Title: Examining the Effectiveness and Validity of Glossary and Read-Aloud Accommodations for English Language Learners in a Math Assessment

Author(s): Mikyung Kim Wolf, Jinok Kim, Jenny C. Kao, and Nichole Rivera, CRESST/University of California, Los Angeles

Date of Initial Publication: 10/2009

Abstract/Summary

Glossary and reading aloud test items are often listed as allowed in many states' accommodation policies for ELL students, when taking states' large-scale mathematics assessments. However, little empirical research has been conducted on the effects of these two accommodations on ELL students' test performance. Further, no research is available to examine how students use the provided accommodations. The present study employed a randomized experimental design and a think-aloud procedure to delve into the effects of the two accommodations. A total of 605 ELL and non-ELL students from two states participated in the experimental component and a subset of 68 ELL students participated in the think-aloud component of the study. Results showed no significant effect of glossary, and mixed effects of read aloud on ELL students' performance. Read aloud was found to have a significant effect for the ELL sample in one state, but not the other. Significant interaction effects between students' prior content knowledge and accommodations were found, suggesting the given accommodation was effective for the students who had acquired content knowledge. During the think-aloud analysis, students did not actively utilize the provided glossary, indicating lack of familiarity with the accommodation. Implications for the effective use of accommodations and future research agendas are discussed.

Subject/Keywords: ELL, ELL Accommodations, Read aloud accommodations

Document Type: Paper

Document Archive Number: 0019cdewo2009

Examining the Effectiveness and Validity of Glossary and Read-Aloud Accommodations for English Language Learners in a Math Assessment

Final Deliverable - October 2009

Mikyung Kim Wolf, Jinok Kim, Jenny C. Kao, and Nichole Rivera
CRESST/University of California, Los Angeles

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the National Research and Development Centers, PR/Award Number R305A050004, as administered by the U.S. Department of Education, Institute of Education Sciences.

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

ACKNOWLEDGEMENTS

We would like to thank the following people for their valuable contributions to this study in many ways:

Joan Herman and Noelle Griffin for their guidance and invaluable feedback throughout the entire process of this study.

Sandy Chang, for her generous contributions to instrument development, data collection, transcription, and data analysis.

Patina Bachman, Julie Nollner, Hyewon Shin, and Tim Farnsworth, for their help with instrument development and data collection.

Rita Pope, for her help with data collection and leading the math test alignment study.

Carol Ann Ramirez, Socorro Shiels, and Lisa Sullivan, for taking part in the alignment study.

Jean Jho and Stella Tsang Li, for their help with data entry and transcription.

Kim Hurst and Haig Santourian, for administrative support, including scanning answer sheets and processing payments.

Robert Kaplinsky and Belinda Thompson, for providing feedback on our test and glossary based on their experiences as middle school math teachers.

Bruin Partners/Marina Del Rey Middle School students, for allowing us to test run our think aloud instruments.

This study relied on the cooperation of many people from the two participating states, for which we are indebted: State Title I and Title III department directors and staff, district English language acquisition units, the principals from all 13 schools for allowing us to come in to their schools, as well as the teachers for coordinating logistics and proctoring the sessions, and finally, the students, for which this study was conducted. We are deeply grateful for their participation.

TABLE OF CONTENTS

EXAMINING THE EFFECTIVENESS AND VALIDITY OF GLOSSARY AND READ-ALOUD ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNERS IN A MATH ASSESSMENT

Mikyung Kim Wolf, Jinok Kim, Jenny C. Kao, & Nichole Rivera
CRESST/University of California, Los Angeles

Abstract

Glossary and reading aloud test items are often listed as allowed in many states' accommodation policies for ELL students, when taking states' large-scale mathematics assessments. However, little empirical research has been conducted on the effects of these two accommodations on ELL students' test performance. Further, no research is available to examine how students use the provided accommodations. The present study employed a randomized experimental design and a think-aloud procedure to delve into the effects of the two accommodations. A total of 605 ELL and non-ELL students from two states participated in the experimental component and a subset of 68 ELL students participated in the think-aloud component of the study. Results showed no significant effect of glossary, and mixed effects of read aloud on ELL students' performance. Read aloud was found to have a significant effect for the ELL sample in one state, but not the other. Significant interaction effects between students' prior content knowledge and accommodations were found, suggesting the given accommodation was effective for the students who had acquired content knowledge. During the think-aloud analysis, students did not actively utilize the provided glossary, indicating lack of familiarity with the accommodation. Implications for the effective use of accommodations and future research agendas are discussed.

Introduction

Since federal legislation mandated the participation of all students, including English language learner (ELL) students, in state accountability systems, the validity of ELL assessment has gained much attention. It is of particular concern in the field to ensure that the states' high-stakes, large-scale content assessments adequately measure ELL students' content knowledge and skills, without unduly penalizing the students who are still learning English. For instance, a math assessment is, broadly speaking, intended to measure a student's mathematical problem-solving ability. However, linguistic complexities in the math assessment may interfere with ELL students' mathematical problem-solving ability, failing to measure the intended construct for these students. Testing accommodations have been utilized as a way of reducing these types of unintended factors, referred to as construct-

irrelevant variance, so that one can adequately assess ELL students' content knowledge and make appropriate inferences from the assessment results.

For the past decade, a body of research has focused on investigating the effectiveness and validity of accommodations and on providing guidance on the appropriate use of accommodations for ELL students (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Sireci, Li, & Scarpati, 2003). However, previous empirical studies on the effects of accommodations yielded mixed results. Francis et al.'s meta-analysis study indicated that the accommodation effects varied depending on grades, content areas, and assessment types. Among the seven accommodation types in the studies included in their meta-analysis (simplified English, English dictionary/glossary, bilingual dictionary/glossary, extra time, Spanish language test, dual language questions, dual language booklet), only English language dictionary/glossary accommodation was found to have an overall positive effect on increasing ELL students' performance. Inarguably, continuous investigation of an effective accommodation use is warranted in order to provide research-based accommodation guidance for practitioners.

In light of this, the purpose of the present study is to examine the effectiveness and validity of accommodations which are commonly provided to ELL students when taking a large-scale content assessment. This study focused on two particular accommodations for states' standards-based math assessments at Grade 8: English glossary and reading aloud an entire test in English. These two accommodations were selected for a number of reasons. First, these accommodations are assumed to help ELL students because they directly support the students' language limitations. Rivera, Collum, Shafer Willner, and Sia, (2006) shifted the previous accommodation paradigm based on students with disabilities into a new taxonomy for ELL students by grouping accommodations into "direct linguistic support" and "indirect linguistic support" accommodations. While a number of researchers advocate providing accommodations that are responsive to ELL students' specific needs, that is, their limited English language proficiency, few empirical studies are available to prove the effect of direct linguistic support accommodations. Secondly, amongst the direct linguistic support types of accommodations, these two were identified as the most frequently allowed accommodations in states' policies. In the 2006-2007 school year, 43 states allowed a type of vocabulary-support accommodation (e.g., dictionary, glossary) and 39 states allowed reading aloud of test items for ELL students in taking states' large-scale standardized assessments (Wolf, Kao, et al., 2008). However, as noted above, little empirical research-based evidence is available to support the use of these accommodations. Although English dictionary/glossary was identified as the only effective accommodation for ELL students, it is worth revisiting to confirm the previous finding. An investigation of the validity of these

common accommodations will also provide useful information to many policymakers and practitioners who allow these accommodations.

This study focuses on mathematics content area and Grade 8 to examine the accommodation effects. As previous studies indicated, allowable accommodations should depend on the content areas, and consider the construct of an assessment. For example, a dictionary may not be allowed for a reading assessment because students may receive unfair advantages by having access to vocabulary words being tested. Given that a state's mathematics assessment is typically intended to measure mathematical knowledge and skills, not language proficiency, providing glossary and read-aloud accommodations to help ELL students' language difficulty to solve math problems seems reasonable. Grade 8 was chosen because of the practical impact of the study findings. In all states, eighth-grade assessment results are counted for the Adequate Yearly Progress (AYP) reporting purposes. An appropriate assessment of ELL students at Grade 8 is thus critical in all states. Additionally, it was expected that the Grade 8 students would be more capable of using a given accommodation compared to students in lower grades.

Specifically, this study posits the following research questions:

1. Does providing glossary and read-aloud accommodations increase ELL students' performance in a math assessment as compared to the standard testing condition?
2. Does providing glossary and read-aloud accommodations leave non-ELL students' performance unchanged, as compared to the standard testing condition?
3. To what extent do ELL students have difficulty with the language and content in solving math items?
4. How do ELL students utilize a glossary accommodation?
5. What are students' perceptions on the helpfulness of glossary and read-aloud accommodations when taking a math assessment?

This study is part of a large-scale research project dealing with a broad range of ELL assessment and accommodation issues with the purpose of providing practical recommendations for policymakers and practitioners to improve the validity of their ELL assessment systems. As a subset study undertaken in the larger research project, the purpose of the current study is not only to investigate the validity of commonly-allowed accommodations in a state's large-scale math assessment, but also to offer useful guidelines on improving the validity of accommodation practices drawn from the findings. In this study, we will refer to other research conducted under the larger research project, such as research on the policies and practices of accommodation uses (Wolf, Griffin, Kao, Chang, & Rivera, 2009), in order to better understand the results of the present study.

Relevant Literature

In this section, we briefly review relevant literature to provide contextual issues that lead to this study. We first review the literature addressing the need of accommodations for ELL students to take a mathematics assessment. We also review previous studies that examine the validity and effectiveness of read aloud and glossary accommodations.

The Language Demands in Math Assessments for ELL Students

ELL students have historically underperformed in mathematics when compared to their non-ELL peers. As reported by the U.S. Government Accountability Office, the math proficiency level of ELL students across 48 states during the 2003-2004 school year was 20% lower, on average, than the overall population (U.S. GAO, 2006). In the 2007 National Assessment of Education Progress (NAEP) in mathematics, 70% of ELL students in Grade 8 scored Below Basic as compared to 27% of non-ELL students (Lee, Grigg, & Dion, 2007). While many issues can partly explain the large achievement gap, such as opportunity to learn, one can speculate that ELL students' lack of language proficiency and the language characteristics in such math assessments may play some underlying role in the gap.

Past research linked language with mathematics problem solving (Aiken, 1971, 1972; Cummins, Kintsch, Reusser, & Weimer, 1988; De Corte, Verschaffel, & DeWin, 1985). For ELL students in particular, language demands may interfere with their ability to perform on a math assessment (Abedi & Lord, 2001). Past studies also found that linguistic features of math problems can interfere with ELL students' ability to solve the problems (Spanos, Rhodes, Dale, & Crandall, 1988). Abedi (2006) contended that unnecessary linguistic complexities of test items are nuisance variables that confound assessment outcomes. In work by Abedi and colleagues, test items that were modified to reduce the linguistic complexity of non-content in both math and science items were found to increase the performance of ELL students (Abedi, Courtney, & Leon, 2003a; Abedi, Lord, & Hofstetter, 1998). Garcia (1991) found that unknown vocabulary in particular affected ELL students' performance on reading assessments. Furthermore, guides on mathematics instruction for ELL students continue to emphasize the need for building students' vocabulary (Coggins, Kravin, Coates, & Carroll, 2007; Dale & Cuevas, 1987; Rubenstein, 1996). Recently, Wolf, Herman, et al. (2008) examined the language characteristics of three states' mathematics and science assessments. Their study found that some states' mathematics assessments presented comparable language demands to those in science assessments particularly with the amount of academic vocabulary. In the subsequent study, Wolf and Leon (2009) provided empirical evidence of language demands on ELL students' test performance by investigating the

language characteristics of items differentially functioning against ELL students. In their study, mathematics and science items disfavoring ELL students tended to contain more academic vocabulary and be lengthy with little visual cues (e.g., graphics, charts, tables).

Accommodations

Accommodations generally are changes to a test or changes to the way a test is administered. They are intended to help ELL students overcome language barriers when taking an assessment, as well as reduce threats to test score validity. Accommodations can provide ELL students with either direct or indirect linguistic support (Rivera et al., 2006). Some examples of accommodations that provide direct linguistic support include providing bilingual dictionaries, providing native language translations of directions, or reading aloud test items in English. Examples of accommodations that provide indirect linguistic support include providing extended time, administering the test in a small group, or administering the test in a separate location.

Research on accommodations has emphasized the importance of accommodations being both effective and valid (Abedi et al., 2003a; Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005). Accommodations that are effective increase the scores of ELL students and reduce the achievement gap between ELL and non-ELL students. Accommodations that are valid do not affect the scores of non-ELL students. This is also referred to as the "interaction hypothesis" which states that students who need a particular type of accommodation would benefit from it, and those who do not, would not benefit from it (Koenig & Bachman, 2004; Sireci et al., 2003). If this is found to be true for a particular accommodation, then the accommodated test results can be aggregated with the non-accommodated results. An invalid accommodation means it gives an unfair advantage to those receiving it over those not receiving it, which means that test results from an invalid accommodation could be inflated (Sireci et al.).

Using an experimental design including a random assignment of accommodations to both ELL and non-ELL students, as illustrated by Koenig and Bachman (2004), the interaction hypothesis can be tested. In this design, one can test whether the given accommodation has influence on ELL students, but not on non-ELL students. However, even if shown to be effective and valid in an experimental study, ELL students are a heterogeneous group and researchers have cautioned against a "one size fits all" approach to providing accommodations (Abedi, Hofstetter, & Lord, 2004). More recent accommodation research found that ELL students receiving appropriate test accommodations scored higher on a math test than students receiving no accommodations or not-recommended

accommodations (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007). Kopriva et al. suggested that future research using control and treatment groups should consider specific student needs before making direct comparisons between groups. Similarly, Ketterlin-Geller, Yovanoff, and Tindal (2007) emphasized the need to consider the interaction between item features (i.e., language complexity) and student characteristics (i.e., personal attributes) in accommodations research.

Below we summarize previous research on the two accommodations focused in the present study, glossaries/dictionaries and read aloud.

Glossaries/Dictionaries. Glossaries and dictionaries are provided to ELL students to help them understand the meaning of some words. While glossaries and dictionaries serve a similar purpose, there is a distinction between them. Generally speaking, researchers of accommodations have considered dictionaries to be reference books that contain general definitions of a word, and are usually commercially published. Glossaries, however, provide an explanation of a word customized for a particular context and audience (Rivera et al., 2006). Both glossaries and dictionaries can be English only, or bilingual (English to students' native language). However, as Rivera et al. noted, there is no identifiable standard distinguishing the two terms in the literature. In a series of CRESST studies conducted by Abedi and colleagues, variations of dictionary, customized dictionary, glossary, and "pop-up" glossary were investigated for math and science assessments (Abedi et al., 2003a; Abedi, Courtney, & Leon, 2003b; Abedi et al., 2005; Abedi, Hofstetter, Baker, & Lord, 2001; Abedi, Lord, Boscardin, & Miyoshi, 2001). Abedi et al. (2003a) provided "customized English dictionaries" and glossaries of non-content terms as supplemental handouts in science assessments to Grade 4 and Grade 8 students. Both ELL and non-ELL students received either the customized English dictionary, or a glossary, or another accommodation, or no accommodation. ELL students of Spanish-speaking backgrounds were provided with English-to-Spanish glossaries, and non-ELL students were provided with English-to-English glossaries. (Students in the standard condition were also provided a supplemental handout, containing a list of words from the assessment, but with no definitions). No significant results were found with glossary or dictionary, and no impact on test scores was found on non-ELL students. Abedi et al. (2003b) provided "pop-up" glossaries of non-content terms using a computer administration in a mathematics assessment to Grade 4 and Grade 8 students. Results indicated that the "pop-up" glossary was effective in increasing the performance of both Grade 4 and Grade 8 ELL students, but also did not affect the scores of non-ELL students. This study also investigated a customized English dictionary, administered through a traditional paper assessment, but no significant results were found. In another study

involving Grade 8 math assessments, ELL students benefited most from receiving an English glossary of non-technical terms, plus extra time (Abedi et al., 2001). However, non-ELL students' test scores also increased with glossary plus extra time.

Read Aloud/Oral Administration. Prior research on read aloud, or oral administration of test items, has focused on students with disabilities and not English language learners (for example, Bolt & Ysseldyke, 2006; Elbaum, 2007; Kim, Schneider, & Siskind, 2009; Weston, 2003). For instance, Bolt and Ysseldyke found that the read-aloud accommodation was associated with greater measurement problems on a reading/language arts test than on a math test for students with disabilities. Weston's study included both learning disabled and "regular classroom" fourth-grade students who took two matched forms of a mathematics assessment, one standard and one read aloud, and included interviews with a sample of the students in a group discussion format. Both learning disabled and regular classroom students overwhelmingly reported preferring the standard "paper and pencil" test over the read aloud. Students felt that the test "took too much time" and one regular classroom student disliked the read aloud because "they won't let you go ahead" (Weston, 2003). There is a dearth of research focusing on read aloud/oral administration accommodation specifically for the ELL population. One study focusing on ELL students investigated oral presentation of test directions only, not test items (Hafner, 2001). Hafner randomly assigned Grade 4 students to one of three testing conditions for a math test: extra time, standard, and extra time plus oral presentation of directions. The oral presentation of directions included simplifying directions, re-reading directions, providing additional examples, or reading directions in students' native language. Results indicated that the non-ELL students benefited the most from the accommodations. In a study of reading tests, Grade 8 ELL students from Spanish-speaking backgrounds were provided with dual-language test items (items printed in both English and Spanish) as well as the option of listening to the item read aloud in Spanish with an audiocassette (Anderson, Liu, Swierzbin, Thurlow, & Bielinski, 2000). However, results on the accommodated test were not significant, and the majority of students reported not using the option of read aloud. In a study of Grade 3 students in mathematics, with 18% special education and 3% ELL, Ketterlin-Geller et al. (2007) found that students with lower reading abilities benefitted the most from a read-aloud accommodation for test items with high mathematics difficulty and high language complexity, but not for items with low mathematics difficulty and high language complexity, or either high or low math difficulty and low language complexity. These results suggested that the read-aloud accommodation was only beneficial when the language of the test items was complicated enough to interfere with students' ability to access difficult content.

As reviewed, the effectiveness and validity of glossary and read-aloud accommodations for ELL students need further investigation. In the following section, we will describe our methodological approach to investigating this issue.

Method

In order to investigate our research questions described earlier, we utilized both quantitative and qualitative methods. Quantitatively, a randomized experimental design was applied to find the effects of accommodations on ELL and non-ELL students. Detailed quantitative analytic techniques are described below. Qualitatively, a think-aloud and retrospective interview were used to conduct students' verbal protocol analysis. This analysis aimed to closely examine the use of the two accommodations and the problem-solving processes of ELL students.

Participants

A total of 605 students from the two states participated in this study (313 ELL, and 292 non-ELL). We henceforth refer to two states as State X and State Y, respectively, to preserve anonymity. The two states were selected for this study largely due to their interest in collaborating with the researchers on issues related to ELL accommodations. The proportion of ELL students in these two states, in relation to non-ELL students in public schools, are roughly consistent with the nationwide average. These two states are also amongst the states with the fastest and largest ELL growth. All participation was on a voluntary basis, and all necessary consent forms were collected from parents and students. The schools were selected based on state recommendation, and then district approval followed by principal approval.

In State X, 267 Grade 8 students (140 ELL, and 127 non-ELL) from four schools in one urban school district participated in the testing. Of the ELL students, 19 students also participated in the think-aloud interview. The Grade 8 ELL students in these schools comprised roughly 19%, on average, of all Grade 8 students, which is higher than the district average of 11%. In State Y, 338 Grade 9 students (173 ELL, and 165 non-ELL) from nine schools in four school districts (three suburban and one urban) participated in the testing. Of the ELL students, 49 also participated in the think-aloud interview. In these four school districts, the proportion of eighth-grade ELL students averaged between 6% to 14%. (Exact percentages for ninth grade were not available, but typically, higher grades have lower proportions of ELL students due to reclassification). ELL students included in this study were largely from Spanish-speaking backgrounds.

Since the math test was designed to measure Grade 8 standards, students at the end of Grade 8 (for State X) or beginning of Grade 9 (for State Y) were targeted for the sample. The

data collection was conducted first in State X in Spring 2008 with Grade 8 students who had just completed state standards-based assessments. State Y data collection occurred in Fall 2008. In order to obtain comparable data between the two states, Grade 9 students were recruited in State Y. It was expected that Grade 9 students were a more appropriate sample than the then-current eighth-graders because the mathematics assessment of the experimental design contained the entire Grade 8 standards.

Additionally, both states provided background information on the students (e.g., gender, race/ethnicity, language proficiency levels, home language, free or reduced lunch program eligibility, IEP status, ELL status, ESL program participation) as well as 2008 standardized test scores for reading and math (raw, scale, and percentile rankings), as well as scores from 2008 English language proficiency tests.

After reviewing students' background information, some of the students were recategorized into "Former ELLs," which means we considered them as former ELL students exited from ELL services (more detail in the Results section will follow). The total number of participants in the math test is displayed in Table 1 below by condition, state, and ELL status.

Table 1

Participants by Condition, State, and ELL Status

		Condition			
	Status	Standard	Glossary	Read Aloud	Total
State X	ELL	44	36	37	117
	Non-ELL	42	48	37	127
	Former ELL	6	9	8	23
	Total	92	93	82	267
State Y	ELL	43	43	52	138
	Non-ELL	51	55	59	165
	Former ELL	15	12	8	35
	Total	109	110	119	338
Total		201	203	201	605

Note. Former ELL refers to recategorized ELL students who, after meeting certain language proficiency conditions, were exited from ELL services.

For the think-aloud interview component of the study, between one to eight students from each of the 13 schools participated in the think-aloud interview for a total of 68 students

(38 female, and 30 male). For the students' verbal protocol analysis, both current ELL and former ELL students were included. Thirteen of the students were considered former ELL: eleven were still under the two-year monitoring period, and two had been exited for over two years. The most common language other than English spoken by students participating in the think-aloud interview was Spanish (55 students, or 80.9%). Other languages included: Arabic, Bengali, Danish, Hmong, Mandarin Chinese, Punjabi, Tagalog, Turkish, and Vietnamese. The most frequent country of birth was Mexico (34, or 50.0%), followed by the U.S. (22, or 32.4%). Other countries of birth included: Bangladesh, China, Costa Rica, Denmark, El Salvador, Honduras, India, Philippines, Turkey, and Vietnam. The most frequent U.S. school start grade level was kindergarten or earlier (28, or 41.2%). The remaining students started school in the U.S. between Grades 1 through 8, with an average of 4.65.

Instruments

Math Test. In order to have assessment content similar to both states' mathematics assessments, we examined their Grade 8 math standards available online as well as the states' Grade 8 math assessments from the prior year. Once assured that both states' assessments covered common Grade 8 math standards and curriculum, such as those in the Trends in International Mathematics and Science Study (TIMSS) or the National Assessment of Educational Progress (NAEP), an eighth-grade level mathematics test was developed for the study. The test comprised of 37 items (35 multiple-choice and 2 open-ended), including released items from NAEP (1990, 1992, 1996) and TIMSS (1995), as well as released items from standardized math tests from various states (Seven items from the California Standards Test, Grades 6-8, 2003; two items from State X Instructional Materials for Grade 8, 2006). Four items were selected from a previous CRESST research project on algebra (CRESST, 2006). All selected items had undergone a field test and had an acceptable level of item reliability statistics (e.g., high item-total correlation). The test items addressed math standards of number sense/computation and algebra, and also included some items covering geometry, measurement, and data analysis. A few items were slightly modified in wording to improve clarity or remove datedness issues, as well as remove extra distractor choices (when necessary), so that all multiple-choice items consisted of four response choices. The test was designed to be administered in approximately 45 minutes under the Standard test administration. Math test booklets were professionally printed into a two-sided booklet format with saddle stitching with one to three test items appearing on each page.

An alignment study, to ensure alignment with states' math content standards, was conducted by an external review team consisting of doctoral students with expertise in math education, educational psychology, and secondary-level teaching, using Webb's alignment

tool (1997)[1]. The four alignment criteria, as defined by Webb, Alt, Ely, and Vesperman (2005) were: categorical concurrence, depth of knowledge consistency, range of knowledge correspondence, and balance of representation. These criteria were examined for alignment with 2006 Grade 8 math standards for both states (the most recent standards available at the time). Results of the alignment study revealed that some standards/objectives had a high incidence of items, while other standards/objectives had a low incidence of items. To ensure adequate alignment, a few items that had high incidence in a specific objective were removed and replaced with items that had lower incidence in a specific objective.

The math test items were also examined for their linguistic complexity using a linguistic content analysis protocol developed by the researchers (See Wolf, Herman, et al., 2008). For example, the number of words, lexical density, the number of academic words, the number of academic grammatical features (e.g., passive, nominalization), form of presentation (e.g., proportion of language and non-language), reliance (i.e., the level of language knowledge required to solve an item), and visuals (i.e., the amount of language presented in visual images) were examined for every item. The results of the rating scores were compared to ones from the states' mathematics assessments, which had been conducted previously (Wolf et al.). The results were comparable in terms of the range of the rating scores as well as the mean rating scores.

Accommodations. In order to implement the read-aloud accommodation in a standardized manner, an administration script of test items was created for test proctors to read aloud verbatim to the students. The script was developed partially based on State Y's standardized math assessment script[2] and State X's general read-aloud guidelines. State X's guideline specified that numbers and symbols were not allowed to be read aloud in the mathematics assessment, while State Y's script indicated that certain numbers and symbols that were part of the construct were not allowed to be read aloud. Similar to State Y's script, the script of this study selected certain numbers or symbols related to the item construct not to be read aloud. For example, in cases where numbers and symbols were not to be read, those numbers/symbols were replaced with the word "[*pause*]" and the proctor was instructed to pause at those times. In cases where numbers were to be read, the numbers were spelled out. Figures and charts with titles or labels were also narrated in the script, to ensure

[1] For more information on Webb's alignment tool, see http://wat.wceruw.org/index.aspx.

[2] Information on the State Y math assessment script was gathered from a conference call with State Y Title III representatives on March 10, 2008. In State Y, test proctors are instructed to pause at specific content terminology as indicated by an underscore. We chose to write in the word "pause" into the script to facilitate test administration and ensure uniformity in the read aloud. State X did not have a script for its math assessment; however, the state provides the general guideline of "no numbers or mathematical symbols" are to be read aloud.

uniformity in reading across test administrators. The original test items were also printed in the script for their reference (See example of script in Appendix A).

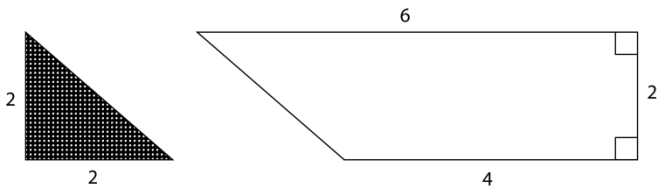
Two versions of the test booklet were created: Standard and Glossary. The Standard version was administered for both the Standard condition and the Read Aloud condition, while the Glossary version was administered only in the Glossary condition. Test items in the Glossary version appeared in the same order and same page layout as the Standard version, with the addition of an English-to-English glossary appearing in the right margin. Only non-content (i.e., non-math) terms were glossed, and glossed words appeared next to their corresponding test item in the order of appearance within the item. Some general academic vocabulary words were glossed, but not specialized or technical terms. In some cases, phrases were also glossed. Glossary definitions were based on *Longman Handy Learner's Dictionary of American English* (2000), with modifications made based on age-appropriateness and relevance to the test item. Thirty of the items contained glossed words, with about one to eight glossed words each. The Glossary version of the test was reviewed by two eighth-grade math teachers with experience teaching ELL students. Feedback was provided on both glossed words as well as test items, and alterations were made based on their feedback. (See Appendix B for a list of glossary terms and definitions).

Student Think-Aloud Test. A sample of five items were selected from the math test to elicit students' think-aloud responses. The items, which included between two to eight glossed words for each item, were reproduced into a separate booklet and used during the think-aloud process. Figure 1 displays the stems of the five think-aloud test items. Note that glossaries for each item are not shown in Figure 1. (See Appendix C for more detail on the five items, including glossaries).

With the goal of examining how students dealt with language in math items, items with different types of linguistic complexities were selected. Table 2 presents the summary of the linguistic rating. Some items were more complex than others in terms of the number of academic vocabulary, grammatical features, and cohesive devices presented. "Form" rating, which captures the amount of the language presented in relation to non-language (e.g., numbers, equations, graphs), received a score of 2 (some non-language) because they contained numerical values or equations in answer choices. Only Item 1 contained a figure in its stem, and the remaining four item stems included sentences and some numbers. On a 4-point scale, "Reliance" rating intends to measure the amount of language that test takers need to process in order to solve an item correctly. A score of 2 indicates that vocabulary knowledge is required to answer the item correctly, and a score of 3 indicates that processing the sentence structure is required in addition to vocabulary knowledge. A score of 4 indicates

that processing cohesive relationships across sentences is also required. As shown in Table 2, the five think-aloud items required test takers to process vocabulary to a high level of sentential relationship.

1.



How many triangles of the shape and size of the shaded triangle can the trapezoid above be divided into?

2. Of the following, which is the closest approximation of a 15 percent tip on a restaurant check of $24.99?

3. A group of students has a total of 29 pencils and everyone has at least one pencil. Six students have 1 pencil each, five students have 3 pencils each, and the rest of the students have 2 pencils each. How many students have only 2 pencils?

4. A group of hikers climbed from Salt Flats (elevation -55 feet) to Talon Bluff (elevation 620 feet). What is the difference in elevation between Talon Bluff and Salt Flats?

5. A plumber charges customers $48 for each hour worked plus an additional $9 for travel. If *h* represents the number of hours worked, which of the following expressions could be used to calculate the plumber's total charge in dollars?

Figure 1. Stems of the five think-aloud test items.

Table 2

Linguistic Analyses Results for the Student Think-Aloud Test

Item Number	No. of Total Words	No. of Sentences	No. of Academic Vocabulary	No. of Grammatical Features	No. of Cohesive Devices	Form	Reliance
1	19	1	3	4	0	2	3
2	17	1	3	1	0	2	2
3	33	3	2	0	7	2	4
4	26	2	2	1	0	2	3
5	36	2	5	5	3	2	4

Student Interview Protocol. A retrospective interview protocol focused on five main areas to prompt students as they completed the items on the Student Think-Aloud Test: comprehension (Does the student understand the question), problem solving (How does the student solve the problem), difficulty (What is the students' perceived difficulty of the problem), accommodation use (Did the student utilize the glossary words printed with the test items), and students' general perceptions on accommodations. Students who were part of the Read Aloud or Glossary conditions for the math test were also asked about their perceptions on the respective accommodation conditions. Prior to data collection, CRESST researchers piloted the Student Interview Protocol on local middle school students, then debriefed and made revisions to the protocol as needed.

Procedure

Data collection occurred April to May 2008 in State X, and October 2008 to January 2009 in State Y. For the math test, students were randomly assigned to one of three testing conditions: Standard, Read Aloud, and Glossary. Effort was made to ensure a roughly equal number of ELL and non-ELL students in each condition. Rooms for Read Aloud contained between 7 to 19 (average of 13.47) students each, while rooms for the other classrooms contained no more than 25 students each. Standard and Glossary conditions were sometimes administered together in the same room, while Read Aloud was always administered in a separate room. Test administration was completed in one to two class periods (approximately 50-90 minutes), depending on the condition. Teachers or school administrators from each school proctored the Read Aloud conditions. They were trained on how to use the script by CRESST researchers either in person or via telephone prior to the testing, and were provided with excerpts from the script to practice, to ensure uniformity across all schools in

administering the read aloud. One to two CRESST researchers were also present in each testing room to assist with proctoring.

The scoring process entailed electronic scanning of answer sheets for multiple-choice items. The two open-ended items were scored by two raters each, using a three-point rating scale (0-2) previously used by Abedi and colleagues (Abedi et al., 2003b; Abedi, Courtney, Leon, Kao, & Azzam, 2006). The raters were trained in the use of scoring rubrics. Inter-rater reliability was computed. On average, the percentage of exact agreement was 80.3% for the first item, and 96.1% for the second item. Disagreements were discussed to reach a consensus score.

For the think-aloud procedure, students met one-on-one with a researcher following the math test. In most cases the think-aloud interview took place within one hour of the student completing the math test. Students were selected on a voluntary basis from those who turned in a parent permission form for the interview (separate from the testing). The average duration of the interview was 20 minutes and 16 seconds per student. In three of the cases, students ran out of time to complete the entire interview. Students were first provided with instructions, and informed that glossary words were printed next to the items. Then students were shown a three-minute video clip demonstrating how to think aloud. Students first performed a "think aloud" while solving the five items in the Student Think Aloud Test (concurrent verbal report), followed by an interview (retrospective verbal report; Ericsson & Simon, 1993). When students struggled with thinking aloud, they were encouraged to continue verbalizing (Ericsson & Simon). Interviews were conducted primarily in English, however, a few students were encouraged to use their native language (Spanish and Mandarin Chinese only) when they struggled with answering interview questions. Students who could not read or speak any English were excluded from the study. All sessions were audio recorded, and then later transcribed.

Data Analysis

Quantitative Analysis for the Experimental Design. In order to examine the effectiveness and validity of read-aloud and glossary accommodations (Read Aloud and Glossary, respectively, hereinafter), the first and second research questions were investigated using regression analyses. The analyses focused on examining: (1) whether there was increased performance of ELL students with the provision of one of the two accommodations compared to ELL students with no accommodation provided; and (2) whether the two accommodations would not affect non-ELL students' performance on the test. Regression analyses were conducted to compare the student scores across different conditions (i.e., Read

Aloud, Glossary, or Standard), separately for ELL and non-ELL students. Since the design of the study was based on randomization, the results were expected to provide fair comparison among the conditions, i.e., unbiased estimates of the effects of the two accommodations.

Another research question of interest was whether accommodation effects varied depending on students' English language proficiency (ELP) levels. Accordingly, the analysis included an examination of the interaction effects between the treatments (i.e., accommodations) and student characteristics (i.e., ELP levels).

Specifically, we used multiple regression models for the sets of analysis. While students within each school were randomly assigned to conditions, the design of this study was a typical multisite randomized trial, as students were nested within schools. In such studies, multilevel models typically provide a good analytical approach (see, e.g., Raudenbush & Liu, 2000; Seltzer, 2004; Shadish, 2002). In this study, the number of schools were fairly small (four in State X and nine in State Y); and thus after controlling for some key predictors in the model, no significant variability remained across schools. Under such settings the results from multilevel models and single-level models (i.e., multiple regressions) will be fairly similar. For the purpose of parsimony, we primarily used multiple regression models and controlled for school membership using binary indicators of schools. In settings where there is a need to check whether the results are robust against such specification of models, multilevel models are fitted in addition to multiple regression models to yield more reliable results.

Student Verbal Protocol Analysis. We conducted multiple close readings of each student interview transcript and developed a coding scheme based on the five targeted areas of interest: comprehension, problem solving, item difficulty, accommodation use, and general perception of accommodation use. Each transcript was coded with the established scheme using Atlas.ti[3] qualitative data analyses software by two researchers with an average of 84.1% exact agreement between the two raters. Disagreements were discussed to reach a consensus. Descriptive statistics were computed in order to find any patterns in the areas of interest listed above. The descriptive analysis was conducted on the two groups of ELL students: current and former. Current ELL students included those who had taken an ELP assessment and were categorized into one of the five ELP levels. Former ELL students included those who had been exited and under a two-year monitoring period as well as those exited for over two years.

[3] ATLAS.ti Scientific Software Development GmbH, Nassauische Str. 58, D-10717 Berlin, Germany.

Results

In the following section, we first present the results of the experimental study by each state, then the results from the students' verbal protocol analysis. Although the sample and utilized methods were generally described above, more detailed description about the sample and specific models fitted to test the accommodation effects are also included in this section. Note that the statistical analyses focused on current ELL students, excluding students who were reclassified as fluent in English (only descriptive statistics for former ELL students are reported here). This decision was based on the rationale that reclassification means students are able to fully benefit from English-only instruction and thus do not need and are not typically provided with accommodations. However, it is unclear whether former ELL students may still benefit from the accommodation when ELLs benefit, and/or whether they would respond to accommodations more similarly to ELLs or non-ELLs. Since this study involved only a small number of former ELL students, studying the subgroup in such issues was not possible.

Quantitative Results for Experimental Study: State X

Sample Characteristics. As described earlier, 267 Grade 8 students (140 ELL, and 127 non-ELL) from four schools in one urban school district participated in the testing. Among the ELL students, 23 were former ELL based on state assessment data, and thus excluded from the analysis, as described above.

Tables 3, 4, and 5 present the descriptive statistics for the participating students' outcome scores (experimental math test scores) by treatment condition, state assessment scores, socioeconomic status as indicated by free or reduced lunch (FRL) program, ELP assessment scores, and ELP levels, for ELL students, non-ELL students, and former ELL students, respectively. The distributions of student characteristics and scores were in general fairly similar across conditions, which one expects to see in randomized studies. However, this study involved relatively small sample sizes for each subgroup of interest (i.e., ELL and non-ELL), resulting in some differences in student characteristics across conditions. For example, in the ELL student sample (Table 3), students in the Glossary accommodation condition tended to have lower test scores on state content and ELP assessments, and also were more likely to be receiving free or reduced lunch than those in the other conditions, although these differences may not be statistically significant. Also, in the non-ELL sample (Table 4), students in the two accommodation conditions had lower average test scores on state content assessments than students in the Standard condition. These preexisting differences are adjusted in the analysis, as described later.

Table 3

State X Descriptive Statistics for ELL Students by Condition

	Standard			Glossary			Read Aloud		
Variables	n	*M*	*SD*	n	*M*	*SD*	n	*M*	*SD*
Test Score	44	13.14	4.75	36	10.56	3.36	37	13.95	4.92
State Math Scale Score	42	208.29	92.74	35	171.26	71.80	35	213.69	85.29
State Reading Scale Score	42	230.05	63.69	35	228.69	44.95	35	227.60	65.06
FRL	42	0.69	0.47	35	0.91	0.39	35	0.77	0.43
ELP Test Scale Score	37	531.05	55.09	35	519.06	39.67	23	540.57	36.80
ELP Level	37	3.78	1.32	35	3.31	1.11	23	4.13	0.97

Note. Test score refers to scores from the math test in this study, out of a total 39 possible points. The state achievement test scale scores in math and reading range from 100 to 500. FRL refers to the proportion of students participating in the free or reduced lunch program. The state ELP test scale score refers to the overall score, and ranges from 341 to 666. The state ELP levels range from 1 to 5, with 5 being the highest level of proficiency. N sizes are lower due to missing background data.

Table 4

State X Descriptive Statistics for Non-ELL Students by Condition

	Standard			Glossary			Read Aloud		
Variables	n	*M*	*SD*	n	*M*	*SD*	n	*M*	*SD*
Test Score	42	21.98	5.85	48	19.52	6.06	37	19.76	6.63
State Math Scale Score	40	340.95	65.07	46	322.48	62.47	37	318.41	79.16
State Reading Scale Score	40	330.28	44.97	46	314.30	52.58	37	314.14	44.32
FRL	40	0.40	0.50	46	0.39	0.49	37	0.41	0.50

Note. Test score refers to scores from the math test in this study, out of a total 39 possible points. The state achievement test scale scores in math and reading range from 100 to 500. FRL refers to the proportion of students participating in the free or reduced lunch program. N sizes are lower due to missing background data.

Table 5

State X Descriptive Statistics for Former ELL Students by Condition

	Standard			Glossary			Read Aloud		
Variables	n	*M*	*SD*	n	*M*	*SD*	n	*M*	*SD*
Test Score	6	19.33	5.89	9	18.89	7.94	8	20.13	6.77
State Math Scale Score	6	292.17	67.21	9	358.67	58.16	8	344.50	49.07
State Reading Scale Score	6	301.67	44.39	9	322.11	36.64	8	313.25	45.44
FRL	6	0.67	0.52	9	0.78	0.44	8	0.38	0.52

Note. Test score refers to scores from the math test in this study, out of a total 39 possible points. The state achievement test scale scores in math and reading range from 100 to 500. FRL refers to the proportion of students participating in the free or reduced lunch program. N sizes are lower due to missing background data.

Results for ELL Students. We used the following multiple regression model:

$$Y_i = \beta_0 + \beta_1 Glossary_i + \beta_2 ReadAloud_i + \beta_3 Mathscore_i + \beta_4 Mathscore^2_i + \beta_5 School1_i + \beta_6 School2_i + \beta_7 School3_i + \beta_8 Admin_i + r_i \quad [1]$$

$$r_i \sim N(0, \sigma^2)$$

The outcome in the above multiple regression model, Y_i, is the number of items student *i* answered correctly in the math test developed for the current study. The descriptive statistics for the outcome is shown in the tables above, in the row labeled "Test Score." For the overall State X sample, the mean and standard deviation of the outcome were 16.9 and 6.8, respectively, with test scores ranging from a minimum of 5 to a maximum of 37.

In the regression model, *Glossary* is a binary indicator of whether a student *i* was assigned to the Glossary accommodation condition, while *ReadAloud* is an indicator of whether a student *i* was assigned to the Read Aloud accommodation condition. *Mathscore* is the scale scores from the state standardized math assessment in Grade 8; and the quadratic term is also included to capture a curvature of the relationship. School 1, 2, and 3 are binary indicators of whether students *i* was in schools 1, 2, or 3, respectively (an indicator for School 4 was not included in the model because it serves as a baseline). *Admin* is whether student *i* was in an administration setting where students had less than 45 minutes to complete the test (due to various, unexpected logistical challenges, in a few classrooms).

With such coding schemes, the key parameters of interest are β_1 and β_2. The parameter, β_1 represents the expected difference in the outcome between the Glossary and Standard

conditions, while β_2 represents the expected difference in the outcome between the Read Aloud and Standard conditions. In randomized studies, these expected differences can be considered as the effects of treatments (i.e., Glossary and Read Aloud).

We controlled for levels of math content knowledge measured by the state standardized assessment (*Mathscore*). This serves dual purposes: 1) to control for remaining imbalances in terms of the characteristic after the randomization; and 2) to increase the statistical power of estimating the effects of treatments given the relatively small sample sizes and high correlations between the outcome and the *Mathscore* variable. Since the relationships between the outcome and the predictor is not linear but shows curvature, we included the quadratic term ($Mathscore^2$) in the equation as well.

Table 6

State X Multiple Regression Results for Current ELL Students (n=112)

	Estimate	*SE*	*p*
Intercept	15.67	1.90	<.0001
Glossary	-1.34	0.87	0.12
ReadAloud	0.53	0.93	0.57
Mathscore	0.06	0.01	<.0001
Admin	-3.09	2.06	0.14
$Mathscore^2$	0.00	0.00	<.0001
School1	-2.75	1.86	0.14
School2	-2.33	1.86	0.21
School3	-1.93	1.91	0.32
Residual	13.51		

Table 6 presents the results for 112 current ELL students (those without missing background data) from the above multiple regression analysis. All parameters that we controlled for showed the direction of relationships we expected: a positive and significant math content knowledge-outcome relationship, lower performance for students in classrooms that ran out of time (which was not significant after controlling for other variables). The effects of both accommodations relative to no accommodation were not significant, indicating null effects of the accommodations.

Other sets of regressions including more predictors or different sets of predictors than the regression shown in Equation 1 were also conducted, but the result tables are not presented here. Other predictors were added to the equation but were dropped in the final

model shown in Equation 1, because they did not explain much variability in the outcome beyond the predictors that are already in Equation 1. These predictors include Reading scores in the state assessment, free or reduced lunch status, and the ELP scores or levels.

Student ELP levels, as measured by state ELP assessment, was a key predictor of interest, given that the study hypothesizes that ELL students may benefit from treatments (i.e., Glossary and Read Aloud) differentially depending on their ELP levels, as noted above. However, results did not show such interaction effects with student ELP levels. A close look at the data shows that there were more students in the medium to high levels (i.e., Levels 3, 4, and 5, out of a possible 5 for State X's ELP test) than lower levels (see Table 7), which means we may not have enough power to detect such interaction effects. Furthermore, in Grade 8 mathematics, math content knowledge appears to be a dominant factor over other predictors that we expected to be important, such as student ELP levels. The math test used in this study was correlated with students' math scores on the state standardized assessment (Pearson r = .44) and almost as highly with reading scores on the state standardized assessment (Pearson r = .41), but not as highly with ELP scores (Pearson r = .22). Figure 2 shows a scatterplot of the outcome scores against student ELP scores. As one can see, many students were clustered at the medium to higher levels (Levels 3, 4, 5, or a score of over 500), of which the scores ranged from 341 to 666. One can see clearly that, among these students with the same level of ELP, student outcome performance in outcome show substantial variation. The scatterplot displays a reason why student ELP scores may not be as related to the outcome scores as we hypothesized.

Table 7

State X Frequency of ELP Levels of Current ELL Students (n=95)

ELP Level	Frequency	%
1	8	8.4
2	5	5.3
3	23	24.2
4	31	32.6
5	28	29.5

Note. Level 1 is lowest. There were 22 students with missing data not included in this table.

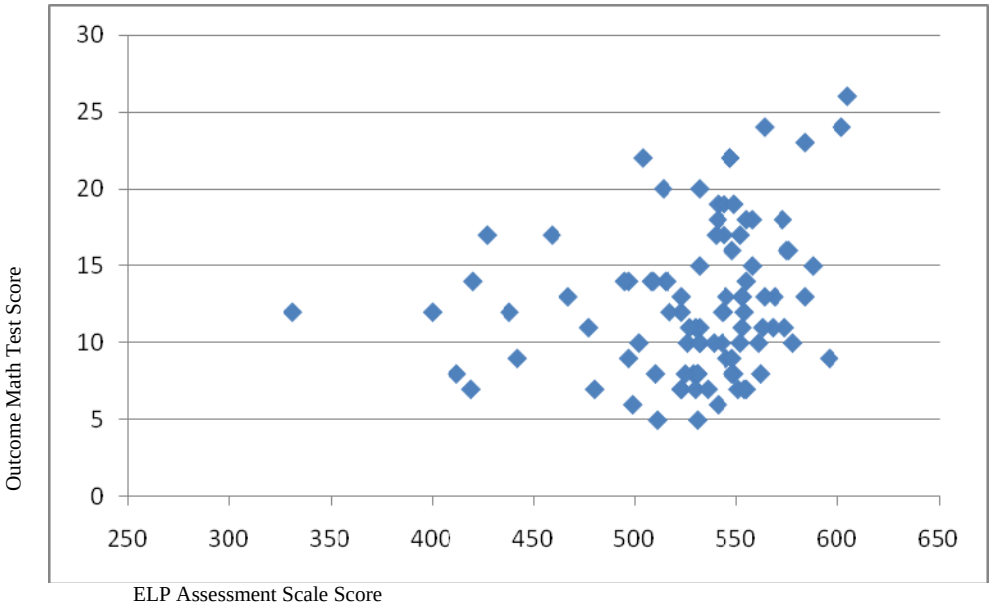


Figure 2. Scatterplot of the outcome math test score against the ELP assessment scale score for current ELL students. Outcome math test has a maximum of 39 possible points. ELP scale score ranges from 341 to 666.

Results for Non-ELL Students. We followed a very similar process in the analysis of non-ELL students to the analysis of ELL students. We controlled for levels of math content knowledge measured by state standardized assessments (*Mathscore*) for the same reasons: to control for remaining imbalances in terms of preexisting characteristics; and to increase the statistical power of estimating the effects of treatments. As with the analysis of ELL students, other sets of regressions with more predictors or different sets of predictors were also conducted. The final model was the same model used for ELL students shown in Equation 1 earlier.

Table 8

State X Multiple Regression Results for Non-ELL Students (n=123)

	Estimate	*SE*	*p*
Intercept	21.20	1.98	<.0001
Glossary	-0.88	0.93	0.35
ReadAloud	-2.51	1.08	0.02
Mathscore	0.05	0.01	<.0001
Admin	-6.27	2.12	0.00
Mathscore2	0.00	0.00	0.00
School1	-3.07	1.99	0.13
School2	-4.39	1.97	0.03
School3	-3.47	2.05	0.09
Residual	18.11		

Table 8 presents the results for 123 non-ELL students (those without missing background data) from the multiple regression analysis. As with the results for ELL students, all parameters that we controlled for showed the direction of relationships we expected: positive and significant math content knowledge-outcome relationship with very slight curvature, lower performance of students who were in classrooms that ran out of time. The effect of the Glossary condition relative to the Standard condition was not significant. However, the Read Aloud condition showed significantly lower performance in the outcome scores relative to the Standard condition, which indicates that the Read Aloud, on average, significantly hampered the performance of non-ELL students in the outcome.

A major criticism of not employing multilevel models in nested settings is that the results may yield erroneously small standard errors, which can make corresponding coefficients statistically significant when in reality they are not. Although such criticism may be unlikely to apply to this particular sample, we also ran a multilevel model that accounts for the nesting nature of the data to see whether the result is sensitive to the differences in model specification. Although the coefficient of *Readaloud* (beta), which captures the expected difference in outcome between Read Aloud and Standard conditions, was of a smaller magnitude and not significant in the traditional sense ($p = .06$), it still approached significance and suggests that Read Aloud may negatively affect the performance of non-ELL students in a Grade 8 mathematics assessment (see Table 9).

Table 9

State X Multilevel Model Results for Non-ELL Students (n=123)

Fixed Effects	Coefficient	*SE*	*p*
Intercept	17.76	0.93	0.00
Glossary	-0.93	0.93	0.32
ReadAloud	-1.97	1.04	0.06
Mathscore	0.04	0.01	<.0001
Admin	-3.03	1.18	0.01
Mathscore^2	0.00	0.00	<.0001
Random Effects	**Variance Component**	**SE**	**p Value**
Between-school in intercept	0.31	1.20	0.40
Within-school residual	18.34	2.44	<.0001

Quantitative Results for Experimental Study: State Y

As mentioned earlier, in State Y, 338 Grade 9 students (173 ELL, and 165 non-ELL) from nine schools in four school districts (three suburban and one urban) participated in the testing. Among the ELL students, 35 students were former ELL based on state assessment data and thus were excluded from the analyses (as described earlier).

Tables 10, 11, and 12 present the descriptive statistics for the participating students' outcome scores (experimental math test scores) by treatment condition, state assessment scores, socioeconomic status as indicated by free or reduced lunch (FRL) program, ELP assessment scores, and ELP levels, for ELL students, non-ELL students, and former ELL students, respectively. The distributions of student characteristics and scores in general were fairly similar across conditions, which one expects to see in randomized studies. However, this study involves a relatively small sample size for each subgroup of interest (i.e., ELL and non-ELL), resulting in some differences in student characteristics across conditions. For example, in the ELL student sample, students in the Read Aloud condition had lower test scores on average on the state standardized math assessment and the ELP assessment as compared to students in the Standard condition, although these differences may not be statistically significant. The analyses adjust for these preexisting differences.

Table 10

State Y Descriptive Statistics for Current ELL Students by Condition

	Standard			Glossary			Read Aloud		
Variables	n	*M*	*SD*	n	*M*	*SD*	n	*M*	*SD*
Test Score	43	13.09	4.03	43	12.86	5.49	52	13.38	4.69
State Math Scale Score	43	481.65	64.23	41	500.51	41.48	44	468.52	62.6
State Reading Scale Score	43	569.09	56.63	41	556.41	56.85	44	554.18	45.43
FRL	43	0.88	0.32	43	0.88	0.32	52	0.83	0.38
ELP Test Scale Score	36	548.17	41.74	38	544.11	34.73	42	541.50	41.57
ELP Level	36	3.67	0.99	38	3.53	0.73	42	3.50	0.99

Note. Test score refers to the math test in this study, out of a total 39 possible points. The state achievement test scale scores range from 310 to 890 for math, and 330 to 990 for reading. FRL refers to the proportion of students participating in the free or reduced lunch program. The state ELP test scale score refers to the overall score, and ranges from 341 to 666. The state ELP levels range from 1 to 5, with 5 being the highest level of proficiency. N sizes are lower due to missing background data.

Table 11

State Y Descriptive Statistics for Non-ELL Students by Condition

	Standard			Glossary			Read Aloud		
Variables	n	*M*	*SD*	n	*M*	*SD*	n	*M*	*SD*
Test Score	51	18.65	6.87	55	19.71	6.06	59	17.97	6.03
State Math Scale Score	42	548.62	48.33	47	555.17	52.42	45	542.96	55.63
State Reading Scale Score	42	635.52	47.47	46	648.89	41.10	45	640.78	50.46
FRL	51	0.22	0.42	55	0.20	0.40	59	0.17	0.38

Note. Test score refers to the math test in this study, out of a total 39 possible points. The state achievement test scale scores range from 310 to 890 for math, and 330 to 990 for reading. FRL refers to the proportion of students participating in the free or reduced lunch program. N sizes are lower due to missing background data.

Table 12

State Y Descriptive Statistics for Former ELL Students by Condition

	Standard			Glossary			Read Aloud		
Variables	n	*M*	*SD*	n	*M*	*SD*	n	*M*	*SD*
Test Score	15	16.20	7.94	12	17.75	6.73	8	17.63	5.15
State Math Scale Score	15	541.60	58.51	11	560.82	42.02	8	554.00	42.37
State Reading Scale Score	15	629.73	28.22	11	631.55	26.79	8	616.13	32.35
FRL	15	0.80	0.41	12	0.67	0.49	8	1.00	0.00

Note. Test score refers to the math test in this study, out of a total 39 possible points. The state achievement test scale scores range from 310 to 890 for math, and 330 to 990 for reading. FRL refers to the proportion of students participating in the free or reduced lunch program. N sizes are lower due to missing background data.

Results for ELL Students. We used the following multiple regression model:

$$Y_i = \beta_0 + \beta_1 Glossary_i + \beta_2 ReadAloud_i + \beta_3 Mathscore_i + \beta_4 Mathscore^2_i + \Sigma_{k=1to8}\, \beta_k School_{ki} + \beta_8 Admin_i + \beta_9 Glossary_i\, Mathscore_i + \beta_{10} ReadAloud_i\, Mathscore_i + r_i, \quad [2]$$

$$r_i \sim N(0, \sigma^2)$$

The outcome in the above multiple regression model, Y_i, is the number of items student *i* answered correctly in the math test developed for this study. The descriptive statistics for the outcome is shown in Tables 10, 11, and 12 above in the row labeled "test score." For the overall State Y sample, the mean and standard deviation of the outcome were 16.3 and 6.4, respectively, with test scores ranging from a minimum of 4 to a maximum of 35.

In the regression model, *Glossary* is a binary indicator of whether a student *i* was assigned to the Glossary condition, while *ReadAloud* is an indicator of whether a student *i* is assigned to the Read Aloud condition. *Mathscore* is the scale score from the state standardized math assessment at Grade 8, and the quadratic term is also included to capture a curvature of the relationship. School1 to School8 are binary indicators of whether student *i* is in schools 1, 2, to 8, respectively (an indicator for School 9 was not included in the model because it serves as a baseline). *Admin* is whether student *i* was in an administration setting where students had less than 45 minutes to complete the test.

With such coding schemes, the key parameters of interest are β_1 and β_2. The parameter, β_1 represents the expected difference in the outcome between the Glossary and Standard conditions, while β_2 represents the expected difference in the outcome between the Read

Aloud and Standard conditions. In randomized studies, these expected differences can be considered as the effects of treatments (i.e., Glossary and Read Aloud).

While β_1 and β_2 represent main effects of the each treatment (i.e., Glossary and Read Aloud), the parameters β_9 and β_{10} represent interaction effects of the treatments. β_9 represents the interaction effect between the Glossary treatment and student math score in the state standardized assessment on the outcome, which captures the expected difference in the math score-outcome relationship in the Glossary condition relative to the Standard condition. Likewise, β_{10} represents the interaction effect between the Read Aloud treatment and student math score in the state standardized assessment on the outcome, which captures the expected difference in the math score-outcome relationship in the Read Aloud condition relative to the Standard condition.

We controlled for levels of math content knowledge measured by the state standardized assessment (*Mathscore*), similar to the model for State X. Since the relationships between the outcome and the predictor is not linear but shows slight curvature, we included the quadratic term ($Mathscore^2$) in the equation as well.

Table 13

State Y Multiple Regression Results for Current ELL Students (n=128)

	Estimate	*SE*	*Z*	*p*
Intercept	13.93	0.81	17.12	<.0001
Glossary	0.67	0.76	0.88	0.38
ReadAloud	3.00	0.88	3.40	<.001
Mathscore	0.07	0.01	6.54	<.0001
Admin	-3.24	1.41	-2.30	0.02
$Mathscore^2$	0.00	0.00	5.11	<.0001
Glossary × Mathscore	0.04	0.01	3.00	<.01
ReadAloud × Mathscore	0.02	0.01	2.19	0.03
School1	0.22	1.01	0.22	0.83
School2	-0.70	1.01	-0.69	0.49
Shcool3	1.77	1.33	1.33	0.19
School4	-0.71	1.35	-0.54	0.59
School5	-1.08	0.93	-1.16	0.24
School6	0.01	0.94	0.01	0.99
School7	2.97	1.38	2.15	0.03
School8	1.37	1.17	1.17	0.24
Residual	8.74			

Table 13 presents the results for 128 current ELL students (those without missing background data) from the above multiple regression analysis, which is shown in Equation 2. All parameters that we controlled for showed the direction of relationships we expected: positive and significant math content knowledge-outcome relationship, lower performance of students in classrooms that ran out of time (which was significant). Read Aloud showed a significant positive effect on the outcome relative to the Standard condition. The expected effect on the outcome was 3.00, reaching almost two thirds of one standard deviation of the outcome. This is considered as a medium to large effect sizes in traditional statistics literature (e.g., Cohen, 1988). However, the main effect of the Glossary accommodation was not significant, indicating null effect of the accommodation for ELL students, on average.

In addition to the main effects, the specified model was a result of further examinations of interactions of both accommodations with ELL pretreatment characteristics. The results indicate that both accommodations interact with student math content knowledge, as

measured by the state standardized assessment. The direction of the interactions indicates that students with higher levels of content knowledge benefit (i.e., scored higher on the state math assessment) from the accommodations more than students with lower levels of content knowledge (i.e., scored lower on the state math assessment).

Similar to the analysis for State X, we also ran a multilevel model for State Y that accounts for the nesting nature of the data to see whether the result is sensitive to the differences in model specification. The results from the multilevel model, as shown in Table 14, show similar findings to the multiple regression results earlier (which did not account for nested settings): a significant main effect of Read Aloud; and positive interaction effects of both accommodations with math content knowledge, as measured by the state standardized assessment.

Table 14

State Y Results from Multilevel Models for Current ELL Students (n=128)

Fixed Effects	Coefficient	*SE*	*p*
Intercept	14.31	0.57	<.0001
Glossary	0.42	0.83	0.61
ReadAloud	2.54	0.73	0.00
Mathscore	0.07	0.01	<.0001
Admin	-2.87	1.01	0.01
$Mathscore^2$	0.00	0.00	<.0001
Glossary × Mathscore	0.04	0.02	0.03
ReadAloud × Mathscore	0.02	0.01	0.05
Random Effects	**Variance Component**	***SE***	***p***
Intercept	0.42	0.78	0.29
Standard condition Residual	8.41	2.02	<.0001
Glossary condition Residual	13.34	3.08	<.0001
Read Aloud condition Residual	4.95	1.16	<.0001

Figure 3 shows the estimated relationships between math score in the state's standardized assessment and the outcome score, respectively for each treatment condition

(i.e., Standard, Glossary, and Read Aloud). As the figure shows, compared to the Standard condition (the line connecting small diamonds), the fitted line for Read Aloud (the line connecting triangles) is above the fitted line for the Standard condition, which is from the significant main effect of Read Aloud. However, due to the interaction effect, the difference between the fitted lines between Read Aloud and Standard becomes greater for students with higher prior math score (in the state standardized assessment). For example, in the left end of the fitted lines (students whose scores are 2 SDs below the average in *Mathscore*), the expected difference between the conditions in the outcome was .5 points, while in the right end of the fitted lines (students whose scores are 2 SDs above the average in *Mathscore*), the expected difference between the two conditions was about 5 points in the outcome, which was about ten times the difference at the lower end.

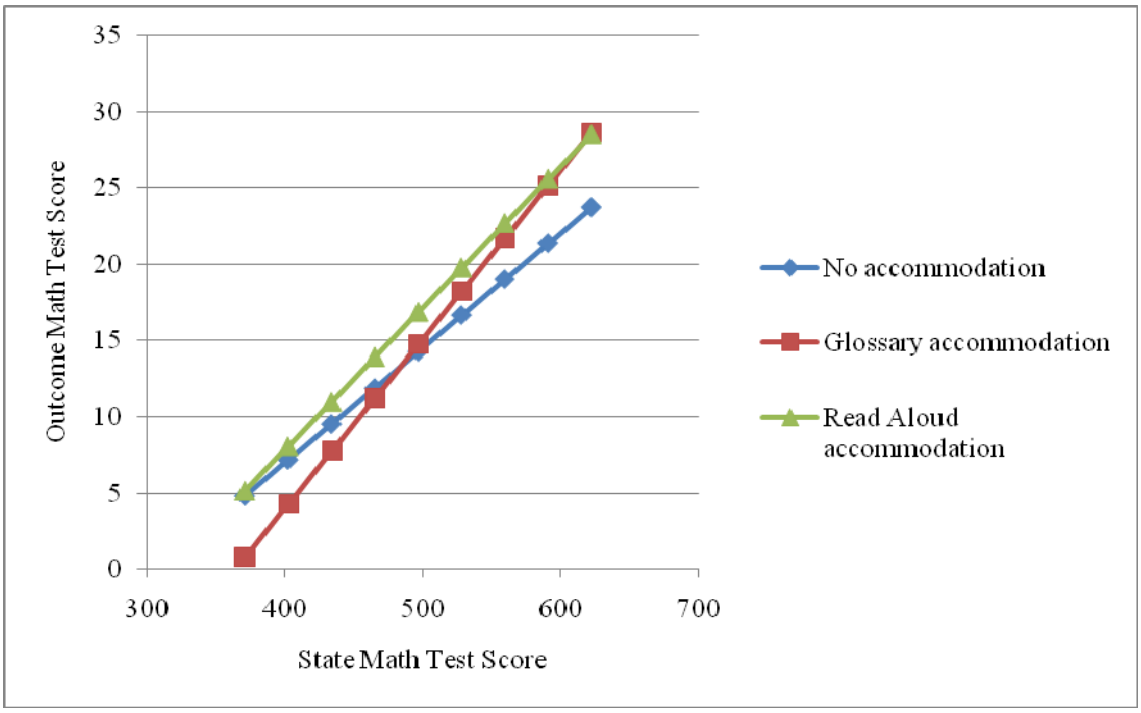


Figure 3. Fitted lines for three conditions showing the estimated relationships between state math test score and the outcome.

For the Glossary accommodation, the fitted lines represent no main effect but only interaction effects, since the fitted line for the Glossary condition (the line connecting large squares) was above the fitted line for the Standard condition for about half of the students, and below for the other half of the students. Students who had scored lower on the state math assessment performed worse with the Glossary accommodation than students who received no accommodation. However, students who had scored higher on the state math assessment

performed better with the Glossary accommodation than students who received no accommodation.

Similar to the analysis conducted for State X, we conducted analysis addressing whether the accommodation effects varied depending on students' ELP levels. We did not find such interaction effects in this study. A close look at the data shows that the majority of students were clustered at two levels, with few students at other levels (see Table 15), and thus we may not have enough power to detect such interaction effects. Furthermore, in Grade 8 mathematics, prior math score appears to be a dominant factor more than any other predictors that we expected to be important. The math test used in this study was correlated with students' math scores on the state standardized assessment (Pearson r = .60), but not as highly with either reading scores in state standardized assessment or with ELP scores (Pearson r = .23 and .22, respectively).

Table 15

State Y Frequency of ELP Levels of Current ELL Students (n=116)

ELP Level	Frequency	%
1	4	3.5
2	10	8.6
3	30	25.9
4	61	52.6
5	11	9.5

Note. Level 1 is lowest. There were 22 students with missing data not included in this table.

In summary, we found that student prior math scores tended to moderate the effects of both accommodations, Glossary and Read Aloud, benefiting students with higher math scores more than students with lower math scores. However, we did not find evidence that student ELP level relates to the accommodations effects. As noted above, there is a possibility that the test lacks statistical power, since more than half of the students were clustered in one level, Level 4.

To examine the extent to which the magnitude of effects were moderated by English language skills rather than math content knowledge, we focused on students' reading scores on the state standardized assessment and used it as a proxy for skills and knowledge related to ELP, since the reading scores are distributed with a bell-shaped curve, unlike student ELP

scores. The results showed no significant interaction between either treatment and student reading score, and are thus not reported here.

Results for Non-ELL Students. We followed a very similar process in the analysis of non-ELL students to the analysis of ELL students. We controlled for levels of math content knowledge measured by state standardized assessments (*Mathscore*) for the same reasons: to control for remaining imbalances in terms of pre-treatment; and to increase the statistical power for estimating the effects of treatments. As with the analysis of ELL students, other sets of regressions with more predictors or different sets of predictors were also conducted. The final model used was equivalent to the model for the ELL students, shown in Equation 2, with the exception of the interactions terms. We used the following regression model:

$$Y_i = \beta_0 + \beta_1 Glossary_i + \beta_2 ReadAloud_i + \beta_3 Mathscore_i + \beta_4 Mathscore^2_i + \Sigma_{k=1to8}\ \beta_k School_{ki} + \beta_8 Admin_i + r_i, \quad [3]$$

$$r_i \sim N(0, \sigma^2)$$

Table 16
State Y Multiple Regression Results for Non-ELL Students (n=134)

	Estimate	*SE*	*p*
Intercept	17.35	1.07	<.0001
Glossary	-0.17	0.86	0.84
ReadAloud	-0.16	0.92	0.86
Mathscore	0.09	0.01	<.0001
Admin	0.09	2.53	0.97
$Mathscore^2$	0.00	0.00	0.01
School1	-3.79	1.76	0.03
School2	-0.20	2.03	0.92
School3	-1.50	1.27	0.24
School4	-1.02	1.43	0.48
School5	0.87	1.47	0.56
School6	0.10	1.23	0.94
School7	-1.88	1.16	0.11
Residual	15.76		

Table 16 presents the results for the 134 non-ELL students (without missing background data) from the above multiple regression analysis. The math content knowledge-

outcome relationship tended to be positive and significant with curves. The effects of both accommodations relative to no accommodation were not statistically significant, indicating null effect of the accommodations. We also examined whether the two accommodations interact with any student pretreatment characteristics, but no interaction was found to be significant.

Qualitative Results: Students' Verbal Protocol Analysis

As described earlier, the students' verbal protocol analysis aimed to identify the difficulties ELL students encountered while taking a math assessment. That is, whether the students' difficulty stemmed from limited English language proficiency or lack of mathematical content knowledge was a focus of the study. The qualitative analysis also focused on the students' use of the given accommodation and their perception about the helpfulness of accommodations in taking a math assessment. The results are presented corresponding to the research questions: language and content difficulties, the use of the glossary accommodation, students' prior experience with the focal accommodations, and students' perceptions about the focal accommodations.

Language Difficulty in Items. In order to examine the extent to which the students had difficulty in understanding the language in the sample items, students were asked to paraphrase what the question was asking in their own words. Based on the students' think-aloud and retrospective interview responses, four codes were assigned including "Yes: Students comprehended the question," "Partial comprehension: There were some parts that students were unable to paraphrase or they said they did not understand about certain parts," (in other words students comprehended the gist of the story in an item, but did not adequately paraphrase parts of the story in an item), "No: Students did not comprehend the question or were unable to paraphrase the question at all," and "Not sure: There was not enough evidence to judge students' comprehension of the language." There were also a few cases where students did not have enough time to complete each item (as described in the Method section earlier). These cases were coded as missing responses with "Not sure" cases.

Table 17 presents the summary of students' comprehension of the five sample items. The results are presented by the students' ELL status: current and former ELL students.

Table 17

Language Comprehension Results for Each Item by ELL Status

Item	Yes	Partial	No	Total
		Current ELL		
1	40 (76.9)	8 (15.4)	4 (7.7)	52 (100.0)
2	20 (44.4)	10 (22.2)	15 (33.3)	45 (100.0)
3	23 (47.9)	22 (45.8)	3 (6.3)	48 (100.0)
4	26 (51.0)	22 (43.1)	3 (5.9)	51 (100.0)
5	23 (51.1)	16 (35.6)	6 (13.3)	45 (100.0)
		Former ELL		
1	13 (100.0)	0 (0.0)	0 (0.0)	13 (100.0)
2	11 (84.6)	1 (7.7)	1 (7.7)	13 (100.0)
3	11 (78.6)	3 (21.4)	0 (0.0)	14 (100.0)
4	11 (78.6)	3 (21.4)	0 (0.0)	14 (100.0)
5	13 (92.9)	1 (7.1)	0 (0.0)	14 (100.0)

Note. Former ELL students included those who were under a two-year monitoring period as well as those who had been exited for over two years.

As shown in Table 17, former ELL students generally comprehended all five items by being able to rephrase the items in their own words. Although some students demonstrated difficulty explaining Items 3 and 4, they still showed at least partial comprehension for these items. The current ELL students were less able to appropriately paraphrase the items in their own words to demonstrate their comprehension of language in items compared to former ELL students in the study. Yet, the current ELL students also demonstrated at least partial understanding of the items by describing the part of items in their own words. It is notable only a few current ELL students had no comprehension, particularly for Items 1, 3, and 4. It appeared that more students had difficulty in understanding Items 2 and 5.

The students were also asked whether each item included difficult words to understand and what those words were. Table 18 summarizes the students' responses on the vocabulary difficulty in each item.

Table 18

Vocabulary Difficulty Identified by Students in Each Item

Item	n	There Were Hard Words	Knew All Words	Identified Difficult Words/Phrases*
1	59	13 (22.0)	46 (78.0)	shaded (8), trapezoid (4)
2	54	27 (50.0)	27 (50.0)	closest approximation (27), tip (4), of the following (3), charges (2), percent (2), restaurant (2)
3	56	7 (12.5)	49 (87.5)	at least one (3)
4	56	30 (53.6)	26 (46.4)	Salt Flats (16), Talon Bluff (14), elevation (13), hikers (6), climbed (4)
5	49	20 (40.8)	29 (59.2)	plumber (9), additional (7), expressions (7), represents (5), calculate (5), travel (2)

*The number in parenthesis indicates the number of students who identified the given word/phrase as difficult.

As previously shown in Table 17, Item 1 was relatively easily understood (from a language standpoint) by the current ELL students compared to other items. Although the item contained some language demands by its grammatical complexity (complex noun phrase and passive structure), students described the questions as "how many triangles fit in the bigger shape," referring to the trapezoid in the item. Item 2 was found to be the most difficult for the current ELL students to comprehend (33% identified as "No" comprehension). Quite a few words in this item were perceived difficult by the students as shown in Table 18. While many students indicated that "closest approximation" was hard vocabulary, they struggled with appropriately describing the phrase of "a 15 percent tip on a check." The following excerpts indicate that these students were struggling to comprehend the phrase while repeatedly reading the item:

03W1G13 (State X Current ELL)

INT: You want to tell me what the question is asking in your own words? Like how would you, how would you explain this question to me?

STU: Of the following which is the closest … [*reads to self again*] I think it's asking like… oh like, what is the tip of… the... [*quietly re-reading*] the tip of…[*reading silently to self again*] I think they left like… tip for on … twenty four and uh point ninety nine cent… Ah, I can't think… I think they let…left their tip… of like twenty four ninety nine. Twenty four dollar and ninety nine cents like I think they left fifteen dollars with it... like…

18A1S05 (State Y Current ELL)

STU: OK...right now I'm looking at the answers again. So I'm looking for like, which person gives the tip, fifteen percent…[*4 seconds silence*]. I'd pick B, because the question number close to, gives a tip for…[*inaudible*].

INT: Close to? I didn't hear the last part.

STU: People giving out tips? ... I don't know. That's all.

INT: OK. Can you tell me what this question is asking in your own words?

STU: Ah...what percent of tip for? I don't know.

As for Item 3, most ELL students were able to comprehend this item at least partially. The item contained neither complex grammatical structures nor difficult words (i.e., the words were high-frequency, daily words). This item had some language demands in terms of its length and cohesive features by requiring one to process the references and connections within and across sentences. However, as seen in Table 17, only three students were unable to rephrase this item in their own words.

Item 4 was also understood by most students. This item contained an academic word, "elevation" and proper nouns such as "Salt Flats" and "Talon Bluff." Students often identified these words as difficult. Some students identified non-academic words such as "hikers" and common words such as "climbed" as difficult ones. Yet, most current ELL students at least demonstrated partial understanding of the language in the item by describing the problem as finding the "difference between two places." The following excerpt illustrates that an ELL student was still able to comprehend the item without knowing a specific word by inferring the meaning from the context:

18A1R16 (State Y Current ELL)

INT: How about any words that you didn't know?

STU: Yeah, this [*pointing to hikers*].

INT: hikers. Did you look at the side [*pointing to the glossary*]?

STU: No, I didn't look [*chuckles*].

INT: Why didn't you look at?

STU: Because I didn't know it [was there]…

INT: You still explained when I asked you what this question was asking. How did you figure out without knowing these words…hikers or elevation?

STU: Because you had to like, you have to...I read the front and the back and it says climbed...so maybe it's like, hikers, a person who hikes, climbs.

Item 5 was expected to have moderate to high demands of language on account of its length, grammatical structures (i.e., conditional and passive structures), and vocabulary (i.e., academic words). As shown in Table 18, many words in this item were identified as difficult by the students. Item 5 was considered relatively hard to comprehend as 13.3% of the current ELL students were unable to describe this item in their own words, as shown in Table 17. The students in the following excerpts were unable to demonstrate comprehension for what happened in the story of this item:

18A1G12 (State Y Current ELL)

STU: Um, they…[*7 seconds silence*] oh, that he works, um, like, uh, works and they need to pay the hours. And that say that if he works forty eight, I think forty eight hours, he's going to pay nine dollars for the hours?

22V1S07 (State Y Current ELL)

INT: Do you understand what the question is? Can you tell me what it is saying?

STU: Uh , like [*7 seconds silence*], I'm not really sure.

INT: What are they asking you to find? Do you know?

STU: The charges H represents. That's the number of hours he worked.

Content Difficulty in Items The students' problem-solving processes were examined through the students' think-aloud as well as the retrospective interview responses. By doing so, we attempted to unveil whether the students' struggles with solving the given math items was related to their lack of content knowledge. Four codes were assigned to students' verbal reports based on the students' problem-solving processes and answers: (1) correct (demonstrating appropriate mathematical knowledge to correctly solve the given item); (2) incorrect attempt (demonstrating some mathematical knowledge, but arriving at an incorrect answer); (3) guess (demonstrating no mathematical knowledge or the answer was chosen based on non-mathematical reasoning); and (4) no attempt (which includes circling a response but not providing any type of reasoning).

Table 19 summarizes the students' problem-solving results on each item. Overall, the current ELL students performed lower than the former ELL students on all five items, as shown in the percentage of "Correct," which is consistent with the overall test data, as well as ELL students' historical performance on math assessments in general. A number of current ELL students attempted to solve the given items, demonstrating an understanding of the language in the items, but did not use appropriate mathematical procedures to correctly solve the items.

Table 19

Problem-Solving Results for Each Item by ELL Status

Item	Correct	Incorrect Attempt	Guess	No Attempt	Total
		Current ELL			
1	13 (24.1)	33 (61.1)	0 (0.0)	8 (14.8)	54 (100.0)
2	4 (7.4)	29 (53.7)	11 (20.4)	10 (18.5)	54 (100.0)
3	12 (22.2)	35 (64.8)	1 (1.9)	6 (11.1)	54 (100.0)
4	8 (14.8)	42 (77.8)	0 (0.0)	4 (7.4)	54 (100.0)
5	11 (21.6)	30 (58.8)	3 (5.9)	7 (13.7)	51 (100.0)
		Former ELL			
1	6 (42.9)	8 (57.1)	0 (0.0)	0 (0.0)	14 (100.0)
2	4 (28.6)	5 (35.7)	3 (21.4)	2 (14.3)	14 (100.0)
3	7 (50.0)	7 (50.0)	0 (0.0)	0 (0.0)	14 (100.0)
4	6 (42.9)	8 (57.1)	0 (0.0)	0 (0.0)	14 (100.0)
5	10 (71.4)	3 (21.4)	1 (7.1)	0 (0.0)	14 (100.0)

Note. Former ELL students included those still under two years of monitoring status as well as those who had been exited for over two years.

A closer look at the students' verbal reports identified as "Incorrect Attempt" and "Guess" revealed how the students arrived at an incorrect answer. For Item 1, students who solved this item correctly applied a visual assessment approach, understanding that a trapezoid could be split into triangles. These students drew lines inside the trapezoid to arrive at the correct number of triangles. However, students who made an "incorrect attempt" tried to perform an arithmetic calculation, such as division, with the numbers presented in the item. Seemingly, the students, who may have lacked appropriate content knowledge, literally thought that the phrase "divided into" in the item required them to carry out division.

Item 2 was the most difficult for both current and former ELL students to solve correctly. The students who made "incorrect attempt" tended to divide 24.99 by 15. Many students also attempted to guess how much tip should be given based on real-life restaurant experiences. For instance, the following excerpt suggests that the students comprehended the language in the item, but lacked in content knowledge to solve the item correctly:

79B1S01 (State Y Former ELL)

STU: And what I did was I divided fifteen into twenty-four….And I also remember when we go to restaurants and the check is like that and my mom gives two fifty. So yeah that's also how I remembered it.

INT: OK so what answer did you get?

STU: I put two fifty…I got two fifty because I thought about how much my mom gave.

02W1S03 (State X Current ELL)

STU: Because um I remember one time I went to a restaurant and…we left ten percent tip and it was two dollars and something, so it couldn't be A and D was too much.

In order to solve Item 3 correctly, students were required to apply the operation of multiplication, addition, subtraction, and division in the proper sequence. Additionally, it was important to associate the numbers with the correct units (e.g., whether the given number was associated with people or pencils). Although most students were able to perform arithmetic operations, many students did not link the given numbers with the right unit at the last step. One of the distractors included the number that students could choose if they had incorrectly associated the unit.

Item 4 was the second most difficult item among the five items, although the language of the item was understood by most students. Most students recognized that the word "difference" indicated that they needed to subtract two given numbers. However, most students incorrectly calculated the subtraction of a negative number. Instead of adding the two numbers, students often subtracted and ignored the negative sign or assumed that the negative sign was equivalent to subtraction, as in the following excerpt:

79B1S08 (State Y Current ELL)

STU: So all I did was pretty much, I did six hundred and twenty-five minus negative fifty five, but yeah…technically, they don't add. They don't subtract or add. And the question is asking you "what was the difference." So I realized that it was a negative, so yeah, you

would just subtract it. Zero...and then this one's five…twelve, seven-[*Student writes 625-55=570 on paper.*]

Item 5 asked students to formulate an algebraic expression based on text. This item, more than others, required students to be able to translate language into mathematical symbols. That is, content knowledge entailed converting the language into a mathematical expression. To solve this problem, students needed to understand that *h* (representing hours) was a variable, but not an unknown to be solved. Among the students who made "incorrect attempt," some students attempted to solve the unknown variables. There were also students who may have been confused by the language in the item, which lead to an incorrect answer. The following excerpts demonstrate students' confusion about language in the item:

79B1R13 (State Y Current ELL)

STU: …It didn't make sense for me.

INT: Why do you say that?

STU: Because it says it's forty eight for each hour for work plus addition for nine so-like, how does that travel, where does he live, what if he lives close to them?

18A1G12 (State Y Current ELL)

STU: …oh that he works, um, like uh works and they need to pay the hours. And that say that if he works forty eight, I think forty eight hours, he's going to pay nine dollars for the hours?

Use of Glossary Accommodation. The five items in the think aloud contained between two to eight glossary words (as described in the Method section earlier). Through the student think aloud and retrospective interview, we examined whether and how students used the given glossary for each item. Students' verbal reports were categorized into three groups: (1) "No" (student did not look at the glossary at all); (2) "Looked" (student said that s/he looked, but did not use because s/he knew all the words); and (3) "Looked and Used" (student used the meaning shown in the glossary). Table 20 summarizes the students' use of glossary for each item.

Table 20

Students' Use of Glossary for Each Item

Item	n	No (Did Not Look)	Looked (But Knew All the Words)	Looked and Used
1	52	28 (53.8)	18 (34.6)	6 (11.5)
2	53	23 (43.4)	11 (20.8)	19 (35.8)
3	50	32 (64.0)	12 (24.0)	6 (12.0)
4	51	28 (54.9)	9 (17.6)	14 (27.5)
5	46	25 (54.3)	7 (15.2)	14 (30.4)

As shown in Table 20, across all five items, students mostly said they did not look at the glossary words. When prompted for reasons, students responded that they did not need the glossary because they already knew all the words, or because they forgot or did not realize the glossary was there, as in the following excerpts:

79B1S08 (State Y Current ELL)

INT: Did you look at these words on the side?

STU: No. Oops. I didn't realize they were there until like the third problem.

66P1R01 (State Y Current ELL)

INT: Did you look at any of those words on the side?

STU: I didn't look at any one.

INT: How come?

STU: I don't know. I just like, I forgot about it.

Some students recognized that the glossary was not always necessary for words, such as proper nouns, like "Talon Bluff" and "Salt Flats," as this student described:

79B1G19 (State Y Former ELL)

STU: I think I don't think I need to know the words. 'Cause they are just name of the place…I didn't know the--the name of the place. But they don't really count, 'cause like you don't need them to fix--to like, do the problem.

Among the students who participated in the think-aloud and interview, 21 students were in the glossary accommodation condition in the experimental study. One ELL student who had recently arrived in the United States commented that she did not realize what the

glossary words printed on the side were for when she took the test (even though they were mentioned in the test administration directions, which were read aloud to students):

20A1G03 (State Y Current ELL)

INT: Did you have these words here when you took the test over there?

STU: Yeah but I don't know why these words are on here so I never look at.

For students who actively used the glossary, they looked at the words that they identified difficult as presented earlier in Table 18. The glossary words for "approximation," "elevation," "additional," "shaded," "plumber," and "tip" were relatively frequently used. They reported that they were looking for a meaning in the glossary. A few students substituted words in the items with the glossary words. For instance, a current ELL student was observed writing the glossary word, "extra" underneath the word, "additional" in Item 5 as a substitution.

Students' Prior Experience with Glossary/Dictionary and Read Aloud Accommodations. In order to understand whether students were familiar with using the given accommodations, students were asked if they had previous experience with glossary, dictionary, and read aloud accommodations for a state's standardized math assessment. Table 21 presents students' responses by each state, considering different policies and practices for each state.

Table 21

Students' Prior Experience with Accommodations in Math Assessments by State

	State X			State Y		
Type	Yes	No	Total	Yes	No	Total
Dictionary	0 (0.0)	6 (100.0)	6 (100.0)	1 (2.4)	41 (97.6)	42 (100.0)
Glossary	0 (0.0)	15 (100.0)	15 (100.0)	0 (0.0)	42 (100.0)	42 (100.0)
Read Aloud	2 (18.2)	9 (81.8)	11 (100.0)	11 (31.4)	24 (68.6)	35 (100.0)

Note. Glossary was defined by showing the students' the built-in glossary printed in the math test booklet used in the present study.

As shown in Table 21, overall students had little experience with the given accommodations. Almost no students among the sample had previous experience of using a built-in glossary or dictionary for a state's math assessment. Two students in State X reported that the directions were read aloud and that the items were read aloud only when the students

raised their hands to ask. More students reported having experience with read aloud in State Y compared to students in State X. A student in State Y reported that all the entire problems were read aloud to students and the problems were repeated upon student request.

Perceptions on the Helpfulness of Accommodations. Students were asked if they felt glossary or read aloud accommodation was helpful to them. Students were first asked about the given accommodation provided to them during the experimental portion of the study (i.e., read aloud or glossary, when applicable) and then asked about their general perception about the helpfulness of both accommodations. Table 22 shows the results of students' perceptions of helpfulness, for both the Glossary test condition in this study, as well as a provision of a glossary for any mathematics assessment in general. Results indicate that most students felt having a glossary was helpful. Some students had mixed feelings toward a glossary, stating that it would only "sometimes" be helpful, such as only "if you need it."

Table 22

Students' Perceptions of the Helpfulness of Glossary

	Helpful	Not Helpful	Mixed	Total
Glossary Used in This Study	11 (78.6)	2 (14.3)	1 (7.1)	14 (100.0)
In General	30 (81.1)	2 (5.4)	5 (13.5)	37 (100.0)

Note. Glossary Used in This Study refers to the Glossary condition during the experimental portion of the study. In General refers to students' general perceptions, which was a question open to all students, regardless of prior experience or accommodation condition in this study.

The following excerpts demonstrate that students were conscious of their limited English language proficiency, positively thinking about glossary accommodation:

18A1S05 (State Y Current ELL)

INT: What do you think about these words on the side because we're interested in whether these words on the side will be helpful or distra--

STU: I think it'll be helpful, help other people because some people who don't know much English, some people no English, so if someone else is from different [*inaudible word*], it'll help solve the problem.

INT: How about you? Was that helpful to you?

STU: Yes, it was helpful to me because I'm still learning English. There are some words that I still can't pronounce yet.

01W1R07 (State X Current ELL)

INT: Do you think when you take a test and if there is a glossary like this [*pointing to the glossary version test*] it would be helpful or useful to you?

STU: Yeah they would be helpful because then you can understand what the words mean if you get stuck on those words.

Some students described why having a glossary similar to the one from the present study would be better than having a dictionary.

18A1S06 (State Y Current ELL)

INT: Do you think it would be helpful if you had words on the side or if the teacher gave you a dictionary?

STU: I think it was helpful if they was like this [*pointing to an open page of the glossed test booklet*]… Because instead of looking in the dictionary, taking a long time and looking, so you can just like, "Oh yeah, it was like-- Oh yeah, I know what 'travel' was meaning." You can know faster.

With respect to read aloud accommodation, students tended to view reading aloud favorably. Table 23 presents the results of students' perceptions of helpfulness, for both the Read Aloud test condition, as well as in general. Compared to glossary, more students with mixed or negative perception about helpfulness of reading aloud were noted.

Table 23

Students' Perception of the Helpfulness of Read Aloud

	Helpful	Not Helpful	Mixed	Total
Read Aloud Condition	15 (62.5)	4 (16.7)	5 (20.8)	24 (100.0)
In General	22 (62.9)	9 (25.7)	4 (11.4)	35 (100.0)

Note. Read Aloud Condition refers to students who were part of the read aloud condition during the experimental portion of the study. In General refers to students' general perceptions, which was a question open to all students, regardless of prior experience or accommodation condition.

As for the reasons to consider reading aloud helpful, students commented listening would be easier than reading as suggested in the following excerpts:

66P1R07: (State Y Current ELL) That it was easier, 'cause she was reading it so I only had to focus on the problem instead of reading it all.

19V1R14: (State Y Current ELL) Yeah, 'cause it-- sometimes I get stuck on words, and she can just read it faster.

19V1R13: (State Y Current ELL) Ah, it was helpful because when I have something to read, there's too much to read. So when the teacher read, I hear the word correctly, so I know what it, what it was talking about.

Students who felt the read aloud was not helpful attributed their reason to a different pace of solving problems or distracting factor. This was consistent with the student interview results conducted on regular students by Weston (2003).

78B2R10: (State Y Current ELL) I think it was confusing. Because when I was behind a question, I had to like--well just try not to think about the teacher, and just work on the one I was. And then when I was on the question the teacher was, then I was already confused, because I had to read it like two times at least to get the question.

04W1R16: (State X Current ELL) Yeah because sometimes you like finish early and then you have to wait until somebody, like so everybody finishes so they could start the other question.

79B1R16: (State Y Former/Exited ELL) I would say to do it alone…Because it would be much more easy to concentrate, I guess.

Students who had been in the read aloud condition for the experimental study were also asked how they felt about the speed of the teacher's read aloud. Among 16 students who were asked this question, 10 students (or 62.5%) commented that the read aloud speed was fast and that time allotted to solve the problems was not enough. However, almost all students (19, or 90.5%) who were asked reported that they followed along with the teacher during the read aloud.

Discussion

The present study investigated the effectiveness and validity of two accommodations, reading aloud the entire test, and glossary, provided to ELL students during a mathematics assessment. As described earlier, these two accommodation types have been commonly allowed across states' accountability assessments with an assumption that they, by directly supporting ELL students' linguistic barriers, would be effective strategies to be used. By effective, we mean that these accommodations are presumed to help ELL students' overcome some language barriers thereby increasing their assessment outcomes. In this report, we first presented an overview of previous literature, which demonstrates that little empirical research evidence is available to support this assumption. With the purpose of providing empirical evidence to shed light on the effectiveness and validity of these two

accommodations, this study employed a randomized experimental design accompanied by a student verbal protocol analysis.

Regarding the effect of the glossary accommodation, no significant difference of the ELL students' performance on the mathematics assessment was found in either state's samples, compared to the standard condition (i.e., receiving no accommodation). The students' verbal protocol analysis results provided some insight into this result. It was found that the majority of the students who participated in the think aloud did not utilize the provided built-in glossary while completing the five math test items. Several students reported that they "forgot" about the glossary and all students (who were asked) reported that they had never been provided a glossary during mathematics testing. A case study conducted as a subset study of the larger project also provided some insight into understanding the results of the present study. In this study, sample teachers reported that a glossary accommodation was not provided for the state mathematics assessment in either state, and was seldom used during mathematics instruction (Wolf et al., 2009). Teachers expressed that a glossary would require more skills and practice for students to effectively use. Thus, the finding of no glossary effect seems related to the sample students of the study, who were neither familiar with, nor skillful in using the provided glossary. Collective evidence insinuates that students' prior experience and skills in using a glossary may be an important factor for improving the effect of the accommodation.

As for the read-aloud accommodation, the statistical analysis yielded mixed results on its effect on the students' performance on a math test. In the State X sample, there was no significant difference in ELL students' performance on the given math test regardless of accommodation condition. However, a significant positive effect of the read-aloud accommodation was detected in the State Y ELL sample. ELL students who received the read-aloud accommodation tended to perform better on the math test compared to ELL students who were in the standard condition. Although the small sample size in this study limited us in generalizing the results to a bigger population, there are some plausible sources to explain these differences. First, the students' verbal protocol analysis revealed that State Y sample students had more prior experience with the read-aloud accommodation than State X sample students. Secondly, according to State Y policy, State Y provided a test script developed by their test publishers to be used for read aloud. State Y appeared to have a more systematic implementation of the read-aloud accommodation, when implemented. State X in contrast had no script to implement a read-aloud accommodation in a standardized way. The case study described above indicated that the read-aloud accommodation was not used for State X's 2008 mathematics assessment in the school district where State X students were

sampled from. It was also found that State Y sample teachers used the read-aloud accommodation for the state's mathematics assessment more often than State X sample teachers did (Wolf et al., 2009). We speculate that the mixed effect of the read-aloud accommodation was related to ELL students' prior experience, similar to the finding about the glossary accommodation. State Y students were more likely to have received a read-aloud accommodation in the past, and were more likely to have received one in a systematic way.

One thing to note is that the read-aloud accommodation showed positive regression coefficients with the ELL students in both states' samples, while the glossary accommodation showed inconsistent directions of the regression coefficients (i.e., negative coefficient for the State X sample, and positive coefficient for the State Y sample). While the directions of the coefficients were not statistically significant, we can speculate that the trend may suggest that the read aloud accommodation could help ELL students regardless of students' prior experience with read aloud. The glossary accommodation, however, based on the trend, may require both skills and familiarity to be an effective accommodation. To what extent students must acquire such skills to utilize a glossary or other accommodations would require further investigation.

Our analysis, which controlled for various students' characteristics, yielded a notable result regarding the interaction between accommodation effects and students' characteristics. In State Y ELL samples, there was significant interaction effect of both the glossary and read-aloud accommodations and ELL students' prior content knowledge, as measured by the states' mathematics assessments. For instance, ELL students who scored higher in their state mathematics assessment benefited more from having a given accommodation than ELL students who scored lower in their state's mathematics assessment. This result suggests that the given accommodations help ELL students who have acquired content knowledge but cannot help those who have not. This finding signifies the importance of providing accommodations to ensure the accessibility of content assessments for ELL students. The result implies that ELL students who have acquired content knowledge may not completely show their knowledge and skills on content assessments because of their limited English proficiency, and that providing accommodations helps to enhance the validity of content assessments by allowing ELL students to demonstrate what they know.

The analysis also examined whether the given accommodations worked differently for the ELL students depending on their levels of English proficiency. In both states' samples, no significant interaction effect was found between the given accommodation and students' ELP levels. Given that the sample of this study was small and its ELP levels were limited (i.e.,

students were mainly clustered at moderate to higher ELP levels), the interaction effect between the accommodation and ELP levels needs to be further investigated.

While the analysis on ELL students' performance in different accommodation conditions was directly concerned with the effectiveness of a given accommodation, it was also intertwined with a validity issue. That is, effective accommodations that help ELL students reduce any linguistic barriers that interfere with their ability to demonstrate their content knowledge can increase the validity of the test for ELL students. In addition, the analysis of non-ELL students' performance also attempted to examine the validity of the accommodations. Providing accommodations should not change the nature of the construct being measured. Thus, providing accommodations to non-ELL students, who would not need linguistic support, should not increase their test scores, thereby retaining the validity of test scores for non-ELL students. One way of addressing these validity concerns was to provide evidence that non-ELL students did not perform differently regardless of the accommodation condition.

The results of this study showed that there was no significant difference among non-ELL students' performance in the different conditions (i.e., Glossary, Read Aloud, and Standard) in the State Y sample. However, State X non-ELL students in the Read Aloud condition performed lower compared to their peers in the Standard condition, which was statistically significant. One may speculate that reading aloud distracted non-ELL students who had less trouble reading and understanding the questions silently by themselves. Since the results of the read-aloud accommodation on non-ELL students were inconsistent across samples (State X and State Y), validity evidence was somewhat weak in this study. Moreover, the small sample size did not allow for sufficient statistical power to detect the significance of the accommodation effects. Inevitably, further investigation on the validity of these two accommodations needs to be conducted.

Students' verbal protocol analysis provided a deeper understanding of ELL students' problem solving processes and students' use of glossaries provided during a math test. The primary purpose of the verbal protocol analysis was to explore whether ELL students struggled with comprehending the language of math items and thus could not understand the items, or whether they had difficulty in solving items due to lack of content knowledge. If the former was the case, it would suggest a validity threat, in that the test scores might not reflect what students knew and could do in the subject area. If the latter was the case, the validity concern may be more about students' opportunity to learn the content, rather than about providing appropriate accommodations for a test. The verbal protocol analysis also sheds some light on the interaction results from the experimental portion of the study.

The results from the verbal protocol analysis demonstrated that, as expected, ELL students had some difficulty in understanding the language of the five sample math items, as compared to former ELL students. What was noteworthy was that most ELL students comprehended the language at least enough to know what the sample items were asking. It seems that the students in this sample who had been in U.S. schooling since a young age were proficient in English enough to comprehend the test language in the given sample math items. The ELL students in the think-aloud analysis were found to struggle more with the content knowledge needed to solve an item correctly. That is, most ELL students attempted to solve the problems, which suggested understanding of the language in the items, but did not use appropriate mathematical procedures to correctly solve the items. This suggests that these students' lower performance may be attributed more to math knowledge rather than language issues. For example, they used inappropriate operations such as multiplication instead of division, and vice versa. Their computations were often incorrectly performed. Students' content knowledge limitations needs to be further investigated to explore whether their limited English language proficiency interfered with students' learning of the content area. It is also questionable whether the students had appropriate opportunity to learn (OTL) the curriculum materials as compared to their non-ELL peers. From the case study, part of the larger project mentioned earlier, one teacher mentioned that the ELL students were sometimes taught below-grade materials because of their low performance on the content area. The teacher pointed out that the students were too behind and that they needed to learn previous-grade materials first.

Regarding the use of a given accommodation, only a few students actively utilized the provided glossary during the think aloud. This has important implications for the experimental study results, and why we perhaps found no main effect of glossary. The students in the verbal protocol analysis identified difficult words for themselves and substituted the words with the ones found in the glossary while reading the items repeatedly. Although most students listed hard words at the researchers' requests, they tended to ignore the glossary while taking the sample test. As expected, students tended to list both general academic and specialized academic words as hard words (e.g., "approximation," "following," "additional," "expression"). It is also notable that some students identified words with higher frequency or part of daily use as hard. These words included "travel," "restaurant," and "climbed." These results suggest that explicit instruction of both academic and social vocabulary is needed for ELL students even in math class at the secondary level.

With respect to the language complexity rating and students' comprehension of items or students' performance on items, they were not necessarily related to each other for the

given sample items. For instance, Item 1 in the think-aloud test was relatively highly rated in its linguistic complexity, particularly for the vocabulary and grammar categories. However, students generally comprehended this item well and performed better on this item compared to other items. It may be the case that the visual image presented in the item (e.g., shaded triangle and trapezoid illustrated) provided extra cues for ELL students to comprehend and solve the item. Item 5, which was also rated high in its linguistic complexity with a higher number of academic words, was correctly solved by students compared to other items. On the other hand, students did poorly on Item 2, which was rated relatively low in its linguistic complexity and language demand (e.g., reliance score of 2). The verbal protocol analysis results revealed that not only did students have difficulty in applying an appropriate mathematical procedure, but also, the story in the problem had little contextual relevance to students' age and background. The concept of tipping at a restaurant is tied to culture, socioeconomics, and age level. This result suggests that socioeconomic background and age-appropriateness should be considered in addition to language difficulty when examining potential sources of item difficulty for ELL students.

The results of this study create a number of practical implications for policymakers and practitioners to consider in the use of accommodations. As discussed above, students' familiarity and prior experiences with a given accommodation seem to play a key role in the effects of the accommodation. In order for accommodations to be effectively and validly used, they should be part of daily classroom practice. Meanwhile, it is important to consider the students' content knowledge as well as language proficiency when providing accommodations. This study suggests that if students had little content knowledge from the beginning, providing accommodations would make little difference regardless of their language proficiency level. The study results also highlight the importance of examining OTL for ELL students in order to make more appropriate inferences about test scores.

Limitations and Future Studies

A major limitation in the present study is the small sample size, which requires the readers to be cautious in interpreting and generalizing the results. The results of the study also indicate that the effects of accommodations may be contingent upon a specific ELL population and their experiences with accommodations. Thus, future research should include a replication of this study with a different population, such as those who had experience with a glossary in testing, for instance. Students in the present study were largely from Spanish-speaking backgrounds and started school in the U.S. in early elementary grades, and were clustered at moderate to high English language proficiency levels. ELL students with different backgrounds might have yielded different results.

As mentioned earlier, future studies also should include an examination of ELL students' OTL in a content area as well as in their English language instruction. In this study, during the verbal protocol analysis, it was difficult to disentangle language difficulty from content difficulty. In other words, did students' limited English proficiency interfere with their ability to access the content of the test item, or to access content during instruction, or to articulate their knowledge to the researchers? Language is an important factor in all of the above, so the remaining question is how providing an accommodation can better support students' linguistic barriers in a math assessment. An investigation of OTL will provide valuable insight into the difficulty that ELL students have in demonstrating their content knowledge, especially in a math assessment. That is, it will offer a better understanding of how language ability is intertwined with learning mathematical content knowledge and skills. As another future study, it will be interesting to examine non-ELL students' problem-solving processes through a verbal protocol analysis in order to identify ELL-specific difficulties in tackling math items.

As in previous literature, the use of accommodations is advocated in order to increase the validity of content assessments for ELL students. The previous literature also emphasizes that the use of accommodations should be research based by providing empirical evidence on the effectiveness and validity of accommodations. Although the present study is limited to the effects of read-aloud and glossary accommodations, it offers possible sources to consider in future accommodation studies. Additionally, given students' positive perceptions about accommodations and preference for receiving accommodations, continuing efforts should be made to provide appropriate accommodations for ELL students.

References

Abedi, J. (2006). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-398). Mahwah, NJ: Erlbaum.

Abedi, J., Courtney, M., & Leon, S. (2003a). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Tech. Rep. No. 608). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Courtney, M., & Leon, S. (2003b). *Research-supported accommodation for English language learners in NAEP* (CSE Tech. Rep. No. 586). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Courtney, M., Leon, S., Kao, J., & Azzam, T. (2006). *English language learners and math achievement: A study of opportunity to learn and language accommodation* (CSE Tech. Rep. No. 702). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodation for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CSE Tech. Rep. No. 666). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (CSE Tech. Rep. No. 536). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., & 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. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Abedi, J., Lord, C., Boscardin, C. K., & Miyoshi, J. (2000). *The effects of accommodations on the assessment of LEP students in NAEP* (CSE Tech. Rep. No. 537). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Education Research, 42*(3), 359-85.

Anderson, M., Liu, K., Swierzbin, B., Thurlow, M., & Bielinski, J. (2000). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2* (Minnesota Report No. 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Bolt, S. E., & Ysseldyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. *Applied Measurement in Education, 19*(4), 329-355.

Coggins, D., Kravin, D., Coates, G. D., & Carroll, M. D. (2007). *English language learners in the mathematics classroom*. Thousand Oaks, CA: Corwin Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

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

Dale, T., & Cuevas, G. (1987). Integrating language and mathematics learning. In J. Crandall (Ed.), *ESL through content area instruction: Mathematics, science and social studies* (pp. 9-54). Englewood Cliffs, NJ: Prentice Hall.

De Corte, E., Verschaffel, L., & DeWin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology*, *77*(4), 460-470.

Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *Journal of Special Education, 40*(4), 218-229.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments*. Portsmouth, NH: RMC Research Corporation, Center on Instruction. Retrieved November 21, 2006, from http://www.centeroninstruction.org/files/ELL3-Assessments.pdf

Garcia, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly*, *26*(4), 371-391.

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

Ketterlin-Geller, L. R., Yovanoff, P., & Tindal, G. (2007). Developing a new paradigm for conducting research on accommodations in mathematics testing. *Exceptional Children, 73*(3), 331-347.

Kim, D.-H., Schneider, C., & Siskind, T. (2009). Examining equivalence of accommodations on a statewide elementary-level science test. *Applied Measurement in Education, 22*(2), 144-163.

Koenig, J. A., & Bachman, L. F. (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments.* Washington, DC: The National Academies Press.

Kopriva, R. J., Emick, J. E., Hipolito-Delgao, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11-20.

Lee, J., Grigg, W., & Dion. G. (2007). *The nation's report card: Mathematics 2007* (NCES 2007-494). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences.

Longman handy learner's dictionary of American English (new edition). (2000). Harlow, Essex, UK: Pearson Education Limited.

National Center for Research on Evaluation, Standards, and Student Testing (CRESST). (2006). *Algebra Posttest.*

Raudenbush, S., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5,* 199-213.

Rivera, C., Collum, E., Shafer Willner, L., & Sia, J.K., Jr. (2006). An analysis of state assessment policies regarding the accommodation of English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective* (pp. 1-173). Mahwah, NJ: Lawrence Erlbaum.

Rubenstein, R. N. (1996). Strategies to support the learning of the language of mathematics. In P. C. Elliot & M. J. Kenney (Eds.), *Communication in mathematics, K-12 and beyond—1996 yearbook* (pp. 214-218). Reston, VA: National Council of Teachers of Mathematics.

Seltzer, M. H. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The handbook of quantitative methods for the social sciences* (pp. 309-330). Thousand Oaks, CA: Sage.

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods, 7*(1), 3-18.

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Center for Educational Assessment Research Report. No. 485). Amherst: University of Massachusetts, School of Education.

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

U.S. Government Accountability Office. (2006). *No Child Left Behind Act: Assistance from Education could help states better measure progress of students with limited English proficiency* (GAO-06-815). Washington, DC: Author.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Monograph No. 6). Council of Chief State School Officers and National Institute for Science Education Research. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005). *The WEB alignment tool: Development, refinement, and dissemination*. Report to the Council of Chief State School Officers' State Collaborative on Assessment & Student Standards, Technical Issues in Large-Scale Assessment Collaborative.

Weston, T. J. (2003). *NAEP validity studies: The validity of oral accommodation in testing* (Working Paper No. 2003-06). Washington, DC: National Center for Education Statistics.

Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., Bachman, P. L., Chang, S. M., Farnsworth, T., Jung, H., Nollner, J., & Shin, H. W. (2008). *Providing validity evidence to improve the assessment of English language learners* (CRESST Rep. No. 738). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

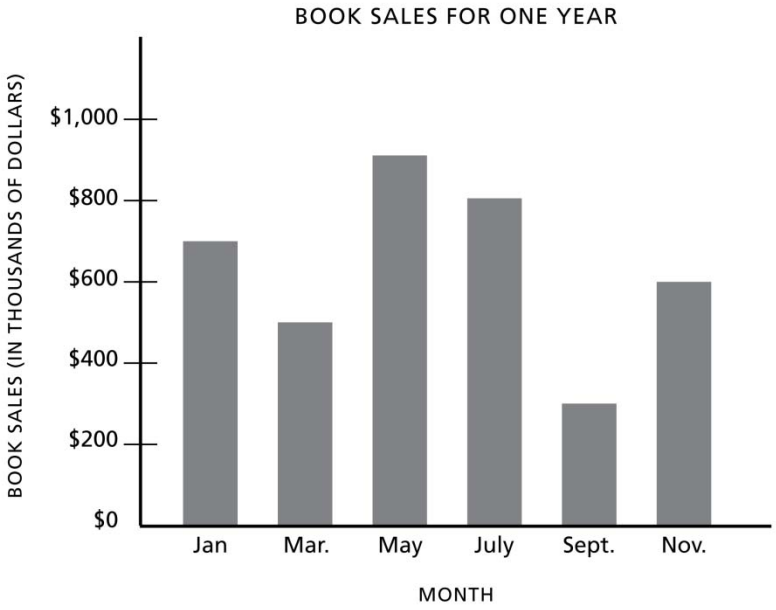
Wolf, M. K., Kao, J., Griffin, N., Herman, J. L., Bachman, P., Chang, S. M., & Farnsworth, T. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses--Practice review* (CRESST Rep. No. 732). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wolf, M. K., Griffin, N., Kao, J., Chang, S. M., & Rivera, N. (2009). *Connecting policy to practice: Accommodations in states' large-scale math assessments for English language learners* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wolf, M. K., & Leon, S. (2009, April). Identifying the test items differentially impacting performance of ELL students and the language demands of the items. In L. L. Cook & J. L. Herman (Chairs), *Validity in ELL assessment: Challenges and promising approaches.* Structured poster session presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Appendix A:
Example of Read-Aloud Script

Item as printed in students' test booklet:	*Please read this way:* *(starting with item number)*
1. In a quadrilateral, two of the angles each have a measure of 110°, and the measure of a third angle is 90°. What is the measure of the remaining angle? A. 20° B. 50° C. 90° D. 130°	In a quadrilateral, two of the angles each have a measure of [*pause*], and the measure of a third angle is [*pause*]. What is the measure of the remaining angle? [*do not read answer choices*]
2. A group of students has a total of 29 pencils and everyone has at least one pencil. Six students have 1 pencil each, five students have 3 pencils each, and the rest of the students have 2 pencils each. How many students have only 2 pencils? A. 4 B. 6 C. 8 D. 9	A group of students has a total of twenty-nine pencils and everyone has at least one pencil. Six students have one pencil each, five students have three pencils each, and the rest of the students have two pencils each. How many students have only two pencils? [*do not read answer choices*]

3. The graph below shows the amount of bi-monthly sales at a local book store during one year.  Based on the graph, which is the closest ESTIMATE of the range in the amount of monthly sales at the store during the year? A. $200,000 B. $300,000 C. $500,000 D. $600,000	The graph below shows the amount of bi-monthly sales at a local book store during one year. Across the top it says, book sales for one year. Along the side it says book sales in thousands of dollars. On the bottom it says month. January, March, May, July, September, November. Based on the graph, which is the closest estimate of the range in the amount of monthly sales at the store during the year. [*do not read answer choices*]

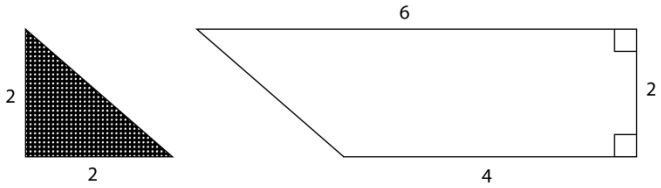
Appendix B:
Glossary Terms Used in Math Test

Original Word	Glossary Definition
above	on top
additional	extra
amount	how much
closest approximation	best guess; nearest amount
at least one	one or more
bake	to cook
baked goods	cakes, cookies, and bread
Bakery	a place where bread is baked and sold
bi-monthly	every other month
bought	buy (past tense)
calculate	find
charges	asks for money; bill
check	a bill
climbed	walked up a mountain
combine	put together
company	a business
consisting	made up of
contains	holds
cost	price; how much money
customers	people who buy things
deliver	take to people's houses
describes	shows
drawer	a box
drawn	make with a pencil
elevation	how high
equivalent	the same as
explain	give a reason
fee	price or cost; how much money
figure	a picture
is given by	is seen in
graph	a drawing used in math
growth	getting bigger
hikers	people who walk in mountains
local calls	phone calls to near places
long-distance calls	phone calls to far away places
measuring	finding the size or amount
most likely	probably
nearest	closest
newspapers	papers printed with news

Of the following	from the choices
on sale	selling
oven	a thing used for cooking or baking
plant	a living thing with roots and leaves
plumber	a person who fixes things
price	cost; how much money
record	put or copy music
remaining	what's left over
renting	paying money to use something
represents	stands for
Salt Flats	name of a place
selecting	choosing
shaded	darker or filled in
shifted	moved
shown	seen
spent	paid or used
Talon Bluff	name of a place
the rest	left over
tip	money for the waiter or waitress
travel	going somewhere
treats	candies
true	correct or right

Appendix C:
The Five Think-Aloud Items

1.



shaded: darker or filled in

above: on top

How many triangles of the shape and size of the shaded triangle can the trapezoid above be divided into?

A. 3

B. 4

C. 5

D. 6

Source:
1995 TIMSS, Population 2, Item R-10
Previously used in Abedi et al. (2003b).

Standard/Objective:
geometry

Student Performance:
TIMSS: International average: 52%

In the present study:
A – 23.0%
B – 26.3%
***C – 39.3%**
D – 11.5%

2. Of the following, which is the closest approximation of a 15 percent tip on a restaurant check of $24.99?

    A. $2.50
    B. $3.00
    C. $3.75
    D. $4.50

Of the following: from the choices

closest approximation: best guess; nearest amount

tip: money for the waiter or waitress

check: a bill

Source:
1996 NAEP, Grade 8, Item 5
Previously used in Abedi et al. (2003b).

Standard/Objective:
Number sense and operations

Student Performance:
In NAEP: 37.7% of students answered it correctly.

In the present study:
A – 21.7%
B – 25.0%
***C – 38.0%**
D – 15.3%

3. A group of students has a total of 29 pencils and everyone has at least one pencil. Six students have 1 pencil each, five students have 3 pencils each, and the rest of the students have 2 pencils each. How many students have only 2 pencils?

at least one: one or more

the rest: left over

E. 4
F. 6
G. 8
H. 9

Source:
1995 TIMSS, Population 2, Item R-11
Previously used in Abedi et al. (2003b).

Standard/Objective:
algebra

Student Performance:
TIMSS: International average: 47%

In the present study:
***A – 36.0%**
B – 12.1%
C – 34.3%
D – 17.6%

4. A group of hikers climbed from Salt Flats (elevation -55 feet) to Talon Bluff (elevation 620 feet). What is the difference in elevation between Talon Bluff and Salt Flats?

   A. 565 feet
   B. 575 feet
   C. 665 feet
   D. 675 feet

hikers: people who walk in mountains

elevation: how high

climbed: walked up a mountain

Salt Flats: name of a place

Talon Bluff: name of a place

Source:
2003 California Standards Test, Grade 6, Item 25
Previously used in Abedi et al. (2006).

Standard/Objective:
Number sense

Student Performance:

In Abedi et al. (2006):
n = 2,354
A – 55.3%
B – 15.0%
C – 6.8%
***D – 22.9%**

In the present study:
A – 46.7%
B – 12.8%
C – 8.5%
***D – 32.0%**

5. A plumber charges customers $48 for each hour worked plus an additional $9 for travel. If h represents the number of hours worked, which of the following expressions could be used to calculate the plumber's total charge in dollars?

A. $48 \times 9 \times h$

B. $48 + (9 \times h)$

C. $(48 \times 9) + h$

D. $(48 \times h) + 9$

plumber: a person who fixes things

charges: asks for money; bill

customers: people who buy things

additional: extra

travel: going somewhere

represents: stands for

of the following: from the choices

calculate: find

Source:
1996 NAEP, Grade 8, Item 9
Previously used in Abedi et al. (2003b).

Standard/Objective:
Algebra and functions

Student Performance:
In NAEP: 57.7% of students answered it correctly.

In the present study:
A – 8.0%
B – 19.2%
C – 11.3%
***D – 61.5%**