testing to thank them for participating.

The ordering of components in both test booklets is as follows:

- Test directions page (English and Spanish versions).
- Multiple-choice items ordered 1–25 (English only).
- Language background survey (English and Spanish versions).

General Instructions for administering the test

Enter the classroom with all test materials described above. Students should be directed to circle their answers directly in the test booklet. Be sure to have a sample test booklet available so you can point out key sections to students when necessary.

Prior to testing,

- If students are accustomed to using calculators when solving math problems in class, be sure they have these available for this test, as calculator use is allowed.
- Hand out a pencil to each student.
- Distribute one test booklet to each student. Because the two versions of the test booklets have been randomly mixed prior to testing, the tests should be distributed to students exactly in the order in which they are arranged.
- Say to the students:

This test is part of a research project that we are working on so we can better understand how you read and solve math problems. Your answers to these items will help us create tests that allow you to show what you know and can do in ways that are fair for all students.

Esta prueba es parte de un estudio educativo en el que estamos trabajando para entender mejor cómo es que los estudiantes leen, interpretan y resuelven problemas de matemáticas. Sus respuestas a estos ejercicios de matemáticas nos ayudarán a crear exámenes que les permitirán demostrar lo que saben y lo que hacen para resolver los problemas, de una manera más justa para todos los estudiantes.

You will not be graded on this test. Your answers will not affect your grade in this class. You will have about 45 minutes to complete the items and questions in this test booklet.

Uds. no serán calificados o no se les darán notas con esta pruebas. Sus respuestas no afectarán sus calificaciones en esta clase. Tendrán 45 minutos para completar este libretto de pruebas y la encuesta final.

Please print your name at the bottom of the outer cover page where it says *Student Name*. This page will be removed and destroyed once all data have been collected.

Por favor escriban su nombre en la parte de abajo de la primera pagina, en la cubierta, donde dice “Student Name—Nombre del Estudiante”. Esta primera página será destruída una vez los datos se hayan colectado.

Please open your test booklet to Pages 2 and 3. A message to students is presented in English and in Spanish. Please read this message—in either language—to yourself now.

(Por favor) Abran sus libretos en las páginas 2 y 3. Hay un mensaje para Uds. La página 2 lo tiene en Inglés y la página 3 está en Español. Por favor léanlo silenciosamente ahorita.

(Watch until all appear to have read message.)

On the next pages, you will find 25 math items. As you solve each problem, please circle the letter of your response to each question directly in your test booklet. Please show your work in the area around each item in the test booklet. Remember to complete all work on your own as we are very interested in how YOU would go about solving these problems.

En las siguientes páginas encontrarán 25 problemas de matemáticas. Por favor hagan un círculo alrededor de la letra que creen es la respuesta para cada pregunta del libreto. Pueden escribir cómo resuelven el problema en el área en blanco de cada ejercicio del libreto.

If you do not know how to solve a problem, you may move on to the next question and then come back to that problem later. I will let you know when there are 20 minutes left for testing.

Do you have any questions about these directions?

Si no saben cómo resolver un problema, pueden seguir a las siguientes preguntas y volver a revisarlo más tarde. Les diré cuando queden 20 mins para organizar su tiempo. ¿Tienen alguna pregunta con respecto a estas direcciones?

(Using the guidelines specified above, respond to any questions the students might have.)

As you work, please raise your hand if you have a question about any test item or survey question. Now open your booklet to page 4. You may begin. (*NOTE start time.*)

A medida que trabajan, si tienen alguna pregunta con respecto a algún ejercicio o pregunta de encuesta, levánten la mano. AHORA, pueden abrir el libreto en la página 4. PUEDEN EMPEZAR.

- (*NOTE start time.*)

- Circulate and check students' work from time to time during the session. When 25 minutes have passed, say to the students:

You have about 20 minutes left for testing. When you have answered all of the items, don't forget to answer the five questions at the end. When you finish, you may go back and check your work, then close your test booklet and sit or read quietly.

Tienen 20 minutos para terminar la prueba. Cuando terminen con todos los ejercicios de la prueba y las preguntas de la encuesta, pueden volver a revisar sus respuestas. Luego cierren el libreto y quédense sentados leyendo silenciosamente.

- When another 20 minutes have passed, say to the students:

The testing time will end soon. If you have not completed the test, you may continue working. If you have completed your test, please raise your hand and I will collect your test booklet.

El período ya casi termina. Si no han completado la prueba, pueden continuar trabajando. Si ya terminaron y completaron el examen, por favor levánten la mano para poder recoger el libreto.

- Collect the test booklets from those students who have completed the test. Confirm that each student's name is printed on the front cover of the test booklet and that the student completed the survey questions at the end. Students who have not yet finished should be allowed to continue as long as they are working productively. Ensure that students who have finished are quietly occupied so they will not disturb students who are still testing.
- Once all tests have been collected, verify that the number of booklets is consistent with the number of tested students.
- At the end of the test, when all students (or nearly all, if the dismissal bell will ring soon) have completed the test, give one pen to each student. Then say to the students:

Thank you very much for your help today. We appreciate your willingness to support this important research study. Have a great day!

Muchas gracias por ayudarnos hoy. Apreciamos mucho el que hayan participado para ayudar en este estudio tan importante. ¡Que tengan un buen día!

- Collect all test materials and return them to WestEd.
- Complete the Debrief Form.

THANK YOU VERY MUCH FOR YOUR TIME AND ASSISTANCE.

Checklist for test administrators

Before testing

- Read this instruction manual in its entirety.
- Become familiar with the student test booklet.
- Check to make sure you have enough of all materials needed for testing (e.g., test booklets, pencils, and pens to give out at the end).
- Organize materials to expedite handing out of test booklets to students once class starts.

During testing

- Follow the provided script for instructions.
- Be sure that all students have comfortable and adequate workspaces.
- Complete the Debrief Form for every class taking the test.
- Maintain test security at all times.
- Monitor students' progress throughout the class period and answer questions as they arise.
- As you collect finished tests, ensure that students answered the questions at the end (language background survey).

After testing

- Verify that a student name appears on the cover of each used test booklet.
- Ensure all test materials are collected from the classroom.
- Thank the students.
- Return all test booklets (used and unused) and completed Debrief Forms to WestEd (see separate instructions below for returning materials).

Instructions for the return of test materials to WestEd

What MUST be returned to WestEd:

- All used test booklets
- All unused test booklets
- All completed Debrief Forms (complete 1 per class)
- All used scratch paper (if any was used)

The return of extra pens and pencils is appreciated, but not necessary.

STEP 1: Following the administration of the test, please place all test materials in the above bulleted list into the same box in which you received the shipment.

STEP 2: Securely close the box using packaging tape.

STEP 3: Affix the return FedEx shipping label that was provided with this shipment on the outside of the box.

STEP 4: Drop off the box at any location that accepts FedEx shipments, use a regularly scheduled FedEx Pickup at your school, or call 888.777.6040 to schedule a pick up with FedEx.

If you encounter any difficulties or have questions about the return of materials, please contact Carole Gallagher at 415.615.3211 or Carol Whang at 415.615.3346.

Appendix C. Student Language Background Survey

This appendix contains the Student Background Survey in English and in Spanish.

English version

We estimate that it will take you about 5 minutes to complete this survey. Remember, there are no right or wrong answers, and you will not be graded on this task. Your participation is voluntary and you may refuse to answer any question. Thank you for your time!

Please place a CHECK ✓ in the box that applies.

1. I am in 7th Grade. or I am in 8th Grade.

2. I am male. or I am female.

Please place a CHECK ✓ in all boxes that apply.

3. I attended these grades in the United States:

Kindergarten 1st Grade 2nd Grade

3rd Grade 4th Grade 5th Grade

6th Grade 7th Grade 8th Grade

4. Did you ever go to school in another country? Yes or No

IF YES, please write the name of that country on this line: _____.

IF YES, please CHECK ✓ all of the grades you attended in that country:

Kindergarten 1st Grade 2nd Grade

3rd Grade 4th Grade 5th Grade

6th Grade 7th Grade 8th Grade

5. We speak these languages in my home:

English Spanish

Other (please write the name of that language) _____.

Your answers to these questions about your language background will be used as part of a research study about testing accommodations sponsored by the U.S. Department of Education and carried out by the Regional Educational Laboratory West at WestEd. If you have questions about the study or this survey, please contact Edynn Sato at (415) 615-3226 or at esato@wested.org.

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless such collection displays a valid Office of Management and Budget (OMB) control number. The valid OMB control number for this information collection is 1850-0849. The time required to complete this information collection is estimated to average

5 minutes per response, including the time to review instructions, search existing data resources, gather the data needed, and complete and review the information collection. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Education, Washington, D.C. 20202-4651. If you have comments or concerns regarding the status of your individual submission of this form, write directly to: Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington, D.C. 20208.

In accordance with The Education Sciences Reform Act of 2002, Title I, Part E, Section 183, responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

Versión en Español

Anticipamos que tomarán 5 minutos para completar esta encuesta. No hay respuestas correctas o incorrectas y no se darán calificaciones por ello. Su participación es voluntaria y puede decidir no contestar cualquier pregunta.

Marque con un \surd lo que le corresponde a Ud.

1. **Estoy en el Grado 7.** o **Estoy en el Grado 8.**

2. **Yo soy un hombre.** o **Yo soy una mujer.**

En las siguientes preguntas, marque con un \surd todas los grados de clase que le corresponden:

3. **Los grados de clase que he atendido en los Estados Unidos son:**

- Kindergarten 1^{er} Grado 2^o Grado
 3^{er} Grado 4^o Grado 5^o Grado
 6^o Grado 7^o Grado 8^o Grado

4. Los grados de clase que he atendido en otro país (o países) son:

- Kindergarten 1^{er} Grado 2^o Grado
 3^{er} Grado 4^o Grado 5^o Grado
 6^o Grado 7^o Grado 8^o Grado

(Nombre del país o países)

5. Marque con un \checkmark los idiomas que se hablan en su casa:

- Inglés Español
 Otro(s) idioma(s): _____.

Tus respuestas a estas preguntas acerca de tu idioma serán parte de un estudio que revisara adaptaciones patrocinados por el Ministerio de Educación de los Estados Unidos y conducidos por el Laboratorio de Educación Regional del Oeste (REL West) en WestEd. Si tienes alguna pregunta acerca del estudio o de esta encuesta, por favor comunícate con Edynn Sato al número telefónico (415) 615-3226 o por correo electrónico a esato@wested.org.

Ninguna persona es requerida a contestar a la colección de información sin tener un número válido de control de OMD de acuerdo al Acto Paperwork Reduction del 1995. El número válido de control de OMD para coleccionar esta información es 1850-0849. El tiempo requerido para terminar esta colección de información se estima que es un promedio de 5 minutos. Esto incluye el tiempo para repasar las instrucciones, buscar datos existentes, obtener los datos necesarios, y terminar y repasar la información que se ha coleccionado. Si tienes comentarios sobre la exactitud del tiempo que se estima para terminar la forma o sugerencias para mejorar esta forma, por favor escriba al: Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington D.C. 20208.

De acuerdo al Acto de Education Sciences Reform del 2002, Título I, Parte E, Sec. 183, respuestas de esta colección de datos serán utilizados solamente para el propósito de estadísticas. Los reportes preparados para este estudio resumirán las conclusiones de la muestra y no serán asociadas con un distrito o individuo. No daremos información que lo identifique a usted o su distrito a cualquier persona que no sea parte del equipo del estudio, a menos que sea obligado por la ley.

Appendix D. Guide for developing a linguistically modified assessment

[This guide was followed to linguistically modify the items used in this study. Experts in mathematics, linguistics, measurement, curriculum and instruction, and the English language learner student population were convened to discuss linguistic modification strategies and their application. These experts possessed advanced degrees (such as an M.A. or Ph.D.), had classroom teaching experience, and assessment development experience. The selection of items, the linguistic modification of items, and the creation of the item sets used in this study occurred over the equivalent of a period of approximately three weeks and followed generally accepted item development procedures including verification of content alignment, appropriateness for the student population, and freedom from bias and sensitivity issues.]

For all students, access to test content is necessary to ensure the validity of assessment results.³⁵ Valid assessments are especially critical if results are used to inform classroom instruction or for accountability purposes. When access is constrained in some way (for example, linguistically or cognitively), students may be prevented from fully demonstrating what they know and can do, and the test score may underestimate or misrepresent students' achievement. To assess English language learner students' knowledge of academic content, it is critical to determine whether their academic performance reflects their understanding of the targeted content or their lack of English language proficiency. There is an interaction between how assessed content is presented in test items and what English language learner students need in order to access that content. This interaction affects the validity of the assessment results and the interpretation of those results.

Linguistic modification of test items is an approach for addressing the particular access needs of English language learner students so that test performance is attributable less to English language proficiency and more to knowledge and skills related to the tested content. The approach outlined below is intended to help researchers in this study consider key characteristics of the content and the student population as they develop linguistically modified test items. The three steps in this process are:

- Define the domain and constructs of tested content.
- Define the English language learner population that will be tested.
- Apply and evaluate linguistic modification strategies to test items.

³⁵ Information in this appendix is drawn from Sato (2008).

Step 1: define the domain and constructs

Articulate the purpose of the assessment. Consider the range of ways the assessment results will be used and the intended outcomes of testing.

Recommended specialists for this step

Given the purpose of the assessment and the population assessed, this step is best conducted by a team that includes content specialists, assessment specialists, curriculum and instruction specialists, English language development specialists, and population specialists (that is, individuals with specialized knowledge about the English language learner student population).

Purpose

The assessment results will be used for the following purpose(s):

Assessed academic content domain

The assessment will measure students' knowledge of:

Considerations

Is this test appropriate for the target content domain? To what degree do content domain characteristics align with the intended purpose of this assessment?

Assessed constructs—content and skills

More specifically, the assessment will measure the following constructs (content and skills) related to the domain:

Considerations

Do the content and skills assessed in the set of linguistically modified test items reflect the intended breadth, depth, and range of complexity of the assessed domain? Are the verbs used in the state standards statements specific enough to guide assessment development (for example, “identify,” “describe,” “compare” vs. the more vague “know,” “understand”)? If the latter, how are students expected to demonstrate their knowledge and skills?

Content-related language—language demands

The following language demands are associated with the content and skills that will be assessed (see tables E1 and E2 in appendix E for a list of language demands—linguistic skills and academic language functions):

Considerations

Have students' linguistic skills and academic language functions both been considered?
Is the range of language demands in the linguistically modified items consistent with the breadth, depth, and range of complexity of the assessed content domain?

Content-related language—specific vocabulary and terminology

The following vocabulary and terminology are specific to the grade-level content assessed; therefore, they should not be linguistically modified:

Considerations

Is the vocabulary and terminology identified consistent with the intent of the grade-level content standards?

Step 2: define the population and student subgroups

Articulate the key characteristics and access needs of the English language learner student population. Since this group of students is especially diverse and heterogeneous, it may be necessary to identify key subgroups of students within the state.

Recommended specialists for this step

Given the purpose of the assessment and the population assessed, this step is best conducted by a team that includes content specialists, assessment specialists, curriculum and instruction specialists, English language development specialists, and population specialists (that is, individuals with specialized knowledge about English language learner students).

Student population

The target English language learner population can be characterized as follows (see appendix E for a description of English language learner students):

Student access needs

Document the access needs of the target English language learner student population, taking into account characteristics such as:

Context

What topics, themes, locations, situations, illustrations, and such are familiar to these students?

Words, phrases, sentences

What written vocabulary is familiar to these students? What phrases are familiar to these students? What sentence structures are familiar to these students? What tenses (for example, present, past) and constructions (for example, plural *_s*, possessive *'s*) are familiar to these students? What proper nouns are familiar to students as a result of their classroom reading?

Format/Style

With what formats/styles are these students familiar (for example, bulleted lists, text boxes, underlining for emphasis)? How is information typically presented to these students during instruction?

Step 3: apply and evaluate linguistic modification strategies

Determine which content and item types lend themselves to linguistic modification. Then develop and evaluate each test item according to the following dimensions: context, graphics, vocabulary/wording, sentence structure, and format/style (see table D1 for linguistic modification guidelines and strategies for each dimension).

Recommended specialists for this step

This step is best conducted by a team that includes content specialists, assessment specialists, curriculum and instruction specialists, English language development specialists, and population specialists (that is, individuals with specialized knowledge of the English language learner population).

Categorize target content and item types

Sort content/test items into one of the following three categories of eligibility for linguistic modification. Within each eligibility category, group content standards and test items by content strand (for example, measurement or algebra for mathematics).

- Definitely eligible.
- Definitely not eligible.
- Possibly eligible.

Considerations

A test item's appropriateness for linguistic modification is associated with the quantity of construct-irrelevant language in that test item; the greater the quantity of construct-irrelevant language, the greater the likelihood that the item can be linguistically modified effectively for English language learner students. There also is a greater likelihood that construct-irrelevant language can be linguistically modified without significantly changing the assessed construct (for example, mathematics achievement).

Apply linguistic modification guidelines and strategies

For content/items that are eligible and possibly eligible for linguistic modification, systematically apply the relevant guidelines and strategies presented in table D1 (that is, context, graphics, vocabulary/wording, sentence structure, format/style).

Considerations

The team of specialists who are linguistically modifying items need specialized training to ensure that they are appropriately applying linguistic modification guidelines. It is important to ensure the guidelines are accurately and consistently applied during item development and that the intended construct, cognitive complexity, and language demands specified in the grade-level standards have not been significantly altered.

Follow checklist for evaluating the linguistically modified items

For each item, verify that:

- The construct being tested has not changed.
- The cognitive complexity of the item is appropriate.
- The following elements in the linguistically modified item maximize English language learner students' linguistic access:
 - Context.

- Graphics.
- Vocabulary/wording.
- Sentence structure.
- Format/style.

Methods used to verify that the test item has been appropriately linguistically modified include:

- Expert verification (for example, by a technical advisory committee, content and bias review committee, or independent external reviewer) that the construct has not changed and that the cognitive complexity of the item is appropriate.
- Statistical analyses (for example, analysis of variance, differential item functioning analysis, or factor analysis).
- Cognitive interviews.

Table D1. Linguistic modification guidelines and strategies

Desirable characteristics	Notes on approaches and criteria
<i>Item context</i>	
<ul style="list-style-type: none"> • Familiar to students. • No cultural or linguistic bias. • Minimal construct (no irrelevant words or phrases). 	<ul style="list-style-type: none"> • The context situates the problem (and may include description of relationship or interaction between location and time). • In the body of the report, context is often described in relation to its complexity and as part of biased or construct-irrelevant information that should be pruned out. Recommendations: <ul style="list-style-type: none"> ○ Remove passive voice construction in original item. ○ Remove past tense and conditional in original item. ○ Break stem into shorter, less complex sentences (sometimes a series of shorter sentences can create a story line or present a more familiar context/situation to students). • Context can provide description that helps make abstract or highly generalized situations more concrete and relevant. Simply stated, it helps to ground the content being tested. Context that facilitates access for English language learner students is expressed in concrete language, illustrative language, and illustrations/graphics.

Desirable characteristics	Notes on approaches and criteria
<i>Item graphics</i>	
<ul style="list-style-type: none"> • Familiar to students. • No cultural or linguistic bias. • Symbols, legends, and key vocabulary relevant to the construct and familiar to English language learner students. • Consistent graphic and labeling/naming conventions • Supportive of English language learner student understanding of assessed content. 	<ul style="list-style-type: none"> • Graphics include diagrams, tables, charts, drawings, graphs, pictures, and maps. • Student knowledge about certain graphics is required and assessed in mathematics. • Graphics allow for reduced amount or complexity of language in a test item. Use of graphics in test items should serve a clear purpose. Otherwise they may be misleading or distracting. For example, graphics may be used to: <ul style="list-style-type: none"> ○ Clarify key aspects of the content/construct assessed. ○ Clarify construct-relevant context. ○ Clarify a mathematical operation. ○ Indicate what the student is expected to do. ○ Help students shift from one context to another within an assessment (for example, from one type of test item to another). ○ Allow students to reinforce or verify understanding of key information in test item. ○ Simplify the structure of a test item that requires a number of operations or steps (for example, through bulleted lists or a diagram of the complete problem that accurately reflects the problem in its totality). • Some criteria that can be used to evaluate the need for a graphic include: <ul style="list-style-type: none"> ○ Does the graphic clarify construct-irrelevant information? If so, it may not be necessary. It might be better to revise or delete the construct-irrelevant information. ○ Does the graphic support the test item context without requiring additional written text? ○ Does the graphic accurately represent the full complexity of the problem? If not, it may be misleading. ○ Is the graphic consistent with the key content/construct of the item?

Desirable characteristics	Notes on approaches and criteria
<i>Item vocabulary/wording</i>	
<ul style="list-style-type: none"> • High-frequency words. • Common and familiar words. • Relevant technical terms that reflect language of the content standards and academic English language. • Technical terms defined, as appropriate. • Naming conventions consistent with graphics/stimuli. • Construct-irrelevant vocabulary/phrases at or below grade level. 	<ul style="list-style-type: none"> • Careful selection of vocabulary and phrases can simplify sentence structure. The amount and complexity of language should be balanced with the amount of information necessary for student to understand/access the item. The goal is to make the language as clear and straightforward as possible, while still providing the amount and complexity of information necessary to communicate the targeted content of the test item. • Some general guidelines: <ul style="list-style-type: none"> ○ Use precise language. Appropriate language modification does not simply mean using common or familiar vocabulary. ○ Consider language used in the content standards and academic English language . ○ Repeat key words/phrases in the test item that students need to understand the item and respond to it. ○ Do not automatically provide synonyms for a key word. This may not be helpful, especially if a test item is already long or complex. Although providing synonyms may be helpful during instruction, it may not be useful in assessment items. ○ Use words/phrases consistently within the context of the item and consider consistency of terms within a strand—for example, reading or measurement). Support this use with context-familiar content-based abbreviations and make explicit connections between terms/abbreviations. • If possible, avoid using: <ul style="list-style-type: none"> ○ Ambiguous words or unnecessary words with multiple meanings. ○ Irregularly spelled words. ○ Proper nouns that are irrelevant or not meaningful to the population. ○ Words that are both nouns and verbs (for example, carpet, value, cost); however, if a choice needs to be made, use the word only as a noun. ○ Hyphenated and compound words ○ Gerunds. ○ Relative pronouns (for example, which, who, that) without a clear antecedent.

Desirable characteristics	Notes on approaches and criteria
<i>Item sentence structure</i>	
<ul style="list-style-type: none"> • Familiar, common sentence structure. • Complexity of sentence structure at or below grade level. • Key information presented first or early in the test item. • One sentence per idea for complex test items. 	<ul style="list-style-type: none"> • To reduce the complexity of a sentence in a test item: <ul style="list-style-type: none"> ○ Identify the agent (that is, the person or object carrying out the action) to construct sentences that use active voice (and avoid passive voice). ○ Make sure that the verb in a sentence follows the subject as closely as possible. ○ Remove introductory phrases that are irrelevant to the construct being tested. ○ Use conventional constructions (for example, apostrophes for possessives and “s” or “es” for plurals). ○ Use proper nouns that students are familiar and are grade-level appropriate. ○ Use clear grammatical structures. • To reduce language load: <ul style="list-style-type: none"> ○ Change past or future tense verb forms to present tense. ○ Change passive verb forms to active verb forms. ○ Change complex sentence structure to subject-verb-object structure. ○ Shorten any long nominals/names/phrases (for example, “last year’s class vice-president” to “a student leader”). ○ Replace compound sentences with two separate sentences, especially when making comparisons. ○ Shorten or delete long prepositional phrases. ○ Replace conditional clauses with separate sentences. ○ Change the order of a clause within a sentence. ○ Remove or rephrase relative clauses. ○ Rephrase questions framed in negative terms. • Make sure the following are clear. <ul style="list-style-type: none"> ○ Noun-pronoun relationships. ○ Antecedent references.

Desirable characteristics	Notes on approaches and criteria
<i>Item format/style</i>	
<ul style="list-style-type: none"> • Clear parts of the item/question. • Explicit order of operations. • Relevant and appropriate distinctions. • Segmented or shortened long problem statements. 	<ul style="list-style-type: none"> • Place test item elements in the following order: (1) text that introduces the graphic; (2) graphic; and (3) the test item stem. • Format for emphasis of key words/terms (highly construct-relevant), using bold, ALL CAPS, and <u>underline</u> to call English language learner students' attention to them. • Consider whether blocks of text (that is, a paragraph) may be necessary and appropriate for presenting a test item. This depends on the construct assessed, the complexity of the information needed by the student to respond to the item, and the centrality of the context to the construct. Suggested strategies to help English language learner students process such text include: <ul style="list-style-type: none"> ○ Bulleted lists. ○ Indenting key information. ○ Emphasizing key words/terms. ○ Using graphics.

Source: Sato 2008.

Appendix E. Workgroup training materials

This appendix provides information on language demands on students, particularly English language learner students, and definitions of key terms used in the discussion of linguistically modified assessments for training item developers.³⁶

Language demands

This section presents an overview of the domain of language needed by students, particularly English language learner students, in the context of academic learning. Mastery of academic language allows students to access and meaningfully engage with academic content. The operationalization of language demands is intended to facilitate systematic analysis of the specific linguistic skills and language functions embedded in standards, curriculum and instructional materials, and assessments.

Categorization

A language demand is categorized as either a linguistic skill or an academic language function based on whether the demand is fundamental to developing and using language or is a contextual application of language.

Modalities

Linguistic skills are labeled to indicate the language modalities (listening, speaking, reading, and writing) believed to be associated with each language demand (table E1). These are the receptive and productive modalities in which the language demand is experienced. Related language modalities are not listed for the academic language functions because each function can involve any or all four domains, depending on application.³⁷

Grouping

Academic language functions are grouped according to three levels, representing the minimum language skills required to perform the given language function: word level, phrase/sentence level, and discourse level (table E2). In implementation a language demand might require more language, but performance of the demand would not be possible without the minimum level listed.³⁸ Some academic language functions are grouped; this occurs when the language demands are very similar or represent multiple levels of essentially the same demand.

³⁶ Information in this appendix is drawn from Sato et al. (2005); Bailey, Butler, & Sato (2007); and Sato (2008).

³⁷ For example, identification taken alone could be a productive function (speaking and/or writing). In many common implementations, however, identification also requires receptive ability. For example, a student would identify something in a reading, listening, or viewing context.

³⁸ For example, classifying can be done at a low language level, such as sorting words by affix, or at a much higher language level, such as classifying characters from multiple literary sources according to their motives or

Key terms

This section described key terms used in the discussion of linguistically modified assessments for training item developers.

Access

To maximize student access to the content being assessed on an achievement test (for example, mathematics), text in the item that is not directly related to the targeted construct (that is, construct-irrelevant text) is minimized or removed. Doing so facilitates students' ability to demonstrate their construct-relevant knowledge and skills and reduces or eliminates sources of construct-irrelevant variance (construct irrelevance) in test results among students. In other words, when access is constrained, it can result in the measurement of sources of variance that are not related to the intended test content. If student access to tested content is restricted, students cannot fully demonstrate what they know and can do; subsequently, test results underestimate their level of content achievement (underrepresentation).

In this study the construct-irrelevant factors that constrain access to tested content for English language learner students are examined to support development of mathematics test items that maximize students' ability to show what they know and can do in mathematics.

Accommodation vs. modification

An accommodation is a change in testing conditions that is implemented to increase accessibility of test content to a specific student population. Such changes are deemed fair and reasonable when standardized administration conditions do not provide an equal opportunity for all students to demonstrate what they know and can do (Abedi & Lord 2001; Butler & Stevens 2001; Holmes & Duron 2000; National Research Council 2002, 2004). It is assumed that the same construct is being assessed with and without the accommodation. An accommodation is intended to minimize or remove the effects on test performance of construct-irrelevant factors that may contribute to, for example, the underrepresentation of student achievement in the content area.

A modification is an adjustment to the test itself, the administration conditions, or the content standards for assessment. While modification may improve access to the test content for a specific student population in a fair and reasonable manner, it significantly alters the construct being assessed. Examples of test modifications include allowing students with specific disabilities to use calculators on mathematics computation items (when general education students cannot) or allowing the reading comprehension portions of a test to be read aloud to English language learner students.

In traditional psychometric practice, accommodations may affect the performance of its intended referent group only, while remaining construct-neutral to nonaccommodated students—that is,

characteristics. However, evaluation can be done only at the discourse level. A critical reading and assignment of meaning requires minimum language beyond the word or sentence level.

the accommodation should benefit the student needing the accommodation but should have no effect on those not needing the accommodation.

However, research-based test design practices (for example, universal design, simplified language in items and associated text) suggest that all student groups may benefit from item development strategies designed to minimize construct-irrelevant variance. So, for this study an accommodation may be considered valid, even if all groups benefit from its use, if evidence collected suggests that:

- The construct/content assessed was not significantly altered.
- The performance of the group targeted for accommodation (that is, English language learner students) improves at a greater rate than that of their English-proficient counterparts.

English language learner students

English language learner students are “national-origin-minority students³⁹ who cannot speak, read, write, or comprehend English well enough to participate meaningfully in and benefit from the schools’ regular education program” (U.S. Department of Education, Office of Elementary and Secondary Education 1999, p. 60). No Child Left Behind legislation (including Title III) refers to this population as “limited English proficient” (U.S. Department of Education, Office of Elementary and Secondary Education 2000).

This study’s analyses included only students in grades 7 and 8 who identified themselves as “Hispanic” or who identified Spanish as their first language or the language spoken in their home. Recruitment efforts targeted Spanish-speaking English language learner students who scored at the mid- to high range of English language proficiency to ensure that their command of the English language was at a level sufficient to benefit from the linguistic modification.

Linguistic modification

Linguistic modification is a theory- and research-based process in which the language in test items, directions, and response options is modified in ways that clarify and simplify the text without simplifying or significantly altering the construct assessed. To facilitate comprehension, linguistic modification reduces construct-irrelevant language demands (for example, semantic and syntactic complexity) of text through strategies such as reduced sentence length and complexity, use of common or familiar words, and use of concrete language (Abedi et al. 2005; Abedi, Lord, & Plummer 1997; Sireci, Li, & Scarpati 2002).

Linguistic modification is not simply good editing practice and does not result in simpler items. Rather, it is a linguistically based, systematic means for targeting, reducing, and removing the irrelevant variance in test performance that is attributable to individual differences in English proficiency so that English language learner students can fully demonstrate what they know and

³⁹ “National origin minority” can include students born in the United States.

can do in that content area. By minimizing the language load, a source of construct-irrelevant variance, English language learner students' access to construct-relevant content is enhanced.

Table E1. Linguistic skills

Language demand	Definition/application	Linguistic area of study
Morphemes	The ability to identify and distinguish the smallest unit of meaningful sound in language (that is, words, roots, or affixes). (L, R)	Morphology
Phonemes	The ability to identify, manipulate, and produce the individual sounds that make up spoken words. (L, S)	Phonology
Phrases and sentences	The ability to determine the meaning of spoken and written phrases and sentences and to generate original phrases and sentences using grammatical forms. (L, R, S, W)	Syntax
Sound-symbol correspondences	The ability to identify the relationship between letters of written language (graphemes) and the individual sounds (phonemes). (R, W)	Orthography
Syllables	The ability to identify the division of words into the smallest units of sequential speech sounds, composed of a vowel sound or a vowel-consonant combination. (L, R)	Phonology
Vocabulary words	The ability to identify and determine the meaning of spoken or written words or short phrases in context and to produce spoken or written words relevant to a particular context. (L, R, S, W)	Lexicon
Written English conventions	The ability to recognize and apply written English conventions (such as punctuation, capitalization, spelling, paragraph structure, and format, including text features). (R, W)	Writing conventions

Note: Letters in parentheses refer to related language modalities: L is listening, R is reading, S is speaking, and W is writing.

Source: Sato, Lagunoff, Worth, Bailey, & Butler 2005.

Table E2. Academic language functions

Language demand	Definition/application	Application/interpretation decision rules
<i>Minimum language demand: word-level</i>		
Classifying	The ability to divide things or their attributes or properties into groups according to type.	
Enumeration	The ability to name things separately, one by one.	
Identification	The ability to identify a problem, need, or fact explicit in a text; recognize it; and show that it exists.	Includes information and important details, fact, and opinion.
Labeling	The ability to produce the term corresponding to a given definition.	In labeling a picture, the picture may be interpreted as a definition.
Organization	The ability to give structure to something such as information or data.	
Sequencing	The ability to arrange, or order things.	
<i>Minimum language demand: phrase- or sentence-level</i>		
Comparison/contrast	The ability to examine or look for differences and similarities between two or more things.	
Definition	The ability to say what the meaning of something, especially a word, is.	
Description	The ability to say or write what someone or something is like.	Used to code standards requiring narrative writing.
Explanation	The ability to offer reasons or a cause.	Includes supporting details. Used to code standards requiring expository writing.
Generalization	The ability to infer a trend, an opinion, principle, or make a conclusion based on facts, statistics, or other information.	
Hypothesis	The ability to form an idea or explanation for something that is based on known facts but has not yet been proved.	
Inference	The ability to reason from circumstance or surmise.	

Language demand	Definition/application	Application/interpretation decision rules
Inquiry	The ability to seek information by forming questions.	
Interpretation	The ability to determine or demonstrate understanding of the intended meaning of something, as distinct from the literal meaning.	
Prediction	The ability to say that an event or action will happen in the future, especially as a result of knowledge or experience.	
Retelling	The ability to relate or tell again, possibly in a different form.	Includes restating in own words.
Summary	The ability to express the most important facts or ideas about something or someone in a short and clear form.	
<i>Minimum language demand: discourse-level</i>		
Analysis	The ability to identify the parts of a whole and their relationship to one another.	
Argument	The ability to discuss a point of view with the purpose of creating agreement around a position or conviction.	
Critique	The ability to review or analyze critically.	Includes understanding and knowledge of main idea, context, purpose, audience, point-of-view.
Evaluation	The ability to use critical reading and thinking to judge and assign meaning or importance to a particular experience or event.	
Negotiation	The ability to engage in a discussion with the point of creating mutual agreement from two or more different views.	
Persuasion	The ability to convince others of something.	
Synthesis	The ability to identify the relationships between two or more ideas or other textual elements.	

Source: Sato et al. (2005).

Appendix F. Overview and protocol for cognitive interviews

This protocol is intended to guide researchers in conducting cognitive interviews in conjunction with Regional Education Laboratory West (REL-W) for the study entitled *Assessment Accommodations for English Language Learners* (Study G). It includes an overview of the study, rationale for including cognitive interview data, details about the student sample, the step-by-step protocol and script for the cognitive interviews, and approved protocol adaptations to address the special linguistic needs of English language learners. Companion documents include the WestEd Study G data collection rubric and the WestEd Study G staff training modules.

Study overview

This study is designed to examine the effects of linguistic modification on the validity of assessments for English language learner students. Specifically, it systematically investigates the ways in which linguistic modification, used as a test accommodation, affects students' ability to access math content during testing. This study aims to increase understanding of the effects of a test accommodation that holds promise as a means of decreasing the achievement gap between English language learner students and non-English language learner students.

Linguistic modification is a theory-based process in which the language in test items, directions, and response options is modified in ways that clarify and simplify the text without simplifying or significantly altering the construct tested (Abedi et al. 2005). To facilitate comprehension, linguistic modification reduces the language demands of text through strategies such as reduced sentence length and complexity, use of common or familiar words, and use of concrete language (Abedi, Lord, & Plummer 1995; Sireci, Li, & Scarpati 2002). Increased access, via linguistic modification, is believed to minimize the effects of construct-irrelevant language demands on English language learner students. In this way, the accommodation facilitates English language learner students' ability to demonstrate their content-related and construct-relevant knowledge and skills, without simplification of the content or significant alteration of the construct tested. Findings from this study may advance current understanding of technically sound assessment practices by presenting empirical evidence about the ways in which increasing English language learner student access to tested content may yield more valid measures of what students know and can do.

The first phase of this study is focused on the development and validation of the instruments used to measure the effectiveness of the accommodation. An initial step in this process was to convene a panel of experts, which included specialists in educational measurement, math content, applied linguistics, English language development, and the English language learner student population. This workgroup developed guidelines for linguistic modification that then were applied to the concurrent development of two versions of a math test—an original version and a linguistically modified version.

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

Experts performed a multistep review process to develop the two versions of the test used during the cognitive interviews. First, they conducted a general screening process, collecting a large pool of released math National Assessment of Educational Progress and state test items with known psychometric qualities. Second, from among these items, experts then selected a range of items in terms of content (for example, measurement, algebra, and number operations) and cognitive complexity assessed. Third, the experts applied the linguistic modification guidelines to each item, eliminating those that did not strongly lend themselves to linguistic modification. Finally, from among the pool of linguistically modified items, they selected a subset of items that were judged to have the greatest potential to yield rich information from students about the effectiveness of particular linguistic modification strategies.

Rationale for including cognitive interview data

Cognitive interviews will be used to examine the effectiveness of the linguistic modification strategies and to identify the most appropriate items for pilot testing. Specifically, the cognitive interviews are intended to help answer two of the study's research questions:

- What are the cognitive processes by which test items are understood by students?
- Do these processes differ for linguistically modified items as compared to original items?

These data will be used to help researchers (1) better understand the degree to which linguistic modification increases access to items for English language learner students, (2) identify the most promising items for pilot testing, and (3) refine questions on the language background survey.

Cognitive interviewing strategies are drawn from the family of process-tracing or verbal protocol models that can be used as to confirm or verify hypotheses about linguistic access. They provide a forum for researchers to test assumptions about the intent of an item or question by microanalyzing the items (Solano-Flores & Trumbull 2003) while simultaneously gathering information about students' understanding of task expectations; their level of mastery of the content; and the reasoning processes, problem-solving strategies, and adaptive skills students use when answering test questions (Ericsson & Simon 1980, 1993; Paulsen & Levine 1999).

During a cognitive interview, researchers observe students individually as they respond to test questions (Ericsson & Simon 1993). As students attempt to answer each item or solve each problem, they are encouraged to articulate, or think out loud, about their interpretation of the task required and the steps or processes needed to complete the task (concurrent data collection). Once the student has responded to all test items, the researcher asks each student a set of follow-up questions to clarify or verify comments collected earlier and to probe deeper into the student's thinking processes about that item (retrospective data collection).

This multistep process helps reveal the types of prior/background knowledge and requisite skills that may support students' ability to respond to the item and to assess the consequences of their

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

decisions (Kopriva 2001). Data collected through cognitive interviews provide validation for inferences about performance outcomes by indicating the degree to which student understanding concurs with the construct intended to be measured by the item.

Study logistics

This section describes the main elements of the study logistics.

Student sample

The cognitive interviews will be conducted with a convenience sample of nine middle-school students. Characteristics of the final sample of eight students⁴⁰ will include the following:

- All will be enrolled in a California middle or junior high school.
- Five will be Spanish-speaking English language learner students, whose level of English language proficiency is intermediate or advanced.
- Four will be non-English language learner general education students (students who do not have an Individualized Education Plan⁴¹).
- Both genders will be represented.

Student participants will be recruited through a two-step process: (a) contacting a small district that has expressed interest in participating in a study about English language learner students and securing their cooperation and (b) identifying eligible students for study participation from among the student population in that district. Recruitment will be ongoing until the target sample is reached. Although this is a relatively small sample, the participants will be selected purposefully to best represent the target population. From these interviews, highly specific, finely grained, richly descriptive data will be collected that can be used to inform the linguistic modification process.

Informed consent

Parents of sampled students will receive information about the study (available in English and Spanish) and be asked to sign and return a consent form only if they do not want their students to participate in the study (passive consent). On the day of testing, eligible students will be asked to indicate their assent by signing a consent form (available in English and Spanish) that is separate from their test booklet. Students will be informed that participation is voluntary but appreciated

⁴⁰ Research suggests that student sample sizes as small as five will yield sufficient information about problem-solving strategies (see, for example, Nielsen 1994).

⁴¹ Special education students whose Individualized Education Plans require eligible test accommodations will be included in the pilot and operational test administrations.

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

and that they may refuse to answer any item or question. Each student will receive a pen to thank him/her for participating.

Assignment of test booklets

Once the nine participating students have been confirmed, each will be assigned a unique identifier. Prior to the day of testing, one of the nine cognitive interview test booklets will be randomly assigned to each student. Only the student's unique identifier will appear on the test booklet. All test booklets will include instructions in both English and Spanish. The ordering of materials in each test booklet is as follows:

- Test directions page.
- Item 1: multiple-choice item.
- Item 2: multiple-choice item.
- Item 3a: original version of a matched pair of items.
- Item 3b: linguistically modified version of a matched pair of items.
- Item 4a: original version of a matched pair of items.
- Item 4b: linguistically modified version of a matched pair of items.
- Item 5: multiple-choice item.
- Language background survey.

Time and location

The interviews are scheduled to take place in March during school hours. To encourage participation, interviews will be conducted in the student's school. Test items will be administered in paper-pencil format. On average, total interview time is estimated to be approximately 70 minutes (10 minutes for practice + 10 minutes per item + 10 minutes for the survey). The audio portion of the interview will be recorded and transcribed following the interview.

Spanish translations

The interviews may be conducted in Spanish, although the math test items are presented to students in English only. The scripted portions of the interview (including prompts and probes) will be translated as directly as possible from English to Spanish. In the test booklet, general

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

directions and language background survey questions are presented to the students in side-by-side English and Spanish versions.

Training of researcher-interviewers

The interviews will be conducted by two WestEd researchers who have experience in working with middle-school students. The researchers will work in teams of two; one will serve as the interviewer (R-1) and the other as the silent observer (R-2). One team of researchers will conduct the interviews in English with the four non-English language learner students. The second team of researchers will be bilingual and will conduct the interviews in Spanish and English with the five English language learner students.

Each researcher will be trained to elicit and record student responses during an interview using the WestEd protocol and will receive specialized training in using the data collection rubric. Specifically, the training will include practice in each of the following:

- Modeling the think aloud process.
- Establishing rapport with participants.
- Taking notes while observing.
- Following a script.
- Eliciting responses, including strategic use of prompts and probes.
- Identifying cues for when a student has completed an item.
- Completing the *Data Summary Rubric* during post-interview debriefing.

Training sessions continue until accuracy and consistency in the implementation of procedures are assured.

The researchers will observe the students as they work on each item and, to the extent possible, record student comments via handwritten notes. To ensure reliability of findings, both researchers will record observations and student responses throughout the interview (as possible) and the entire interview will be recorded (audio only). The R-1 primarily will record observations, while the R-2 will try to capture student comments as well as observations. During the course of each interview, the R-2 will manage the audio recording and document the starting time for the interview, the starting and ending time for each item attempted, and the time at which the interview ended. Following each interview, the researchers will conduct a debriefing session in which they collaboratively review the R-2's notes and agree upon an official transcript of the interview. At this time, they will begin completing the *Data Summary Rubric*. All audio recordings will be transcribed for post-interview analyses.

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

Cognitive interview protocol and script

The cognitive interviews are conducted in four steps. In the first step, the student is introduced to the interview process and allowed to practice thinking aloud. In the second step, data are collected concurrently as the student thinks out loud as he/she attempts to solve each math item. Via prompts, the researcher may interact with the student to elicit verbal responses that describe his/her understanding of the problem and strategies for solving it. In the third stage, the retrospective stage of data collection, students are asked specific questions (probes) about individual items after solving the full set of problems. Students are encouraged to look back, recall, and discuss what they did to solve a problem or draw conclusions about similar items; in this way, they are verifying or clarifying their earlier comments while also providing additional information about their thought processes. In the final step, students are asked to respond to the questions on the language background questionnaire.

Step 1. Training and practice

Each student will be interviewed individually. The student will be invited to sit down at a table in a quiet room. To facilitate communication, the R-1 should be seated so that his/her chair is perpendicular to the student's chair. The R-2 should be seated in a position that facilitates observation and recording of student comments. The R-2 will document the starting time.

The researchers will converse for a few minutes with each student to establish rapport, saying, for example, "How are you today?" or "Isn't it a lovely day?" Then the R-1 will say:

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

Thank you for joining us today. We are interested in learning more about how students solve problems on math tests, so we would like to listen to you as you think about how to answer five test questions. We would like to see and hear how you decide to solve each problem, but you do not need to worry about whether or not you answer the item correctly. You can help us by thinking out loud as you consider your plan for answering each item. What you say is really important to us, so we will take notes as you work. Now I'm going to demonstrate this process by thinking out loud while I work on a test item. This means that I'm going to talk about what I'm thinking while I work. Then we'll let you practice on a few items.⁴²

The R-1 then will demonstrate the think aloud process using Sample Item 1 while the R-2 takes notes. The R-1 will read the item stem out loud, talk about what she thinks the item is asking, describe what she thinks she should do to solve the problem, then solve it/answer the question on paper while verbalizing her thoughts. The R-1 then will ask the student if he/she has any questions about the process just observed.

Each student then will have the opportunity to practice "thinking out loud" with Sample Item 2. Students continue to practice, using different sample items as needed, until the student confirms that he/she understands the process.

Step 2. Concurrent data collection

When the researchers agree that the student appears to understand the process, the R-1 will ask the student:

Do you have any questions before we begin? Okay, let's start with this item. [R-1 shows the student the first item.] Just do like we practiced: read the item out loud, take a few moments to think about the problem, and then talk out loud while you work on that problem. You may ask questions as you work, but remember not to worry about whether or not you have the right answer. Are you ready to see the first math test item?

Using test booklets developed for this process, the R-1 will show each student three different multiple-choice test items (original or linguistically modified) and two sets of matched pairs of items (original and linguistically modified). Researchers will monitor the student's comfort level throughout the process and provide appropriate reassurances if the student appears anxious or confused.

Researchers may reinforce verbalization by nodding approvingly. The R-1, however, will avoid influencing or guiding the student's response through visual or verbal cues or by indicating whether the student's response was correct.

If students are silent for more than 10 seconds, the R-1 may remind them to verbalize their thoughts, using prompts such as, "I wonder what you are thinking now. . .," or "Can you tell us

⁴² Script modified from Johnstone, Bottsford-Miller, & Thompson 2006; van Someren 1994; and Willis 1999.

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

more about what you are thinking?” or “Remember to keep talking.” For the matched pairs of items, the R-1 will ask the student to consider both items at the same time to explore his or her thinking about possible differences between the two versions of the item. The R-1 will try to time the prompts to spur articulation without disrupting the student’s thinking processes (Ericsson and Simon 1993).

If the student appears reluctant to speak, makes a comment that needs clarification, or only partially verbalizes a strategy, the R-1 may prompt the student to extend his or her thinking out loud. Prompts may include statements such as, “Can you tell me more about what that item was asking you to do?” or “How did you know what to do first?”

When it appears that the student has completed an item, the R-1 will ask “Would you like to tell me anything else about that item (or item pair)?”

If the student says yes, the R-2 will continue to record all comments. If the student says no, the R-1 shows the student the next item in the test booklet and repeats the steps above. When the student has completed Item 5, the researchers move to Step 3.

Step 3. Retrospective data collection

When it appears that the student has completed the last test item, the R-1 will ask each of the following questions (with pauses after each for student comment) to elicit additional information about (1) one particular item or item type (MCs); (2) a matched pair of items; or (3) the whole set of items:

1. “Let’s look again at Item 1 *or* Item 2 *or* Item 5” (*all single MC items*).
 - Which word(s) helped you solve this problem?
 - Were there words you didn’t understand in this item?
 - What did you need to know or be able to do to solve this problem?
 - Were you unsure how to solve this problem?
 - What was tricky about solving this/these problem(s)?
2. “Let’s look again at Items 3a and 3b *or* 4a and 4b” (*matched pairs*). These items are similar in certain ways but also different in special ways.”
 - Which item was easier for you to understand?
 - (if saw a difference) What parts of that item helped you understand the problem?
 - (if did *not* see a difference) How were these items alike?
3. “Do you have any other questions or comments about this set of items?”

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

Step 4. Student language background questionnaire

When the student has completed Step 3, the R-1 will say,

Thank you for letting us listen to you today. Your comments have been very helpful. Now we would like you to answer a few questions about your language background. There are no right or wrong answers to these questions. This information will be used only by the researchers to help us understand the ways in which language may help students when they take tests.⁴³

The R-1 and R-2 should continue to monitor the student's responses to the questions on the survey. When the student has completed the language background survey, the researchers provide each student with a pen to thank him/her for participating. The researchers then escort the student to the agreed-upon location.

Special adaptations to the protocol for English language learner students

To ensure that the special linguistic needs of the four English language learner students are addressed, research-based adaptations to the existing protocol may be implemented (Goerman 2006). Two Spanish-speaking bilingual researchers will be trained to administer the cognitive interviews to the English language learner students. Researchers will be especially attentive to the student's comfort level so they may provide appropriate verbal reassurances and/or clarifications to support the English language learner student as he/she works. The researchers will be asked to provide notes that have been translated into English to supplement the audio recording and facilitate post-interview analyses.

The interviews may be conducted in Spanish, although the math test items are presented to students in English only. The scripted portions of the interview (including allowable prompts and probes) will be translated as directly as possible from English to Spanish. Test directions and the language background survey questions will be available in side-by-side English and Spanish versions.

Post-interview debriefing

The researchers will examine their handwritten notes and compare comments. The R-2 will have recorded primarily observations, while the R-2 likely will have captured student comments and responses to prompts or probes. The researchers will use their notes and the audio recording to clarify comments, verify observations, and elucidate their sets of notes. Together, they will complete a *Post-Interview Data Summary Rubric* for each interviewee.

⁴³ Data from surveys are intended to gather information about factors known to covary with test performance, such as primary language spoken in the home and number of years in U.S. schools.

This protocol and script were developed by WestEd. Any use of this protocol or script requires WestEd's consent.

Using this collaborative process, a final transcript of the interview will be constructed that includes direct quotes and all other comments or observations in an item-by-item format; comments recorded during the administration of the language background survey; and preliminary discussion of overall themes in the data. In conjunction with the data charted on the rubric, these data will be used to inform expert judgment related to the effectiveness of the linguistic modification strategies, to help identify the most promising items for pilot testing, and to refine questions on the language background survey.

Appendix G. Item parameter estimates for IRT models

As discussed in chapter 3, the random group equating method based on the EP student subgroup was used to place the item and person parameter estimates on a common metric so that the comparisons between two item sets by student subgroups would be meaningful.

Table G1 presents the resulting item parameter estimates for the one-parameter logistic item response theory (IRT) model (1-PL). Only the item difficulty parameter estimates (b) were reported for the 1-PL model. The range of item difficulty parameters is typically between -3 and $+3$ (recall that the origin of the scale was set at zero, the mean theta score of EP students), with a higher value indicating a more difficult item. Using Item 1 as an example, the difficulty parameter before linguistic modification was estimated to be -1.41 (with a standard error of 0.1098). After linguistic modification, it became -1.37 . Thus, for EP students, the linguistically modified version of Item 1 was slightly more difficult than the original item, as the difficulty parameter for the linguistically modified version is 0.04 units greater than that for the original item (that is, $-1.37 - [-1.41] = 0.04$).

Table G2 presents similar information as in table G1 but includes additional item parameter estimates, the slope parameter estimates (a), under the two-parameter logistic IRT model (2-PL). The slope under both 2-PL and three-parameter logistic IRT model (3-PL) can be any nonzero positive real number, with a higher number indicating being more able to distinguish the high-performing students from the low-performing students. As seen in table G2, Item 19 appeared to have the most discriminating power before and after linguistic modification ($a = 1.05$ and 1.00 , respectively).

For the 3-PL model (table G3), the pseudo-guessing parameter estimates (c) were added in addition to the item difficulty and discrimination parameter estimates. The range of c is typically between 0 and 1 , with a higher number indicating a higher probability of occurrence of guessing behavior. Since there were four response choices per item, one could use 0.25 (guessing totally by random) as a yardstick to examine the guessing behavior for each item. Overall, it appeared that (1) the guessing behavior was about the same before and after linguistic modification for most items and (2) only three items in either item set had the guessing parameter estimate larger than 0.25 .

Table G1. Item parameter estimates for 1-PL model

Item	Original item set		Linguistically modified item set	
	<i>b</i>	se_ <i>b</i>	<i>b</i>	se_ <i>b</i>
1	-1.41	0.1098	-1.37	0.1098
2	-2.12	0.1385	-1.65	0.1205
3	-0.91	0.0970	-1.11	0.1032
4	-0.29	0.0883	-0.28	0.0897
5	-0.29	0.0921	-0.30	0.0933
6	-2.13	0.1335	-2.02	0.1320
7	-1.36	0.1087	-1.14	0.1054
8	-2.95	0.1751	-3.16	0.1942
9	-0.38	0.0903	-0.52	0.0954
10	-1.53	0.1120	-1.57	0.1156
11	0.51	0.0891	0.57	0.0927
12	0.98	0.0958	0.69	0.0930
13	-0.45	0.0913	-0.61	0.0958
14	0.39	0.0874	0.29	0.0902
15	-1.33	0.1064	-1.29	0.1071
16	0.01	0.0871	0.18	0.0885
17	-0.93	0.0984	-0.18	0.0900
18	-0.98	0.0979	-0.81	0.0970
19	0.55	0.0933	0.50	0.0958
20	-0.62	0.0925	-0.49	0.0929
21	-0.23	0.0893	-0.11	0.0892
22	0.74	0.0927	0.69	0.0944
23	-0.57	0.0915	-1.60	0.1181
24	-0.72	0.0937	-1.00	0.1000
25	-1.17	0.1039	-1.10	0.1024

Note: *b* refers to the item difficulty parameter estimate, and se_ *b* is the corresponding standard error for the parameter estimate. In this table the item difficulty parameter estimates were obtained using the ConQuest software (Wu et al. 2007), while the standard errors reported here were reproduced by PARSCALE (Muraki & Bock 2003) software. This was to compare the standard errors under the 2-PL and 3-PL models, where PARSCALE was used for model estimation.

Source: Authors' analyses of primary data.

Table G2. Item parameter estimates for 2-PL model

Item	Original item set				Linguistically modified item set			
	<i>a</i>	se <i>a</i>	<i>b</i>	se <i>b</i>	<i>a</i>	se <i>a</i>	<i>b</i>	se <i>b</i>
1	0.62	0.0709	-1.37	0.1474	0.73	0.0811	-1.20	0.1201
2	0.36	0.0633	-3.21	0.5447	0.30	0.0530	-2.90	0.5064
3	0.39	0.0560	-1.26	0.1947	0.53	0.0660	-1.21	0.1513
4	0.38	0.0517	-0.40	0.1244	0.42	0.0527	-0.37	0.1157
5	0.93	0.0900	-0.23	0.0662	0.87	0.0865	-0.25	0.0677
6	0.98	0.1257	-1.54	0.1407	0.91	0.1169	-1.54	0.1468
7	0.61	0.0714	-1.35	0.1493	0.57	0.0674	-1.19	0.1437
8	0.80	0.1173	-2.37	0.2655	0.56	0.0952	-3.30	0.5011
9	0.44	0.0562	-0.48	0.1167	0.41	0.0561	-0.69	0.1372
10	0.41	0.0584	-2.07	0.2911	0.48	0.0661	-1.84	0.2354
11	0.61	0.0639	0.50	0.0935	0.66	0.0674	0.52	0.0880
12	0.56	0.0617	1.02	0.1272	0.41	0.0527	0.94	0.1531
13	0.50	0.0583	-0.51	0.1087	0.56	0.0656	-0.64	0.1061
14	0.47	0.0554	0.46	0.1106	0.43	0.0532	0.38	0.1140
15	0.73	0.0800	-1.16	0.1168	0.55	0.0682	-1.36	0.1628
16	0.39	0.0515	0.01	0.1130	0.46	0.0552	0.21	0.1024
17	0.71	0.0738	-0.83	0.0975	0.49	0.0561	-0.21	0.0985
18	0.94	0.0952	-0.75	0.0791	0.76	0.0781	-0.70	0.0856
19	1.05	0.0990	0.38	0.0652	1.00	0.0896	0.35	0.0665
20	0.46	0.0577	-0.75	0.1287	0.40	0.0538	-0.67	0.1394
21	0.61	0.0664	-0.23	0.0841	0.55	0.0600	-0.12	0.0885
22	0.74	0.0683	0.63	0.0853	0.80	0.0746	0.56	0.0787
23	0.39	0.0529	-0.79	0.1503	0.43	0.0615	-2.06	0.2798
24	1.01	0.0973	-0.53	0.0697	0.93	0.0953	-0.77	0.0807
25	0.67	0.0767	-1.08	0.1219	0.71	0.0776	-0.98	0.1064

Note: *a* is the item discrimination parameter estimate, and *b* is the item difficulty parameter estimate; se_*a* and se_*b* are the corresponding standard errors for the parameter estimates.

Source: Authors' analyses of primary data.

Table G3. Item parameter estimates from 3-PL model

Item	Original item set						Linguistically modified item set					
	<i>a</i>	se <i>a</i>	<i>b</i>	se <i>b</i>	<i>c</i>	se <i>c</i>	<i>a</i>	se <i>a</i>	<i>b</i>	se <i>b</i>	<i>c</i>	se <i>c</i>
1	0.70	0.0969	-0.90	0.2406	0.22	0.0884	0.82	0.1159	-0.80	0.2081	0.21	0.0849
2	0.37	0.0655	-2.68	0.5492	0.21	0.0921	0.32	0.0567	-2.21	0.5438	0.21	0.0911
3	0.44	0.0764	-0.57	0.3578	0.22	0.0898	0.59	0.0897	-0.70	0.2681	0.21	0.0866
4	0.54	0.1158	0.52	0.3012	0.26	0.0793	0.53	0.0936	0.26	0.2671	0.20	0.0754
5	1.42	0.2493	0.17	0.1012	0.20	0.0465	1.51	0.2879	0.23	0.1013	0.23	0.0452
6	1.03	0.1441	-1.36	0.1741	0.17	0.0763	0.99	0.1382	-1.32	0.1756	0.17	0.0753
7	0.65	0.0904	-0.94	0.2439	0.20	0.0854	0.60	0.0833	-0.81	0.2407	0.18	0.0789
8	0.83	0.1176	-2.20	0.2654	0.18	0.0837	0.57	0.0920	-3.05	0.4630	0.19	0.0869
9	0.81	0.1808	0.55	0.2018	0.33	0.0614	0.62	0.1361	0.35	0.2922	0.31	0.0808
10	0.41	0.0639	-1.61	0.3727	0.19	0.0852	0.49	0.0725	-1.46	0.3145	0.19	0.0842
11	0.94	0.1824	0.84	0.1209	0.16	0.0410	0.97	0.1763	0.80	0.1118	0.13	0.0384
12	1.08	0.2421	1.22	0.1168	0.15	0.0307	0.61	0.1377	1.35	0.2026	0.16	0.0495
13	0.63	0.1073	0.09	0.2376	0.21	0.0750	1.00	0.2056	0.27	0.1744	0.34	0.0606
14	0.83	0.1820	1.00	0.1526	0.21	0.0464	1.07	0.2605	1.06	0.1332	0.28	0.0384
15	0.77	0.1033	-0.86	0.1958	0.18	0.0769	0.59	0.0805	-1.00	0.2436	0.17	0.0779
16	0.49	0.0933	0.62	0.2653	0.18	0.0694	0.70	0.1468	0.79	0.1857	0.21	0.0582
17	0.81	0.1094	-0.48	0.1767	0.17	0.0705	0.55	0.0800	0.15	0.1916	0.13	0.0565
18	1.05	0.1408	-0.49	0.1359	0.15	0.0629	0.93	0.1368	-0.30	0.1625	0.20	0.0685
19	1.64	0.2918	0.56	0.0724	0.11	0.0277	2.06	0.4171	0.58	0.0689	0.14	0.0265
20	0.55	0.0922	-0.17	0.2665	0.20	0.0799	0.53	0.1017	0.15	0.3096	0.24	0.0847
21	1.14	0.2267	0.43	0.1290	0.27	0.0494	0.79	0.1474	0.45	0.1741	0.22	0.0606
22	1.34	0.2298	0.84	0.0854	0.13	0.0269	1.25	0.2119	0.77	0.0866	0.12	0.0292
23	0.46	0.0788	-0.13	0.3120	0.19	0.0808	0.45	0.0682	-1.56	0.3535	0.20	0.0891
24	1.11	0.1318	-0.35	0.1037	0.10	0.0462	1.09	0.1514	-0.47	0.1382	0.17	0.0650
25	0.75	0.1045	-0.71	0.2051	0.19	0.0782	0.84	0.1289	-0.50	0.2054	0.23	0.0820

Note: *a* is the item discrimination parameter estimate, *b* is the item difficulty parameter estimate, and *c* is the item guessing parameter estimate; se_*a*, se_*b*, and se_*c* are the corresponding standard errors for the parameter estimates.

Source: Authors' analyses of primary data.

Appendix H. Descriptive statistics from four scoring approaches

Table H1. Mean math raw scores, by grade, student subgroup, and item set

Grade and student subgroup	Original item set			Linguistically modified item set			Total		
	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation
<i>Grade 7</i>									
EL	290	8.04	3.25	281	8.98	3.76	571	8.50	3.54
NEP	396	9.96	3.88	364	10.73	4.01	760	10.33	3.96
EP	438	15.89	4.62	441	15.66	4.56	879	15.77	4.59
Total	1,124	11.78	5.26	1,086	12.28	5.07	2,210	12.02	5.17
<i>Grade 8</i>									
EL	316	8.72	3.73	327	9.32	4.03	643	9.02	3.90
NEP	425	10.48	3.80	440	10.66	4.09	865	10.57	3.95
EP	445	15.29	4.69	454	15.59	4.61	899	15.44	4.65
Total	1,186	11.82	4.98	1,221	12.14	5.06	2,407	11.98	5.02
<i>Grades 7 and 8 combined</i>									
EL	606	8.40	3.52	608	9.16	3.91	1,214	8.78	3.74
NEP	821	10.23	3.85	804	10.69	4.05	1,625	10.46	3.95
EP	883	15.59	4.66	895	15.63	4.58	1,778	15.61	4.62
Total	2,310	11.80	5.12	2,307	12.20	5.07	4,617	12.00	5.10

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

Source: Authors' analyses of primary data.

Table H2. Mean theta estimates from the 1-PL model, by grade, student subgroup, and item set

Grade and student subgroup	Original item set			Linguistically modified item set			Total		
	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation
<i>Grade 7</i>									
EL	290	-1.58	0.46	281	-1.40	0.55	571	-1.49	0.52
NEP	396	-1.18	0.55	364	-1.04	0.58	760	-1.11	0.57
EP	438	0.06	0.84	441	0.01	0.82	879	0.03	0.83
Total	1,124	-0.08	0.96	1,086	-0.71	0.91	2,210	-0.76	0.94
<i>Grade 8</i>									
EL	316	-1.49	0.52	327	-1.35	0.59	643	-1.42	0.56
NEP	425	-1.10	0.53	440	-1.05	0.60	865	-1.08	0.57
EP	445	-0.06	0.84	454	0.00	0.82	899	-0.03	0.83
Total	1,186	-0.81	0.90	1,221	-0.74	0.90	2,407	-0.78	0.90
<i>Grades 7 and 8 combined</i>									
EL	606	-1.53	0.49	608	-1.37	0.57	1,214	-1.45	0.54
NEP	821	-1.14	0.54	804	-1.05	0.59	1,625	-1.09	0.57
EP	883	0.00	0.84	895	0.00	0.82	1,778	0.00	0.83
Total	2,310	-0.81	0.93	2,307	-0.73	0.90	4,617	-0.77	0.92

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

Source: Authors' analyses of primary data.

Table H3. Mean theta estimates from the 2-PL model, by grade, student subgroup, and item set

Grade and student subgroup	Original item set			Linguistically modified item set			Total		
	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation
<i>Grade 7</i>									
EL	290	-1.29	0.56	281	-1.16	0.65	571	-1.23	0.61
NEP	396	-0.98	0.66	364	-0.88	0.69	760	-0.93	0.68
EP	438	0.07	0.87	441	0.02	0.87	879	0.04	0.87
Total	1,124	-0.65	0.94	1,086	-0.59	0.92	2,210	-0.62	0.93
<i>Grade 8</i>									
EL	316	-1.19	0.64	327	-1.11	0.71	643	-1.15	0.68
NEP	425	-0.90	0.65	440	-0.88	0.71	865	-0.89	0.68
EP	445	-0.07	0.88	454	-0.02	0.87	899	-0.04	0.87
Total	1,186	-0.67	0.88	1,221	-0.62	0.90	2,407	-0.64	0.89
<i>Grades 7 and 8 combined</i>									
EL	606	-1.24	0.61	608	-1.13	0.68	1,214	-1.19	0.65
NEP	821	-0.94	0.66	804	-0.88	0.70	1,625	-0.91	0.68
EP	883	0.00	0.88	895	0.00	0.87	1,778	0.00	0.87
Total	2,310	-0.66	0.91	2,307	-0.61	0.91	4,617	-0.63	0.91

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

Source: Authors' analyses of primary data.

Table H4. Mean theta estimates from the 3-PL model, by grade, student subgroup, and item set

Grade and student subgroup	Original item set			Linguistically modified item set			Total		
	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation
<i>Grade 7</i>									
EL	290	-1.40	0.62	281	-1.28	0.71	571	-1.34	0.67
NEP	396	-1.05	0.72	364	-0.92	0.74	760	-0.99	0.73
EP	438	0.07	0.88	441	0.02	0.88	879	0.05	0.88
Total	1,124	-0.70	0.99	1,086	-0.63	0.97	2,210	-0.67	0.98
<i>Grade 8</i>									
EL	316	-1.26	0.68	327	-1.20	0.76	643	-1.23	0.72
NEP	425	-0.94	0.69	440	-0.92	0.75	865	-0.93	0.72
EP	445	-0.07	0.90	454	-0.02	0.88	899	-0.04	0.89
Total	1,186	-0.70	0.92	1,221	-0.66	0.95	2,407	-0.68	0.94
<i>Grades 7 and 8 combined</i>									
EL	606	-1.33	0.65	608	-1.24	0.74	1,214	-1.28	0.70
NEP	821	-0.99	0.71	804	-0.92	0.75	1,625	-0.96	0.73
EP	883	0.00	0.89	895	0.00	0.88	1,778	0.00	0.89
Total	2,310	-0.70	0.96	2,307	-0.65	0.96	4,617	-0.67	0.96

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

Source: Authors' analyses of primary data.

Appendix I. ANOVA findings across four scoring approaches

Below are the detailed analysis of variance (ANOVA) summary tables based on raw scores and a one-parameter logistic item response theory (IRT) model (1-PL), a two-parameter logistic IRT model (2-PL), and a three-parameter logistic IRT model (3-PL). As indicated in chapter 4, the interaction of item set by student subgroup (bolded in the tables) was the focus of the analysis. If there would be any effect of linguistic modification—particularly on English language learner students (EL) and non-English-language-arts-proficient who are not English language learners (NEP) but not on English language arts–proficient students who are not English language learners (EP)—one should expect the difference between the two item sets to vary by student subgroup (indicated by a significant interaction effect). If so, three post-hoc comparisons (EL versus NEP, NEP versus EP, EL versus EP) were then be conducted to further examine the subgroup difference on the effect of linguistic modification. Also, a profile plot would be produced to demonstrate this interaction effect. Since only the analysis based on the 1-PL model indicated a significant interaction effect, the post-hoc comparisons and a profile plot were presented for the 1-PL model only (see table 7 and figure 4 in chapter 4).

Table I1. Analysis of variance for linguistic modification effects on student subgroups (based on raw scores)

Source	Partial sum of squares	Degrees of freedom	Mean squares	F-ratio	p-value ^a
Model	40,069.22	11	3,642.66	210.16	0.0000
Form	202.43	1	202.43	11.68	0.0006
Subgroup	39,643.31	2	19,821.65	1,143.57	0.0000
Grade level	20.29	1	20.29	1.17	0.2794
Item set by subgroup	99.44	2	49.72	2.87	0.0569
Item set by grade level	4.61	1	4.61	0.27	0.6059
Subgroup by grade level	140.26	2	70.13	4.05	0.0176
Item set by subgroup by grade level	71.77	2	35.89	2.07	0.1262
Residual	79,818.78				

Note: $N = 4,617$, R-squared = 0.3342.

a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true.

Source: Authors' analyses of primary data.

Table I2. Analysis of variance for linguistic modification effects on student subgroups (based on 1-PL model)

Source	Partial sum of squares	Degrees of freedom	Mean squares	F ratio	p-value ^a
Model	1,803.33	11	163.94	363.08	0.0000
Item set	8.17	1	8.17	18.09	0.0000
Student subgroup	1,787.10	2	893.55	1,979.00	0.0000
Grade level	0.21	1	0.21	0.46	0.4969
Item set by subgroup	4.89	2	2.44	5.41	0.0045**
Item set by grade level	0.01	1	0.01	0.03	0.8640
Group by grade level	3.53	2	1.77	3.91	0.0200
Item set by subgroup by grade level	2.03	2	1.02	2.25	0.1052
Residual	2,079.24	4,605	0.45		

** Significant at $\alpha=0.01$ level

Note: $N = 4,617$, R-squared = 0.4645.

a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true.

Source: Authors' analyses of primary data.

Table I3. Analysis of variance for linguistic modification effects on student subgroups (based on 2-PL model)

Source	Partial sum of squares	Degrees of freedom	Mean squares	F ratio	p-value ^a
Model	1,220.89	11	110.99	196.97	0.0000
Item set	3.50	1	3.50	6.21	0.0128
Student subgroup	1,210.26	2	605.13	1073.93	0.0000
Grade level	0.06	1	0.06	0.11	0.7392
Item set by subgroup	2.13	2	1.07	1.89	0.1509
Item set by grade level	0.07	1	0.07	0.12	0.7268
Group by grade level	5.43	2	2.71	4.82	0.0081
Item set by subgroup by grade level	1.95	2	0.97	1.73	0.1774
Residual	2,594.79	4,605	0.56		

Note: $N = 4,617$, R-squared = 0.3200.

a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true.

Source: Authors' analyses of primary data.

Table I4. Analysis of variance for linguistic modification effects on student subgroups (based on 3-PL model)

Source	Partial sum of squares	Degrees of freedom	Mean squares	F ratio	p-value ^a
Model	1,403.47	11	127.59	207.16	0.0000
Item set	3.14	1	3.14	5.1	0.0240
Student subgroup	1,391.94	2	695.97	1,130.04	0.0000
Grade level	0.75	1	0.75	1.22	0.2691
Item set by subgroup	1.74	2	0.87	1.41	0.2432
Item set by grade level	0.20	1	0.20	0.32	0.5731
Group by grade level	8.39	2	4.19	6.81	0.0011
Item set by subgroup by grade level	2.23	2	1.12	1.81	0.1632
Residual	2,836.14	4605	0.62		

Note: $N = 4,617$, R-squared = 0.3310.

a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true.

Source: Authors' analyses of primary data.

Appendix J. Cross-approach comparisons

Because differences across English language learner students (EL), non-English-language-arts-proficient students who are not English language learners (NEP), and English language arts-proficient students who are not English language learners (EP) in the effects of linguistic modification on student math performance depended on the scoring approach, additional analyses were conducted to explore factors that may have contributed to the disparity in findings.⁴⁴

The differences in effect sizes between raw scores and the one-parameter logistic item response theory (IRT) model (1-PL) estimates (0.15 and 0.17, respectively, as shown in table 6 in chapter 4), while very close, may reflect differences in the precision of estimating student math performance between the classical test theory and IRT approach. Raw scores are based on the number of items answered correctly while the 1-PL model estimation takes into account not only the number of correct responses but also the difficulty level of the items. In comparison with the raw score approach, IRT modeling offers the potential to yield more reliable and accurate student ability estimates (such as math understanding) as well as information about item characteristics (see, for example, Crocker & Algina 1986; Hambleton, Swaminathan, & Rogers 1991). Because of this, the student math understanding estimates obtained from the 1-PL model may be less adversely affected by random error, thereby increasing power to detect group differences in the effect of linguistic modification.

However, factors that contribute to discrepant results across IRT models are less clear. The authors first examined which model (1-, 2-, or 3-PL) would better fit the data using global model fit statistics (log likelihood function, Akaike information criterion [AIC], and Bayesian information criterion [BIC]). A typical chi-square difference test was used to test whether the difference in $-2 \times \log$ likelihood function was statistically significant. If the test were significant, then a model with a lower $-2 \times \log$ likelihood function fits the data better. For both AIC and BIC, the model with a smaller value is preferred. Table J1 below summarizes the model comparisons based on these model fit statistics.

⁴⁴ A constraint in this process is that one cannot directly observe which model yields more reliable and accurate estimates because the true model underlying these data is unknown. In practice, the researcher would choose the model that is believed to work best with the collected data. In this study, however, the authors concluded that presenting findings from all approaches commonly used by state agencies for scoring and interpreting results from achievement tests would be most useful in guiding future research.

Table J1. Evaluation of model fit, by item set, for item response theory models

	$-2*\log(L)$	Number of parameters	Difference in $\log(L)$	Parameters difference	p^a	AIC	BIC
<i>Original item set</i>							
1-PL	24,777.259	25				24,827.26	24,946.84
2-PL	24,543.248	50	234.011	25	<.01	24,643.25	24,882.41
3-PL	24,489.790	75	53.458	25	<.01	24,639.79	24,998.54
<i>Linguistically modified item set</i>							
1-PL	25,288.589	25				25,338.59	25,458.51
2-PL	25,098.053	50	190.536	25	<.01	25,198.05	25,437.89
3-PL	25,028.782	75	69.271	25	<.01	25,178.78	25,538.54

a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true.

Source: Authors' analyses of primary data.

As shown in table J1, using global fit criteria, either the 2-PL or 3-PL model is preferred over the 1-PL model. This is consistent across the two item sets.

It is still uncertain, however, whether the added complexity of these models yielded more precise student math understanding estimates. Given the relatively small sample size in this study, the 1-PL model may be the most appropriate model for this data set and provide an estimate that is most sensitive to differences across student subgroups in the effect of linguistic modification. Since additional parameters are estimated in more complex IRT models (i.e., 2-PL and 3-PL models), a larger sample size typically is needed to yield more reliable estimates.⁴⁵ The study's sample size of less than 900 per subgroup was on the borderline for recommended minimum sample sizes for the 2-PL model and did not meet the recommended minimum sample size for the 3-PL model. Moreover, complications in estimating item difficulty and item discrimination parameters simultaneously in 2-PL and 3-PL models may further reduce the precision of student math understanding estimates. The standard error associated with the difficulty parameter was larger for more items in the 2-PL model than in the 1-PL model (see appendix G for item parameter estimates derived from each IRT model). For both models, the largest standard errors were associated with very easy items. For example, the standard error of the difficulty parameter for Item 8 (an easy item on both item sets under both models) is 0.27 on the original item set and 0.50 on the linguistically modified item set under the 2-PL model whereas they were 0.18 and 0.19, respectively, under the 1-PL model. Note that only 7 percent and 6 percent of examinees did not answer this item correctly on the original and linguistically modified item sets, respectively. Under the 2-PL model, limited data about those students who did not answer this item correctly could have contributed to instability during estimation of the item difficulty

⁴⁵ For the 1-PL model, it is recommended to have at least 200–500 cases; for the 2-PL model, it is recommended to have at least 800–1,000 cases; and for the 3-PL model, it is recommended to have at least 1,200–1,500 cases (Crocker & Algina 1986; Embretson & Reise 2000; Reckase 1979; Hambleton & Jones 1993; Harris 1989; Thissen & Wainer 2001).

parameter as the computer program worked simultaneously to estimate the item discrimination parameters. Several items behaved similarly to this item.

This type of instability is less problematic for 1-PL models, as the feature of “separability of parameters” in the 1-PL model allows estimation of the person parameters without reliance on the item parameters (and vice versa). This property is associated with the sufficient statistics. For the 1-PL model, the sufficient statistic for estimating any student’s latent ability is that student’s total summed score across items; the sufficient statistic of the difficulty parameter for a given item is the sum of item responses (0 or 1) across persons. Although there is a sufficient statistic for the ability estimate under the 2-PL model, it is dependent on the true item discrimination parameters that are usually unknown. With only limited information available about those students who did not answer certain items correctly, the item parameter estimates under the 1-PL model rely less on the student data to yield reliable item estimates. This may explain why the standard errors associated with those items tended to be smaller under the 1-PL model than under the 2-PL model.

Although the more complex IRT models may technically fit the data better than the 1-PL model, there is greater uncertainty about person estimates in the more complex models with limited data. This can result in greater within-cell error variance, and thus less power to detect subgroup differences. This may explain why the interaction effect in the analysis of variance was only statistically significant when the item set score was based on the 1-PL model.

In summary, the disparity among three IRT models in the effect of linguistic modification across EL, NEP, and EP students may have been due to two possible sources. First, the sample size was relatively small for obtaining reliable estimates for the 2-PL or 3-PL model. Second, related to the first the study sample may not provide sufficient information for some items. Together, these factors may introduce increased error variance for (at least) some items when estimating item parameters under the 2- and 3-PL models, which in turn affect the precision of student ability estimates used in the primary analyses.

Appendix K. Results of the classical item-level analyses

For each of the three student subgroups, reported statistics in tables K-1 and K-2 include item p-values (i.e., the percentage of students within each of the subgroups who answered the item correctly) and point biserial correlations (item-total score correlations).

Table K1. Item-level statistics for original item set

Item	Non-English language learner students					
	EL students		NEP		EP	
	p-value ^a	Point biserial correlation	p-value ^a	Point biserial	p-value ^a	Point biserial correlation
1	0.42	0.28	0.55	0.32	0.77	0.32
2	0.58	0.23	0.71	0.29	0.86	0.14
3	0.38	0.22	0.47	0.15	0.68	0.23
4	0.32	0.23	0.40	0.20	0.56	0.24
5	0.18	0.25	0.26	0.29	0.56	0.47
6	0.55	0.30	0.68	0.31	0.86	0.36
7	0.37	0.25	0.47	0.25	0.76	0.31
8	0.61	0.34	0.79	0.38	0.93	0.25
9	0.32	0.18	0.37	0.12	0.58	0.27
10	0.52	0.26	0.61	0.25	0.79	0.21
11	0.10	0.23	0.16	0.25	0.39	0.35
12	0.13	0.03	0.19	0.16	0.30	0.30
13	0.40	0.15	0.46	0.15	0.59	0.31
14	0.25	0.10	0.26	0.09	0.42	0.29
15	0.46	0.10	0.53	0.30	0.76	0.36
16	0.31	0.06	0.35	0.12	0.50	0.25
17	0.31	0.25	0.41	0.26	0.69	0.38
18	0.36	0.26	0.40	0.22	0.70	0.45
19	0.16	0.06	0.17	0.23	0.39	0.49
20	0.33	0.19	0.41	0.21	0.63	0.28
21	0.25	0.13	0.30	0.17	0.55	0.36
22	0.18	0.09	0.17	0.07	0.35	0.39
23	0.26	0.05	0.35	0.17	0.62	0.24
24	0.22	0.23	0.27	0.30	0.65	0.49
25	0.42	0.19	0.50	0.25	0.73	0.36
Mean	0.34		0.41		0.62	
Standard deviation	0.14		0.17		0.17	

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

a. Proportion of students who answered test item correctly.

Source: Authors' analyses based on primary data.

Table K2. Item-level statistics for linguistically modified item set

Item	Non-English language learner students					
	EL students		NEP		EP	
	p-value ^a	Point biserial correlation	p-value ^a	Point biserial	p-value ^a	Point biserial correlation
1	0.42	0.28	0.56	0.33	0.76	0.35
2	0.55	0.23	0.62	0.26	0.80	0.14
3	0.38	0.26	0.50	0.29	0.72	0.30
4	0.34	0.30	0.42	0.10	0.56	0.26
5	0.27	0.32	0.30	0.32	0.56	0.44
6	0.55	0.30	0.65	0.31	0.85	0.36
7	0.39	0.29	0.54	0.37	0.73	0.30
8	0.71	0.36	0.83	0.25	0.94	0.17
9	0.39	0.14	0.43	0.17	0.61	0.25
10	0.56	0.34	0.63	0.28	0.79	0.25
11	0.10	0.25	0.18	0.27	0.38	0.37
12	0.17	0.18	0.19	0.21	0.36	0.25
13	0.40	0.21	0.45	0.20	0.62	0.33
14	0.29	0.05	0.30	0.14	0.44	0.25
15	0.47	0.16	0.57	0.26	0.75	0.30
16	0.35	0.21	0.33	0.15	0.46	0.29
17	0.29	0.27	0.33	0.29	0.54	0.30
18	0.36	0.17	0.43	0.24	0.66	0.40
19	0.18	0.17	0.18	0.25	0.40	0.48
20	0.38	0.18	0.38	0.17	0.60	0.24
21	0.27	0.01	0.31	0.18	0.52	0.33
22	0.20	0.06	0.18	0.09	0.36	0.42
23	0.41	0.30	0.54	0.26	0.80	0.22
24	0.26	0.40	0.33	0.36	0.70	0.44
25	0.44	0.26	0.51	0.26	0.72	0.37
Mean	0.37		0.43		0.63	
Standard deviation	0.14		0.17		0.16	

EL is English language learner students. NEP is non-English-language-arts-proficient students who are not English language learners. EP is English language arts-proficient students who are not English language learners.

a. Proportion of students who answered test item correctly.

Source: Authors' analyses based on primary data.

Appendix L. Summary of differential item functioning findings

Table L1. Summary of findings from analysis of differential item functioning, NEP students versus EP students

Item	Form = original				Form = linguistically modified			
	MH chi-square	<i>p</i> -value ^a	Odds ratio ^b	Class ^c	MH chi-square	<i>p</i> -value ^a	Odds ratio ^b	Class ^c
1	0.0247	0.87508	0.97971	A+	1.4736	0.22477	0.85375	A+
2	0.5637	0.45278	1.12098	A-	2.899	0.08863	1.25596	A-
3	3.5062	0.06114	1.25045	A-	0.7021	0.40209	1.10952	A-
4	0.7263	0.39408	0.90281	A+	0.6747	0.41143	0.9092	A+
5	0.7194	0.39634	1.11692	A-	0.1132	0.73654	1.04352	A-
6	0.1056	0.74527	0.95281	A+	0.2352	0.62773	1.07223	A-
7	11.295	0.00078	1.51952	A-	1.9507	0.16251	0.83291	A+
8	0.0887	0.76579	1.05809	A-	3.0731	0.0796	1.4121	A-
9	1.0313	0.30985	1.12823	A-	0.3834	0.53578	1.07437	A-
10	0.4181	0.5179	1.08938	A-	0.0037	0.95127	0.99194	A+
11	1.8818	0.17013	1.21452	A-	0.2434	0.62175	1.06949	A-
12	2.3642	0.12415	0.80213	A+	0.6355	0.42534	1.11483	A-
13	6.401	0.01141	0.73527	A+	0.3036	0.58164	0.93662	A+
14	0.2474	0.61894	0.93856	A+	0.1396	0.7087	0.95542	A+
15	0.2158	0.64227	0.94026	A+	0.2044	0.65116	0.94398	A+
16	0.8039	0.36992	0.89573	A+	3.1554	0.07568	0.8054	A+
17	1.6074	0.20486	1.16683	A-	0.1876	0.66494	0.94844	A+
18	1.6502	0.19893	1.17405	A-	0.2821	0.59531	1.06696	A-
19	1.2928	0.25554	0.83796	A+	0.0035	0.95302	0.99177	A+
20	0.626	0.42882	1.09887	A-	6.5982	0.01021	1.34813	A-
21	1.2399	0.2655	1.14271	A-	0.6577	0.41738	1.10169	A-
22	0.1372	0.71111	1.05546	A-	0.237	0.62636	1.07203	A-
23	10.274	0.00135	1.4625	A-	15.504	0.00008	1.62985	B-
24	13.003	0.00031	1.60168	B-	24.915	0.00000	1.8807	B-
25	0.1065	0.7442	0.95926	A+	0.0268	0.86992	0.98028	A+

a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true.

b. EP students are the reference group and the NEP students are the focal group. If the odds ratio is greater than 1, the odds favor the EP students—that is, the probability of getting this item right is higher for EP students than for NEP students. If the odds ratio equals 1, the odds of getting the item right is about the same between groups. If the odds ratio is less than 1, the odds favor the NEP students—that is, the probability of getting this item right is higher for NEP students than for EP students.

c. Class was determined based on the delta scale transformed from the odds ratio where A indicates “negligible DIF,” B indicates “moderate DIF,” and C indicates “moderate to large DIF.” A positive sign indicates that the item favors NEP students; a negative sign indicates that the item favors EP students.

Source: Authors’ analysis of primary data.

Table L2. Summary of findings from analysis of differential item functioning, EL students versus EP students

Item	Form = original				Form = linguistically modified			
	MH chi-square	p-value ^a	Odds ratio ^b	Class	MH chi-square	p-value ^a	Odds ratio ^b	Class
1	1.2914	0.25579	1.18804	A-	3.2741	0.07038	1.29453	A-
2	10.305	0.00133	1.68966	B-	5.2963	0.02137	1.40693	A-
3	3.4997	0.06138	1.31159	A-	6.3974	0.01143	1.44237	A-
4	1.0314	0.30984	0.85972	A+	2.2021	0.13783	0.80988	A+
5	0.62	0.43104	1.1368	A-	0.7887	0.37449	0.8719	A+
6	0.4069	0.52355	1.11081	A-	1.1378	0.28612	1.18649	A-
7	10.372	0.00128	1.59053	B-	3.987	0.04585	1.33035	A-
8	14.177	0.00017	2.01938	C-	6.8112	0.00906	1.81098	B-
9	0.154	0.69475	1.05837	A-	0.1769	0.67408	1.05873	A-
10	0.405	0.52451	1.10498	A-	0.7243	0.39473	0.87467	A+
11	3.3213	0.06839	1.3935	A-	6.6211	0.01008	1.62124	B-
12	0.048	0.82663	1.04208	A-	0.0845	0.77134	1.04885	A-
13	4.7096	0.03	0.72599	A+	2.1517	0.14241	0.81587	A+
14	2.0713	0.1501	0.79504	A+	0.1775	0.67352	0.94006	A+
15	0.3336	0.56352	1.08694	A-	2.6809	0.10156	1.26126	A-
16	0.4885	0.48461	0.90018	A+	19.691	0.00001	0.52498	C+
17	2.1178	0.1456	1.24252	A-	0.7933	0.37312	0.87881	A+
18	2.007	0.15658	0.80343	A+	0.6291	0.42767	1.11929	A-
19	2.403	0.1211	0.72804	A+	0.7657	0.38154	0.85473	A+
20	0.6432	0.42256	1.12441	A-	0.0376	0.84624	1.02764	A-
21	0.3479	0.55533	1.09442	A-	4.3448	0.03712	1.3477	A-
22	6.6459	0.00994	0.62488	B+	2.1238	0.14503	0.77435	A+
23	18.349	0.00002	1.87356	B-	24.478	0.00000	2.04172	C-
24	2.3459	0.12561	1.28129	A-	10.752	0.00104	1.64858	B-
25	0.0796	0.77779	1.04205	A-	0.1525	0.69616	0.94565	A+

a. Probability of obtaining a test statistic of the same or larger magnitude as the one observed, given that the null hypothesis is true.

b. EP students are the reference group and the NEP students are the focal group. If the odds ratio is greater than 1, the odds favor the EP students—that is, the probability of getting this item right is higher for EP students than for NEP students. If the odds ratio equals 1, the odds of getting the item right is about the same between groups. If the odds ratio is less than 1, the odds favor the NEP students—that is, the probability of getting this item right is higher for NEP students than for EP students.

c. Class was determined based on the delta scale transformed from the odds ratio where A indicates “negligible DIF,” B indicates “moderate DIF,” and C indicates “moderate to large DIF.” A positive sign indicates that the item favors NEP students; a negative sign indicates that the item favors EP students.

Source: Authors’ analysis of primary data

Appendix M. Exploratory factor analysis results

For each item set within each student subgroup, a series of exploratory factor analyses was conducted to examine the number of underlying factors. Table M1 provides factor loadings for all items, by item set (original or linguistically modified) and student subgroup (English language learner students [EL], non-English-language-arts-proficient non-English language learner students [NEP], and English language arts-proficient non-English language learner students [EP]). Note that these factor loadings were derived based on a one-factor solution.

Table M1. Estimated factor loadings based on one-factor solution, by item set and student subgroup

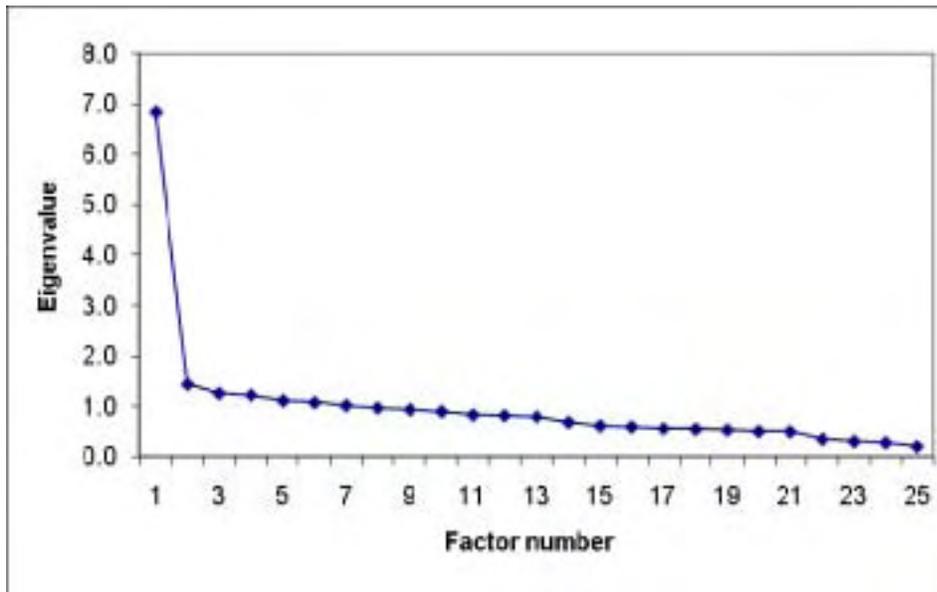
Item	Original item set			Linguistically modified item set		
	EL students	Non-English language learner students		EL students	Non-English language learner students	
		NEP	EP		NEP	EP
1	0.467	0.530	0.487	0.431	0.514	0.541
2	0.399	0.482	0.250	0.407	0.425	0.218
3	0.371	0.230	0.329	0.395	0.439	0.437
4	0.395	0.338	0.336	0.495	0.164	0.369
5	0.452	0.447	0.682	0.484	0.496	0.646
6	0.500	0.525	0.626	0.476	0.517	0.601
7	0.454	0.389	0.476	0.435	0.584	0.458
8	0.626	0.697	0.517	0.636	0.478	0.355
9	0.278	0.181	0.382	0.207	0.229	0.363
10	0.432	0.405	0.326	0.562	0.437	0.385
11	0.450	0.444	0.510	0.478	0.486	0.549
12	0.068	0.272	0.466	0.289	0.329	0.361
13	0.223	0.223	0.435	0.344	0.292	0.474
14	0.146	0.150	0.406	0.054	0.208	0.377
15	0.132	0.474	0.545	0.230	0.401	0.449
16	0.092	0.199	0.352	0.340	0.230	0.407
17	0.416	0.404	0.553	0.453	0.459	0.436
18	0.412	0.361	0.644	0.243	0.367	0.579
19	0.083	0.399	0.734	0.245	0.429	0.713
20	0.326	0.340	0.393	0.276	0.250	0.344
21	0.205	0.276	0.504	0.008	0.266	0.469
22	0.145	0.119	0.580	0.084	0.163	0.625
23	0.079	0.240	0.340	0.476	0.411	0.349
24	0.460	0.490	0.686	0.640	0.558	0.643
25	0.285	0.363	0.525	0.404	0.378	0.544

Source: Authors' analyses based on primary data.

Figures M1–M6 provide scree plots for all items. For both the original and linguistically modified item sets, one dominant factor emerged within each student subgroup.

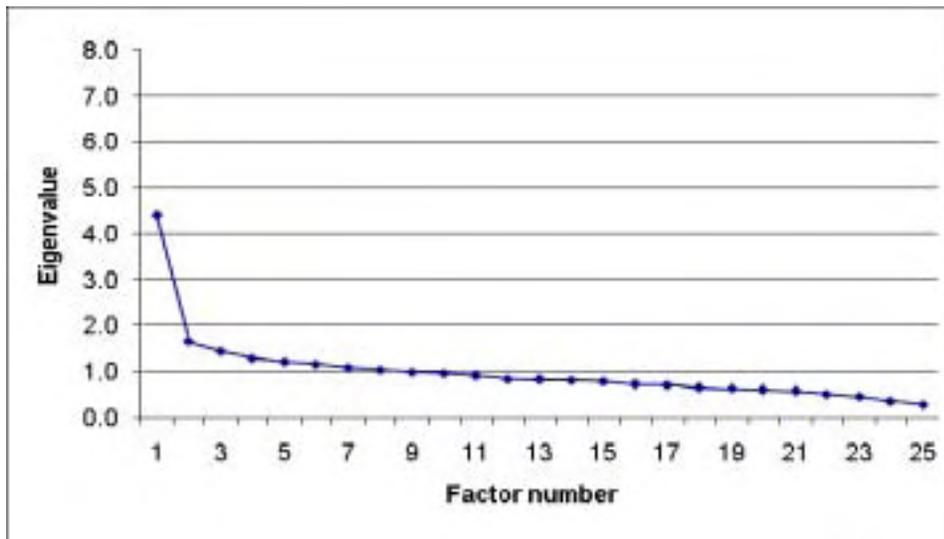
For the original item set responses of EP students resulted in a first factor with an eigenvalue of 6.9 (explaining 27 percent of the total variance), while responses for NEP students resulted in a first factor with an eigenvalue of 4.4 (explaining 18 percent of the total variance), and responses for English language learner students resulted in a first factor with an eigenvalue of 3.9 (explaining 16 percent of the total variance).

Figure M1. Scree plot for non-English language learner students who are proficient in English language arts, taking original item set



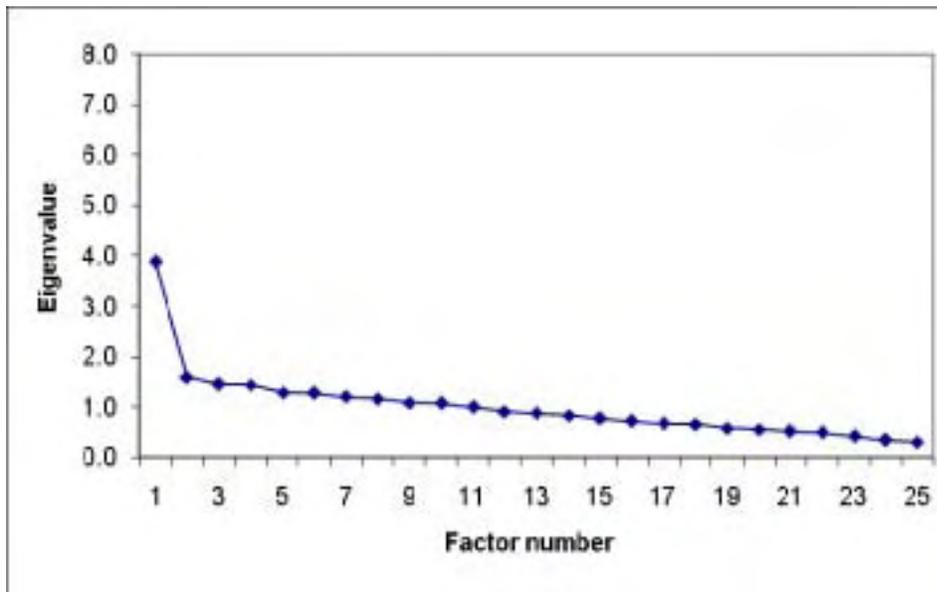
Source: Authors' analyses based on primary data.

Figure M2. Scree plot for non-English language learner students who are not proficient in English language arts, taking original item set



Source: Authors' analyses based on primary data.

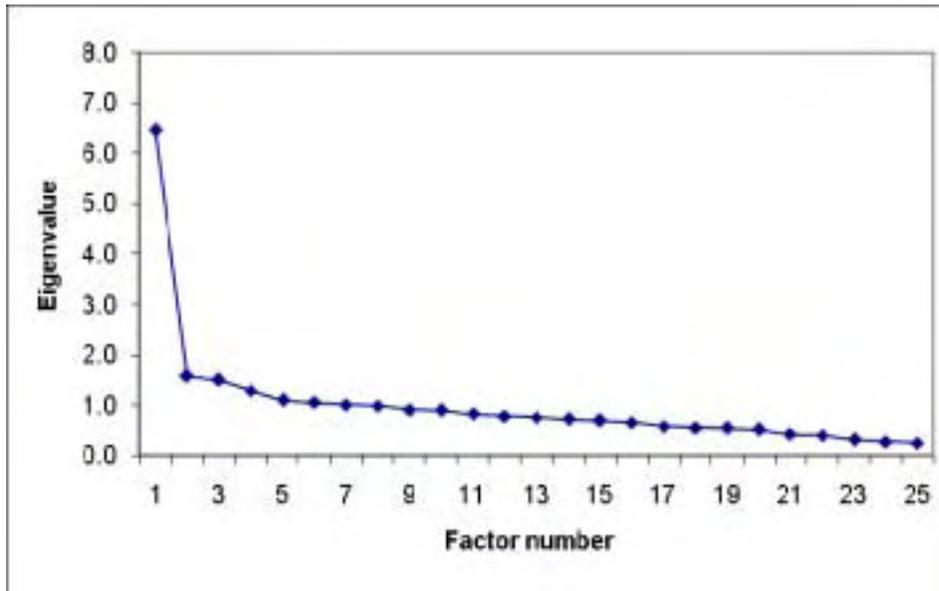
Figure M3. Scree plot for English language learner students taking original item set



Source: Authors' analyses based on primary data.

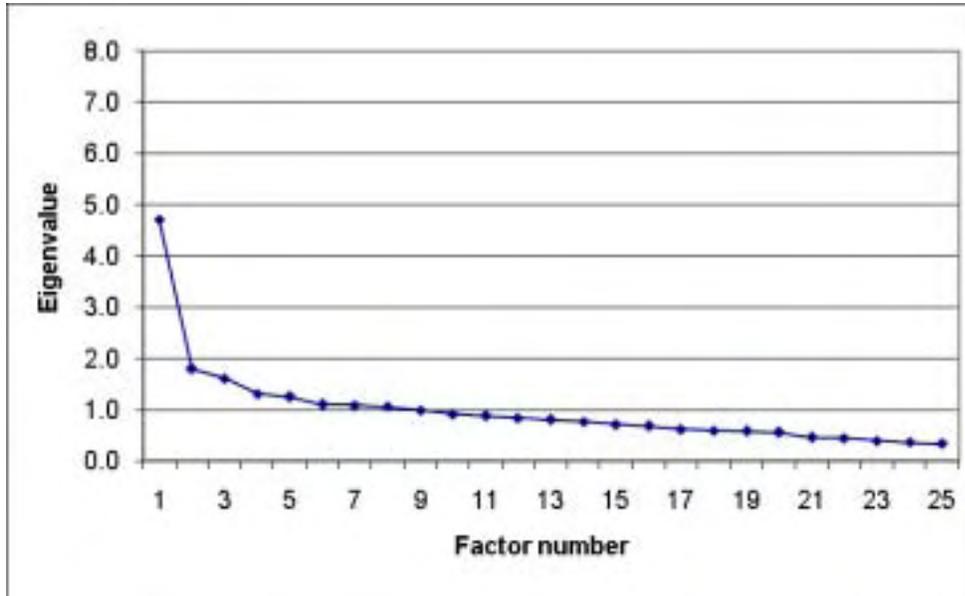
For the linguistically modified item set responses of non-English language learner students who were proficient in English language arts resulted in a first factor with an eigenvalue of 6.5 (explaining 26 percent of the total variance), while responses for non-English language learner students who were not proficient in English language arts resulted in a first factor with an eigenvalue of 4.7 (explaining 19 percent of the total variance), and responses for English language learner students resulted in a first factor with an eigenvalue of 4.7 (explaining 19 percent of the total variance).

Figure M4. Scree plot for non-English language learner students who are proficient in English language arts, taking linguistically modified item set



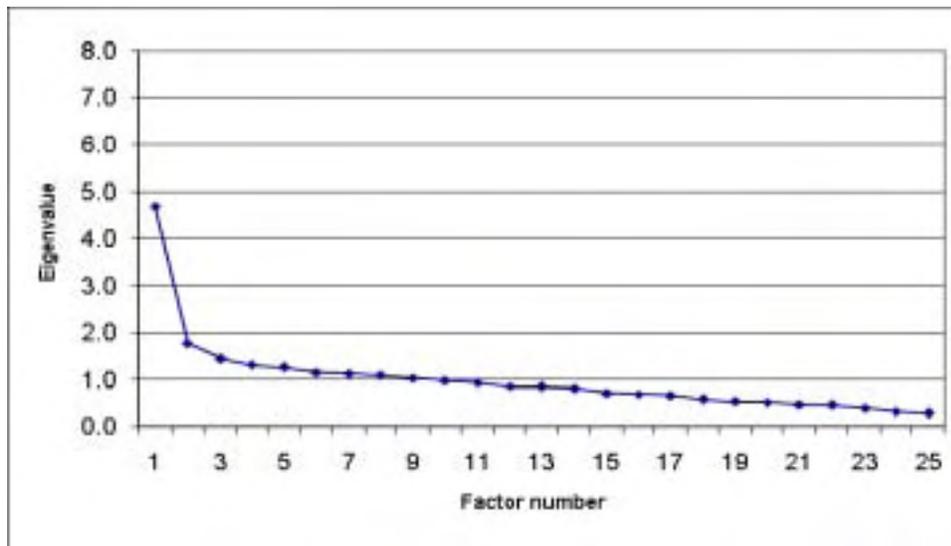
Source: Authors' analyses based on primary data.

Figure M5. Scree plot for non-English language learner students who are not proficient in English language arts, taking linguistically modified item set



Source: Authors' analyses based on primary data.

Figure M6. Scree plot for English language learner students taking linguistically modified item set



Source: Authors' analyses based on primary data.

Research Study

OPERATIONAL TEST FORM-0

Math Test

Grades 7&8

2008

Student Name: _____



Thank you for your participation!

This math test is part of a research project that is studying the ways in which students read and solve problems on different types of tests. Your answers will help us create test items that are fair for all students.

Please try to answer each item on this test just as you would in your classroom. You may raise your hand if you have any questions and a test administrator or teacher will come and assist you.

You will **not** be graded on your answers and this test will **not** be used as part of your grade in your math classroom. You may rest at any time. We estimate that it will take you about 45 minutes in total to complete the test items and answer the questions at the end.

When you have finished all of the test items, **please remember to answer the 5 questions at the end of the test booklet.** You will see that we have provided these 5 questions both in English and in Spanish. Please answer each question only once, either on the English version *or* on the Spanish version.

Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

If you have any comments or suggestions for improving this form, please write to: U.S. Department of Education, Washington, DC, 20202-4651. If you have comments or concerns about your participation in this study, please write to Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington, DC, 20208.

¡Gracias por su participación!

Este examen de matemáticas es parte de un estudio para entender mejor cómo estudiantes leen y resuelven problemas de diferentes tipos de exámenes. Sus respuestas nos ayudarán a crear exámenes justos para todos los estudiantes.

Por favor trate de contestar cada pregunta en el examen y la encuesta, así como lo haría en el salón de clases. Ud. puede levantar la mano si tiene alguna pregunta y el administrador del examen o un maestro podrá ayudarle.

No se calificarán las respuestas que Ud. dé, y este examen **no** será parte de sus calificaciones en la clase de matemáticas. Puede descansar en cualquier momento. Se estima que para terminar la prueba, y las preguntas de la encuesta acerca del idioma tomará aproximadamente una hora.

Cuando haya terminado el examen, **por favor recuerde contestar las 5 últimas preguntas al final del folleto.** Hay dos versiones de estas 5 preguntas, una en Español y otra en Inglés. Por favor tan solo responda a una u otra de las versiones, pero no las dos.

Los datos que se coleccionarán de las respuestas serán utilizados solamente con el propósito de estudiar las estadísticas correspondientes. Los reportes finales para este estudio resumirán las conclusiones de la muestra y no serán asociados con un distrito o un individuo. No daremos a cualquier persona que no sea parte del equipo del estudio información que lo identifique a Ud., o a su distrito, al menos que seamos obligados por la ley.

Si tiene Ud. algún comentario o sugerencia para mejorar esta forma, por favor escriba al: U.S. Department of Education, Washington, DC, 20202-4651. Si tiene comentarios o alguna preocupación sobre su participación en este estudio, por favor escriba al: Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington, DC, 20208.

1. How many hours are equal to 150 minutes?

A. $1\frac{1}{2}$

B. $2\frac{1}{4}$

C. $2\frac{1}{3}$

D. $2\frac{1}{2}$

2. How much change will John get back from \$5.00 if he buys 2 notebooks that cost \$1.80 each?
- A. \$1.40
 - B. \$2.40
 - C. \$3.20
 - D. \$3.60

-
3. Fifteen boxes each containing 8 radios can be repacked in 10 larger boxes each containing how many radios?
- A. 8
 - B. 12
 - C. 80
 - D. 120

-
4. If a measurement of a rectangular box is given as 48 cubic inches, then the measurement represents the
- A. distance around the top of the box.
 - B. length of an edge of the box.
 - C. surface area of the box.
 - D. volume of the box.

5. There were 90 employees in a company last year. This year the number of employees increased by 10 percent. How many employees are in the company this year?
- A. 9 workers
 - B. 81 workers
 - C. 99 workers
 - D. 100 workers

6. Tasha is buying a toy that is regularly \$12.99 and is on sale for $\frac{1}{4}$ off. Which expression can she use to estimate the discount on the toy?
- A. $0.0025 \times \$13$
 - B. $0.04 \times \$13$
 - C. $0.25 \times \$13$
 - D. $0.40 \times \$13$

7. What is 4 hundredths written in decimal notation?

A. 0.004

B. 0.04

C. 0.400

D. 4.00

- 8.** Alba needed to know about how much the sum of 19.6, 23.8, and 38.4 is. She correctly rounded each of these numbers to the nearest whole number. What three numbers did she use?
- A.** 19, 23, 38
- B.** 19, 24, 38
- C.** 20, 24, 38
- D.** 20, 34, 39

9. Jason bought a jacket on sale for 50% off the original price and another 25% off the discounted price. If the jacket originally cost \$88, what was the final sale price that Jason paid for the jacket?
- A. \$22
- B. \$33
- C. \$44
- D. \$66

- 10.** If Jill is driving at 65 miles per hour, what is her approximate speed in kilometers per hour? (1 mile \approx 1.6 kilometers)
- A.** 16
 - B.** 41
 - C.** 104
 - D.** 173

- 11.** A certain reference file contains approximately one billion facts. About how many millions is that?
- A.** 1,000,000
 - B.** 100,000
 - C.** 10,000
 - D.** 1,000

- 12.** A car odometer registered 41,256.9 miles when a highway sign warned of a detour 1,200 feet ahead. What will the odometer read when the car reaches the detour? (5,280 feet = 1 mile)
- A.** 42,456.9
- B.** 41,261.3
- C.** 41,259.2
- D.** 41,257.1

- 13.** A calculator that is regularly priced \$20 is on sale for 40% off. What is the sale price of the calculator?
- A.** \$8
 - B.** \$12
 - C.** \$15
 - D.** \$16

- 14.** The mean distance from Venus to the Sun is 1.08×10^8 kilometers. Which of the following quantities is equal to this distance?
- A.** 10,800,000 kilometers
 - B.** 108,000,000 kilometers
 - C.** 1,080,000,000 kilometers
 - D.** 10,800,000,000 kilometers

15. If the values of the expressions below are plotted on a number line, which expression would be closest to five?

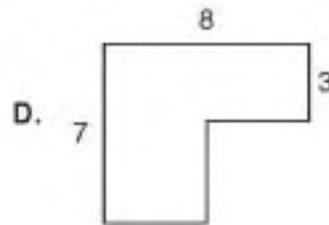
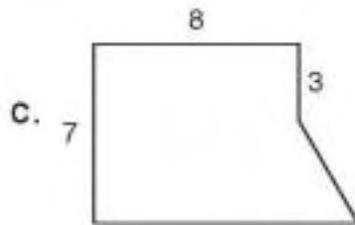
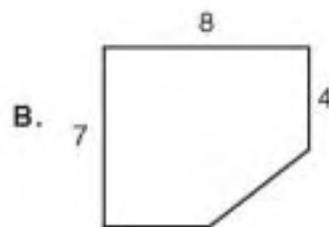
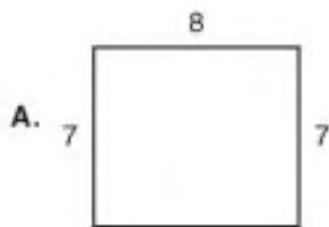
A. $|-4|$

B. $|-18|$

C. $|7|$

D. $|16|$

16. For each figure below, the lengths of 3 sides are given. Which figure could have a perimeter of 28?



17. A sweater originally cost \$37.50. Last week, Moesha bought it at 20% off.



How much was deducted from the original price?

- A. \$7.50
- B. \$17.50
- C. \$20.00
- D. \$30.00

18. Jim has $\frac{3}{4}$ of a yard of string which he wishes to divide into pieces, each $\frac{1}{8}$ of a yard long. How many pieces will he have?
- A. 3
- B. 4
- C. 6
- D. 8

19. If the price of a can of beans is raised from 50 cents to 60 cents, what is the percent increase in the price?
- A. 83.3%
 - B. 20%
 - C. 16.7%
 - D. 10%

- 20.** A landscaper estimates that landscaping a new park will take 1 person 48 hours. If 4 people work on the job and they each work 6-hour days, how many days are needed to complete the job?
- A.** 2 days
 - B.** 4 days
 - C.** 6 days
 - D.** 8 days

-
- 21.** Consider the statement “If n is an even number, then n is two times an odd number.” For which of the following values of n is the statement FALSE?
- A.** 2
 - B.** 6
 - C.** 8
 - D.** 10

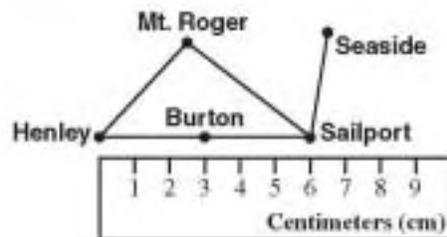
22. Ms. Thierry and 3 friends ate dinner at a restaurant. The bill was \$67. In addition, they left a \$13 tip. Approximately what percent of the total bill did they leave as a tip?
- A. 10%
 - B. 13%
 - C. 15%
 - D. 20%

23. Which piece of information is NOT needed to solve the problem below? **You do not have to solve the problem.**

Carlos is planning to buy food for his two dogs. The food he buys must last for 4 weeks. Each dog eats 1 can of dog food and 3 dog biscuits every day. How many cans of dog food does Carlos need to buy?

- A. Carlos has 2 dogs.
- B. The food must last 4 weeks.
- C. Each dog eats 1 can of dog food every day.
- D. Each dog eats 3 biscuits every day.

24. Javier is using a ruler and a map to measure the distance from Henley to Sailport.



The actual distance from Henley to Sailport is 120 kilometers (km). What scale was used to create the map?

- A. 1 cm = 6 km
- B. 1 cm = 12 km
- C. 1 cm = 15 km
- D. 1 cm = 20 km

25. Stuart is buying a pair of jeans that regularly cost \$40. They are on sale for 20% off. If the tax rate is 8%, what is the sale price of the jeans, including tax?
- A. \$21.60
 - B. \$34.56
 - C. \$42.34
 - D. \$44.16

**PLEASE GO ON TO
THE NEXT PAGE**





Student Language Background Survey: English Version

We estimate that it will take you about 5 minutes to complete this survey. Remember, there are no right or wrong answers, and you will not be graded on this task. Your participation is voluntary and you may refuse to answer any question.

Please place a CHECK ✓ in the box that applies.

1. I am in 7th Grade. or I am in 8th Grade.
2. I am Male. or I am Female.

Please place a CHECK ✓ in all boxes that apply.

3. I attended these grades in the United States:

- Kindergarten 1st Grade 2nd Grade 3rd Grade 4th Grade
 5th Grade 6th Grade 7th Grade 8th Grade

4. Did you ever go to school in another country? Yes or No

IF YES, please write the name of that country on this line: _____.

IF YES, please CHECK ✓ all of the grades you attended in that country:

- Kindergarten 1st Grade 2nd Grade 3rd Grade 4th Grade
 5th Grade 6th Grade 7th Grade 8th Grade

5. We speak these languages in my home:

- English Spanish
 Other (please write the name of that language) _____.

Thank you for your time!

Your answers to these questions about your language background will be used as part of a research study about testing accommodations sponsored by the U.S. Department of Education and carried out by the Regional Educational Laboratory West at WestEd. If you have questions about the study or this survey, please contact Edynn Sato at (415) 615-3226 or at esato@wested.org.

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless such collection displays a valid Office of Management and Budget (OMB) control number. The valid OMB control number for this information collection is 1850-0849. The time required to complete this information collection is estimated to average 5 minutes per response, including the time to review instructions, search existing data resources, gather the data needed, and complete and review the information collection. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Education, Washington, D.C. 20202-4651. If you have comments or concerns regarding the status of your individual submission of this form, write directly to: Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington, D.C. 20208.

In accordance with The Education Sciences Reform Act of 2002, Title I, Part E, Section 183, responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.



Encuesta de Antecedentes de Idioma – Versión en Español

Anticipamos que tomarán 5 minutos para completar esta encuesta. No hay respuestas correctas o incorrectas y no se darán calificaciones por ello. Su participación es voluntaria y puede decidir no contestar cualquier pregunta.

Marque con un ✓ lo que le corresponde a Ud.

1. Estoy en el Grado 7 o Estoy en el Grado 8
2. Yo soy un hombre o Yo soy una mujer

En las siguientes preguntas, marque con un ✓ todas los grados de clase que le corresponden:

3. Los grados de clase que he atendido en los Estados Unidos son:
- Kindergarten 1^o Grado 2^o Grado 3^o Grado 4^o Grado
- 5^o Grado 6^o Grado 7^o Grado 8^o Grado

4. Los grados de clase que he atendido en otro país (o países) son:
- Kindergarten 1^o Grado 2^o Grado 3^o Grado 4^o Grado
- 5^o Grado 6^o Grado 7^o Grado 8^o Grado

(Nombre del país o países)

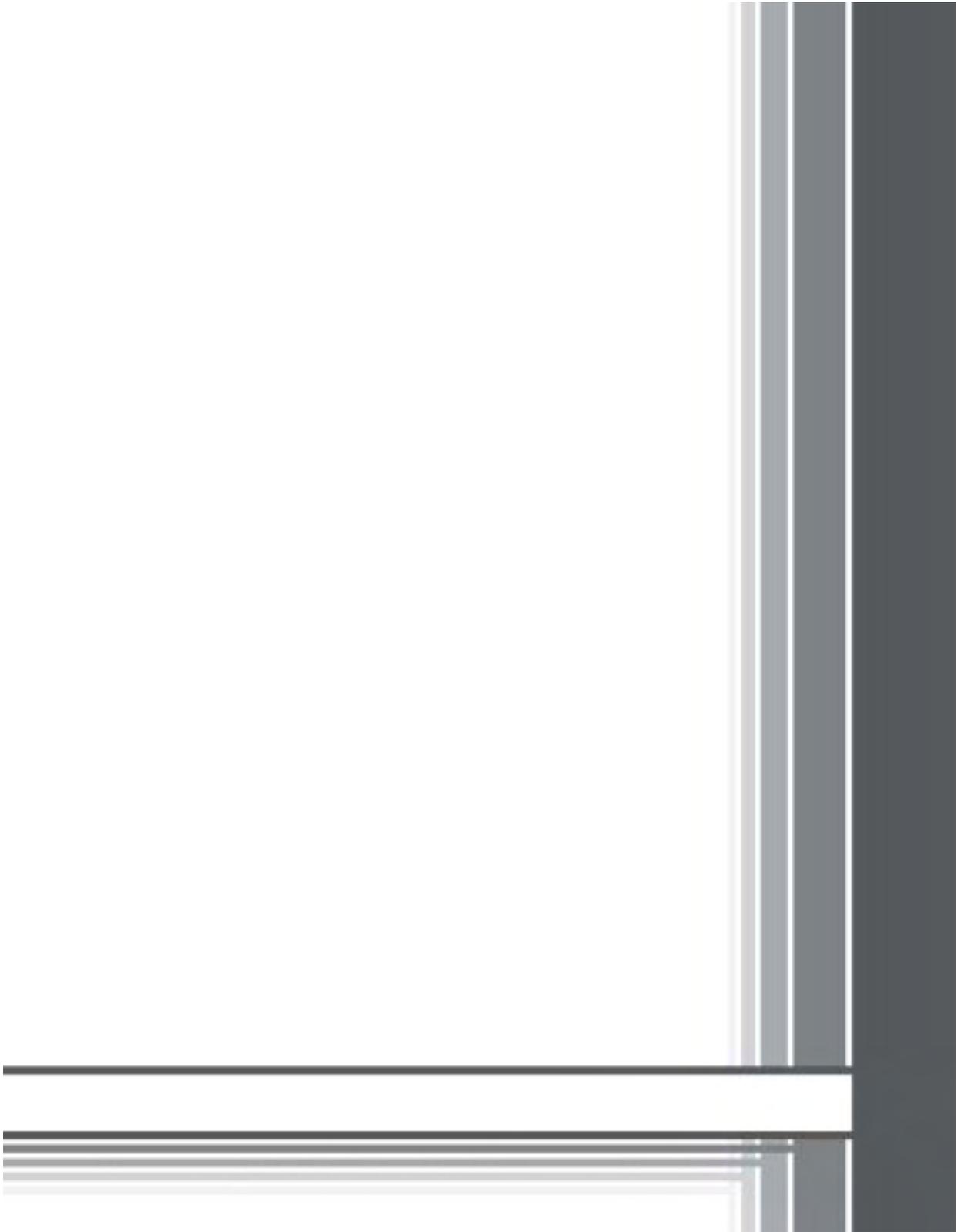
5. Marque con un ✓ los idiomas que se hablan en su casa:
- Inglés Español Otro(s) idioma(s): _____

¡Gracias por tu tiempo!

Tus respuestas a estas preguntas acerca de tu idioma serán parte de un estudio que revisará adaptaciones patrocinados por el Ministerio de Educación de los Estados Unidos y conducidos por el Laboratorio de Educación Regional del Oeste (REL West) en WestEd. Si tienes alguna pregunta acerca del estudio o de esta encuesta, por favor comunícate con Edynn Sato al número telefónico (415) 615-3226 o por correo electrónico a esato@wested.org.

Ninguna persona es requerida a contestar a la colección de información sin tener un número válido de control de OMD de acuerdo al Acto Paperwork Reduction del 1995. El número válido de control de OMD para coleccionar esta información es 1850-0849. El tiempo requerido para terminar esta colección de información se estima que es un promedio de 5 minutos. Esto incluye el tiempo para repasar las instrucciones, buscar datos existentes, obtener los datos necesarios, y terminar y repasar la información que se ha coleccionado. Si tienes comentarios sobre la exactitud del tiempo que se estima para terminar la forma o sugerencias para mejorar esta forma, por favor escríbe al: Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington D.C. 20208.

De acuerdo al Acto de Educación Sciences Reform del 2002, Título I, Parte E, Sec. 183, respuestas de esta colección de datos serán utilizados solamente para el propósito de estadísticas. Los reportes preparados para este estudio resumirán las conclusiones de la muestra y no serán asociadas con un distrito o individuo. No daremos información que lo identifique a usted o su distrito a cualquier persona que no sea parte del equipo del estudio, a menos que sea obligado por la ley.



Research Study

OPERATIONAL TEST FORM-M

Math Test
Grades 7&8

2008

Student Name: _____

 **REL**
WEST

Thank you for your participation!

This math test is part of a research project that is studying the ways in which students read and solve problems on different types of tests. Your answers will help us create test items that are fair for all students.

Please try to answer each item on this test just as you would in your classroom. You may raise your hand if you have any questions and a test administrator or teacher will come and assist you.

You will **not** be graded on your answers and this test will **not** be used as part of your grade in your math classroom. You may rest at any time. We estimate that it will take you about 45 minutes in total to complete the test items and answer the questions at the end.

When you have finished all of the test items, **please remember to answer the 5 questions at the end of the test booklet.** You will see that we have provided these 5 questions both in English and in Spanish. Please answer each question only once, either on the English version *or* on the Spanish version.

Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

If you have any comments or suggestions for improving this form, please write to: U.S. Department of Education, Washington, DC, 20202-4651. If you have comments or concerns about your participation in this study, please write to Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington, DC, 20208.

¡Gracias por su participación!

Este examen de matemáticas es parte de un estudio para entender mejor cómo estudiantes leen y resuelven problemas de diferentes tipos de exámenes. Sus respuestas nos ayudarán a crear exámenes justos para todos los estudiantes.

Por favor trate de contestar cada pregunta en el examen y la encuesta, así como lo haría en el salón de clases. Ud. puede levantar la mano si tiene alguna pregunta y el administrador del examen o un maestro podrá ayudarle.

No se calificarán las respuestas que Ud. dé, y este examen **no** será parte de sus calificaciones en la clase de matemáticas. Puede descansar en cualquier momento. Se estima que para terminar la prueba, y las preguntas de la encuesta acerca del idioma tomará aproximadamente una hora.

Cuando haya terminado el examen, **por favor recuerde contestar las 5 últimas preguntas al final del folleto.** Hay dos versiones de estas 5 preguntas, una en Español y otra en Inglés. Por favor tan solo responda a una u otra de las versiones, pero no las dos.

Los datos que se coleccionarán de las respuestas serán utilizados solamente con el propósito de estudiar las estadísticas correspondientes. Los reportes finales para este estudio resumirán las conclusiones de la muestra y no serán asociados con un distrito o un individuo. No daremos a cualquier persona que no sea parte del equipo del estudio información que lo identifique a Ud., o a su distrito, al menos que seamos obligados por la ley.

Si tiene Ud. algún comentario o sugerencia para mejorar esta forma, por favor escriba al: U.S. Department of Education, Washington, DC, 20202-4651. Si tiene comentarios o alguna preocupación sobre su participación en este estudio, por favor escriba al: Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington, DC, 20208.

1. ____ hours = 150 minutes

A. $1\frac{1}{2}$

B. $2\frac{1}{4}$

C. $2\frac{1}{3}$

D. $2\frac{1}{2}$

2. John buys 2 notebooks. Each notebook costs \$1.80. John pays with \$5.00. How much change should he get back?
- A. \$1.40
 - B. \$2.40
 - C. \$3.20
 - D. \$3.60

3. A student works in a store.

- She unpacks 15 boxes.
- Each box contains 8 radios.
- She repacks the radios in 10 larger boxes.
- Each box contains the same number of radios.

How many radios are in each larger box?

- A. 8
- B. 12
- C. 80
- D. 120

4. A rectangular box measures 48 cubic inches. This measurement represents the
- A. distance around the top of the box.
 - B. length of an edge of the box.
 - C. surface area of the box.
 - D. volume of the box.

-
5. A company has 90 workers. Next year, the number of workers will be increased by 10%. How many workers will the company have next year?
- A. 9 workers
 - B. 81 workers
 - C. 99 workers
 - D. 100 workers

6. A student buys a toy on sale.

- The regular price is \$12.99.
- The discount is $\frac{1}{4}$ of the regular price.

Which expression can he use to estimate the amount of the discount?

- A. $0.0025 \times \$13$
- B. $0.04 \times \$13$
- C. $0.25 \times \$13$
- D. $0.40 \times \$13$

7. 4 hundredths = _____

A. 0.004

B. 0.04

C. 0.400

D. 4.00

8. Which list shows the numbers 19.6, 23.8, and 38.4 rounded to the nearest whole number?

A. 19, 23, 38

B. 19, 24, 38

C. 20, 24, 38

D. 20, 34, 39

9. A student buys a jacket on sale.

- The regular price is \$88.
- The first sale price is 50% off the regular price.
- The second sale price is 25% off the first sale price.

What is the final sale price of the jacket?

- A. \$22
- B. \$33
- C. \$44
- D. \$66

10. 65 miles per hour is about _____ kilometers per hour
(1 mile \approx 1.6 kilometers)
- A. 16
 - B. 41
 - C. 104
 - D. 173

11. How many millions is 1 billion?

A. 1,000,000

B. 100,000

C. 10,000

D. 1,000

- 12.** A car's mileage is 41,256.9 miles.
The car travels 1,200 feet to an exit.
What is the car's mileage at the exit?
(5,280 feet = 1 mile)
- A.** 42,456.9
B. 41,261.3
C. 41,259.2
D. 41,257.1

13. A student buys a calculator on sale.

- The regular price is \$20.
- The sale price is 40% off the regular price.

What is the sale price of the calculator?

- A.** \$8
- B.** \$12
- C.** \$15
- D.** \$16

14. Which distance equals 1.08×10^8 kilometers?

- A.** 10,800,000 kilometers
- B.** 108,000,000 kilometers
- C.** 1,080,000,000 kilometers
- D.** 10,800,000,000 kilometers

15. Which value is closest to five on a number line?

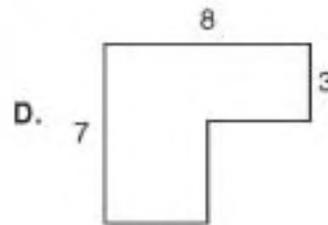
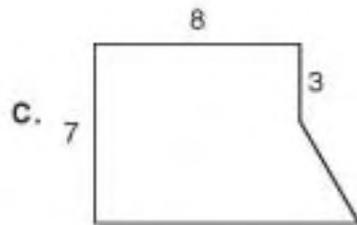
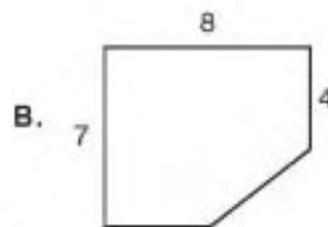
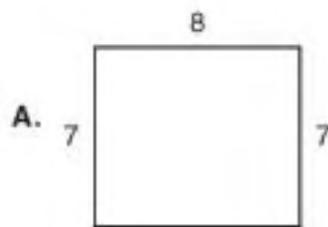
A. $|-4|$

B. $|-18|$

C. $|7|$

D. $|16|$

16. Look at the figures below.
Which figure could have a perimeter of 28?



17. A girl wants to buy a sweater on sale.
- The regular price is \$37.50.
 - The discount is 20% of the regular price.

What is the amount of the discount?

- A. \$7.50
- B. \$17.50
- C. \$20.00
- D. \$30.00

18. Jim divides $\frac{3}{4}$ yard of string into pieces. Each piece is $\frac{1}{8}$ yard long. How many pieces of string does he have?
- A. 3
 - B. 4
 - C. 6
 - D. 8

- 19.** The price of a can of beans is 50¢. The price is increased to 60¢. What is the percent increase in price?
- A** 83.3%
 - B** 20%
 - C** 16.7%
 - D** 10%

20. A manager hires students to do a job.

- She estimates that 1 student needs 48 hours to do the job.
- She hires 4 students to do the job together.
- Each student works 6 hours per day.

What is the total number of days the 4 students need to do the job?

- A.** 2 days
- B.** 4 days
- C.** 6 days
- D.** 8 days

21. Read the statement below.

If n is an even number, then n is two times an odd number.

Which value of n makes the statement FALSE?

- A.** 2
- B.** 6
- C.** 8
- D.** 10

22. A restaurant bill is \$67. The tip is \$13. Approximately what percent of the bill is the tip?
- A. 10%
 - B. 13%
 - C. 15%
 - D. 20%

23. A math problem is shown below. One piece of information is **not** needed to solve the problem.

Math Problem

A student is buying food for her dogs.

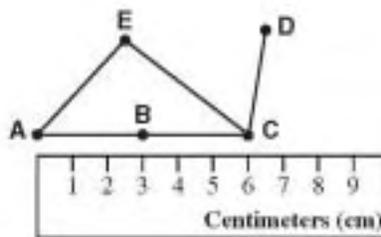
- She has 2 dogs.
- She is buying enough food for 4 weeks.
- Each dog eats 1 can of food every day.
- Each dog eats 3 biscuits every day.

How many cans of dog food does the student buy?

Which piece of information is **not** needed to solve the problem?

- A. The student has 2 dogs.
- B. The student is buying enough food for 4 weeks.
- C. Each dog eats 1 can of dog food every day.
- D. Each dog eats 3 biscuits every day.

24. Look at the map and ruler below. The diagram below shows the distance from Point A to Point C on a map.



The actual distance from Point A to Point C is 120 kilometers (km). What is the scale of the map?

- A. 1 cm = 6 km
- B. 1 cm = 12 km
- C. 1 cm = 15 km
- D. 1 cm = 20 km

25. A boy buys a pair of jeans on sale.

- The regular price is \$40.
- The discount is 20% of the regular price.
- The sales tax rate is 8%.

What is the sale price of the jeans including tax?

- A.** \$21.60
- B.** \$34.56
- C.** \$42.34
- D.** \$44.16

**PLEASE GO ON TO
THE NEXT PAGE**





Student Language Background Survey: English Version

We estimate that it will take you about 5 minutes to complete this survey. Remember, there are no right or wrong answers, and you will not be graded on this task. Your participation is voluntary and you may refuse to answer any question.

Please place a CHECK ✓ in the box that applies.

1. I am in 7th Grade. or I am in 8th Grade.
2. I am Male. or I am Female.

Please place a CHECK ✓ in all boxes that apply.

3. I attended these grades in the United States:

- Kindergarten 1st Grade 2nd Grade 3rd Grade 4th Grade
 5th Grade 6th Grade 7th Grade 8th Grade

4. Did you ever go to school in another country? Yes or No

IF YES, please write the name of that country on this line: _____.

IF YES, please CHECK ✓ all of the grades you attended in that country:

- Kindergarten 1st Grade 2nd Grade 3rd Grade 4th Grade
 5th Grade 6th Grade 7th Grade 8th Grade

5. We speak these languages in my home:

- English Spanish
 Other (please write the name of that language) _____.

Thank you for your time!

Your answers to these questions about your language background will be used as part of a research study about testing accommodations sponsored by the U.S. Department of Education and carried out by the Regional Educational Laboratory West at WestEd. If you have questions about the study or this survey, please contact Edynn Sato at (415) 615-3226 or at esato@wested.org.

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless such collection displays a valid Office of Management and Budget (OMB) control number. The valid OMB control number for this information collection is 1850-0849. The time required to complete this information collection is estimated to average 5 minutes per response, including the time to review instructions, search existing data resources, gather the data needed, and complete and review the information collection. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Education, Washington, D.C. 20202-4651. If you have comments or concerns regarding the status of your individual submission of this form, write directly to: Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington, D.C. 20208.

In accordance with The Education Sciences Reform Act of 2002, Title I, Part E, Section 183, responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.



Encuesta de Antecedentes de Idioma – Versión en Español

Anticipamos que tomarán 5 minutos para completar esta encuesta. No hay respuestas correctas o incorrectas y no se darán calificaciones por ello. Su participación es voluntaria y puede decidir no contestar cualquier pregunta.

Marque con un ✓ lo que le corresponde a Ud.

1. Estoy en el Grado 7 o Estoy en el Grado 8
2. Yo soy un hombre o Yo soy una mujer

En las siguientes preguntas, marque con un ✓ todas los grados de clase que le corresponden:

3. Los grados de clase que he atendido en los Estados Unidos son:

- Kindergarten 1^o Grado 2^o Grado 3^o Grado 4^o Grado
 5^o Grado 6^o Grado 7^o Grado 8^o Grado

4. Los grados de clase que he atendido en otro país (o países) son:

- Kindergarten 1^o Grado 2^o Grado 3^o Grado 4^o Grado
 5^o Grado 6^o Grado 7^o Grado 8^o Grado

(Nombre del país o países)

5. Marque con un ✓ los idiomas que se hablan en su casa:

- Inglés Español Otro(s) idioma(s): _____

¡Gracias por tu tiempo!

Tus respuestas a estas preguntas acerca de tu idioma serán parte de un estudio que revisará adaptaciones patrocinados por el Ministerio de Educación de los Estados Unidos y conducidos por el Laboratorio de Educación Regional del Oeste (REL West) en WestEd. Si tienes alguna pregunta acerca del estudio o de esta encuesta, por favor comunícate con Edynn Sato al número telefónico (415) 615-3226 o por correo electrónico a esato@wested.org.

Ninguna persona es requerida a contestar a la colección de información sin tener un número válido de control de OMD de acuerdo al Acto Paperwork Reduction del 1995. El número válido de control de OMD para coleccionar esta información es 1850-0849. El tiempo requerido para terminar esta colección de información se estima que es un promedio de 5 minutos. Esto incluye el tiempo para repasar las instrucciones, buscar datos existentes, obtener los datos necesarios, y terminar y repasar la información que se ha coleccionado. Si tienes comentarios sobre la exactitud del tiempo que se estima para terminar la forma o sugerencias para mejorar esta forma, por favor escríbale al: Ok-Choon Park, U.S. Department of Education, 555 New Jersey Avenue, NW, Room 506E, Washington D.C. 20208.

De acuerdo al Acto de Educación Sciences Reform del 2002, Título I, Parte E, Sec. 183, respuestas de esta colección de datos serán utilizados solamente para el propósito de estadísticas. Los reportes preparados para este estudio resumirán las conclusiones de la muestra y no serán asociadas con un distrito o individuo. No daremos información que lo identifique a usted o su distrito a cualquier persona que no sea parte del equipo del estudio, a menos que sea obligado por la ley.

