

Smarter Balanced Assessment Consortium:

Summary of Literature on Empirical Studies of the Validity and Effectiveness of Test Accommodations for ELLs: 2005–2012

Prepared for Measured Progress by
The George Washington University
Center for Equity and Excellence in Education
Maria Pennock-Roman and Charlene Rivera

March 2012



THE GEORGE WASHINGTON UNIVERSITY
CENTER FOR EQUITY AND
EXCELLENCE IN EDUCATION

Summary of Literature on Accommodations for ELLs

Table of Contents

Summary of Literature on Accommodations for ELLs	3
Summary of Literature	3
Overview of Findings	3
Translation, Dual Language, and Bilingual Glossaries.....	3
English Language Accommodations: English Glossary	11
English Language Accommodations: Plain English	12
Findings on Other Accommodations	14
Conclusions	15
Methodological approaches.	15
How the Literature Search Was Conducted: Detailed Methodological Notes	16
Scope.....	16
Search Terms by Subject.....	16
Other Restrictions	17
References	21
Appendix A: Annotated Bibliography	24

Smarter Balanced Assessment Consortium:

Summary of Literature on Accommodations for ELLs

Summary of Literature

The goal of the literature summary is to review studies of the validity and effectiveness of test accommodations for English language learners (ELLs) not included in the Pennock-Roman and Rivera (2011) meta-analysis of mean effects. The former meta-analysis included 14 studies through 2005. The current, updated literature review identified articles related to the use of test translation and dual language, glossary and dictionary accommodations, and linguistically simplified English versions of tests.

The summary provides an analysis of testing accommodations for ELLs from both the Pennock-Roman and Rivera (2011) meta-analysis of mean effects and the updated annotated bibliography of accommodation studies. The summary concludes with a discussion of the methodology for conducting the literature review. The annotated bibliography and primary variables for the studies are included in Appendix A.

Overview of Findings

Translation, Dual Language, and Bilingual Glossaries

Findings from Pennock-Roman and Rivera (2011).

Among the 14 experimental studies summarized in the meta-analysis, several included translated or dual-language written versions of tests or bilingual glossaries but none evaluated translations of test directions that were read aloud. Figure 1 provides descriptive information about the 14 experimental studies and accommodations examined in the meta-analysis.

Figure 2 shows the range and variety of individual effect sizes of accommodations examined in the meta-analysis. In the inside-out display, the first number in each cell identifies the study number from which the effect size was calculated. The numbers correspond to the order of studies as shown in Figure 1. Studies with independent samples are listed first in alphabetic order by author, followed by the two repeated-measures designs.

As shown on Figure 2, the highest individual effect sizes by far among all accommodations were found for *Spanish versions* of tests administered to ELLs who were low in English proficiency (+1.45, Aguirre-Muñoz, 2000) or who had received instruction in Spanish for the content area (+0.95, Hofstetter, 2003). On the other hand, the effect sizes were negative and smaller in absolute value for ELLs at intermediate levels of English proficiency (EP) or who had received instruction in English. Specifically the values were -0.02 and -0.11 for ELLs with low intermediate or high intermediate EP, respectively (Aguirre-Muñoz, 2000) and -0.34 for ELLs receiving content instruction in English (Hofstetter, 2003). Hence, the effectiveness of written versions in the native language is very sensitive to students' language proficiency in English and to their literacy skills and content knowledge in the native language. This result implies that written versions of native language tests are effective only for ELLs who have literacy skills in the native language and who are familiar with the content vocabulary in their native language. If administered to an unselected group of ELLs who vary in terms of native language skills, one would expect near zero effect sizes because the positive effects for one subgroup would be cancelled by the negative effects for students who know the

Summary of Literature on Accommodations for ELLs

content knowledge vocabulary better in English. Also, effect sizes for native language accommodations should not be averaged across

Summary of Literature on Accommodations for ELLs

Study #	Author(s)	Report date	Accommodation	Sample Size per Accommodation & Grade				Content	Source	Grade(s)		Design
				ELLs		Non-ELLs						
1	Abedi, Courtney, & Leon, #586	2003b	English Dictionary/G.	64	86	93	87	Math	NAEP & TIMSS	4	8	Random Assignment
			Pop-up English Glossary	35	84	44	68					
			Extra Time	89	0	84	0					
			Small Group Testing	11	0	9	0					
			No Accommodation	80	86	98	131					
2	Abedi, Courtney, & Leon, #608	2003a	English Dictionary/G.	270	206	247	241	Science	NAEP & TIMSS	4	8	Random Assignment
			Plain English	284	209	257	241					
			Bilingual Glossary	135	119	101	129					
			No Accommodation	268	199	241	245					
3	Abedi, Courtney, Mirocha, Leon, & Goldberg	2005	English Dictionary/G. + Time	59	23	62	36	Science	NAEP	4	8	Random Assignment
			Plain English + Time	20	11	23	0					
			Bilingual Glossary + Time	64	16	0	0					
			Extra Time	62	22	85	33					
4	Abedi, Hofstetter, Baker, & Lord	2001	English Dictionary/G.	146		121		Math	NAEP	8		Random Assignment
			Plain English	124		117						
			English Dictionary/G. + Time	29		30						
			Extra Time	30		25						
			No Accommodation	144		130						
5	Abedi, Lord, & Hofstetter ^a	1998	Plain English	117		166		Math	NAEP	8		Random Assignment
			Spanish Version	15		0						
			No Accommodation	115		145						
6	Abedi, Lord, Kim, & Miyoshi	2001	English Dictionary/G.	55		82		Science	NAEP	8		Random Assignment
			Bilingual Glossary	70		75						
			No Accommodation	58		79						
7	Aguirre-Munoz	2000	Plain English	41	44	82	89	History, Aztec task	Based on CA standards	7 (Levels of proficiency in English defined 4 groups)		Random Assignment Classrooms
			Dual Language	26	25	37	0					
			Spanish Version	77	30	25	0					
			No Accommodation	46	52	61	82					
8	Anderson, Liu, Swierzbis, Thurlow, & Bielinski	2000	Dual Language	53		0		Reading	Based on MN standards	8		Random Assignment (Power Test) ^c
			No Accommodation	52		101 ^b						
9	Duncan, Parent, Chen, Ferrara, Johnson, Oppler, & Shieh	2005	Dual Language	127 ^b		74 ^d	0	Math	NAEP	8		Random Assignment
			No Accommodation	0		119 ^d	82 ^{b, e}					

^a In its original form, this study partially overlapped with the Hofstetter study (2003), but only non-overlapping results are included here.

^b An effect size could not be calculated for these groups because there was no corresponding accommodation or control group.

Figure 1. Descriptive Information for 14 Experimental Studies in Meta-Analysis by Pennock-Roman and Rivera (2011)

Summary of Literature on Accommodations for ELLs

Study #	Author(s)	Report date	Accommodation	Sample Size per Accommodation & Grade		Content	Source	Grade(s)		Design		
				ELLs	Non-ELLs							
10	Hofstetter ^a	2003	Plain English	6 ^f	222 ^g	82 ^g	Math	NAEP	8	Random Assignment		
			Spanish Version	63 ^f	147 ^g	24 ^g						
			No Accommodation	9 ^f	229 ^g	61 ^g						
11	Kopriva, Emick, Hipolito-Delgado, & Cameron	2007	Pop-up Bilingual Glossary (BG)	36		None	Math	Based on SC standards	3 & 4	Random Assignment		
			Picture Dictionary (PD)	36		None						
			Read Aloud (RA)	33		None						
			RA + PD	29		None						
			RA + BG	33		None						
			BG + PD	37		None						
			No Accommodation	33		None						
12	Rivera & Stansfield	2004	Plain English	34	15	1412	1368	Science	DE State Assessment	4	6	Random Assignment through spiraling of booklets
						1416	1423					
			No Accommodation	38	22	1430	1415					
						1426	1416					
13	Abedi, Lord, & Plummer (Accuracy Study)	1997	Plain English	471		546	Math	NAEP	8	Counterbalanced Repeated Measures (Power Test) ^c		
			No Accommodation	471		546						
14	Albus, Thurlow, Liu, & Bielinski	2005	English Dictionary/G.	133		69	Reading	Based on MN standards	Middle school	Counterbalanced Repeated Measures (Power Test) ^c		
			No Accommodation	133		69						

^c Designs with power tests either had no time limits or had very generous time limits, allowing all items to be attempted.

^d These groups had a Spanish-language background but were considered non-ELLs because they had had three or more years of instruction in English and therefore met the criteria for inclusion in NAEP assessments.

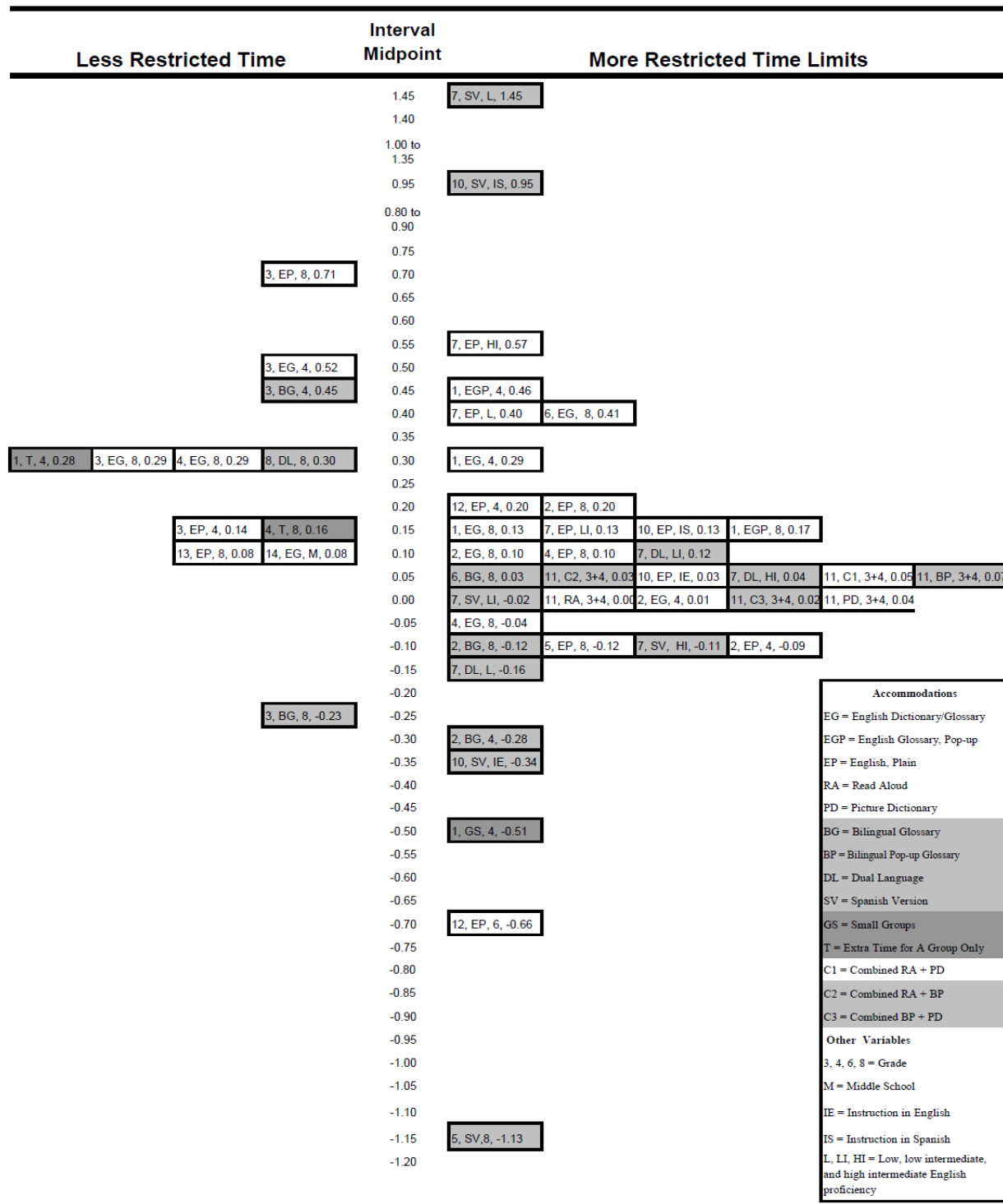
^e These non-ELLs were native speakers of English who had no background in Spanish.

^f These students attended basic 8th grade mathematics classes taught in Spanish.

^g These students attended a variety of mathematics classes taught in English..

Figure 1. Descriptive Information for 14 Experimental Studies in Meta-Analysis by Pennock-Roman and Rivera (2011)

Summary of Literature on Accommodations for ELLs



Note : Each cell above first includes the study number, followed by the accommodation type, other variables, and the effect size d , in that order. Level of proficiency in English or language of instruction is noted when available rather than grade level due to available space on the figure.

Figure 2. Inside Out Display of Glass's Unbiased d Values for Accommodations Administered to ELLs

Summary of Literature on Accommodations for ELLs

ELL groups heterogeneous in language background; to do so results in a near-zero effect size that obscures the interaction between language proficiency and effectiveness of translated versions of tests.

Dual language (DL) tests were found to be very sensitive to the generosity of the time allotted to students to negotiate a test booklet that is often double in size as compared to the original, English-only test booklet. When the original and dual language versions were power tests with essentially unlimited time, the effect size was 0.30 (Anderson, Liu, Swierzbins, Thurlow, & Bielinski, 2000). In contrast, when time limits were constrained and identical to the time allotted for the original test booklet, the effect size was essentially zero for ELLs (an average of 0.003 over three results) and negative for non-ELLs (an average of -0.169 over four results). ELL groups in these studies, were undifferentiated by level of EP, literacy in their native language, and language of instruction. One can speculate that perhaps even higher effect sizes could be found if ELLs low in EP who had received recent content instruction in Spanish were separated from the general group of ELLs, randomly assigned to the DL vs. original test versions, and administered their designated test form with essentially unlimited time.

Like the dual-language test, the paper and pencil *bilingual glossary* accommodation was very sensitive to the generosity of time limits available for testing. Specifically, under restricted conditions the average of three effects for ELLs with paper and pencil Spanish-English glossaries was -0.176 . For non-ELLs, the average of three effect sizes was -0.134 under restricted time conditions. Hence the bilingual glossary accommodation was more difficult for both ELLs and non-ELLs when time limits were constrained. In contrast, Kopriva, Emick, Hipolito-Delgado, and Cameron (2007) found an effect of 0.069 for a pop-up, computer administered version of a Spanish-English glossary accommodation. The larger size of the latter effect could be the result of having a more convenient, time-efficient format with computer administration. In fact, when the paper and pencil versions of Spanish-English glossaries were administered under generous time limits, the average was higher (0.247 mean of two effects for ELLs) than for the pop-up version with restricted time. Taken together, these results imply that both ELL and non-ELL students need generous time limits and/or a time efficient computerized format to utilize bilingual glossaries effectively. Cormier, Altman, Shyyan, and Thurlow, (2010) and Thompson, Blount, and Thurlow (2002) found that generous time limits are typical among state assessments as a way to offer students more opportunity to demonstrate what they know. Hence, the larger effect sizes are probably more representative of the potential effects under actual test conditions.

Findings from Robinson (2010).

Robinson found clear evidence of greater validity of the *Spanish version* of an oral, individually administered mathematics test for students in kindergarten and first grade who had Spanish as a home language and were classified as having low EP. The regression findings demonstrated that the English version was more related to English proficiency and less related to teachers' ratings of students' knowledge about mathematics concepts, particularly in the fall semester of Grade 1. He found that the effect sizes measuring the superiority of the Spanish version for the low EP group and for students near the cut-off value of EP grew from statistically non-significant values in the fall of kindergarten to an effect size with an absolute value of 1.22 in the spring of first grade. These results are in agreement

Summary of Literature on Accommodations for ELLs

with Aguirre-Muñoz (2000). On the other hand, Robinson found that the match between home language and test version was more important than the match between language of instruction and test version for the early grades, contrary to Hofstetter (2003) who studied 8th graders. This discrepancy can be reconciled taking into account grade level. Often, for ELL students in late elementary, middle, or high school there is a decline of native language proficiency over several grade levels and limited development of literacy skills in their native language owing to the rarity of dual language programs in most U.S. schools. The small, exceptional group of students receiving instruction in Spanish in Hofstetter's study was able to maintain their literacy and content knowledge skills in Spanish, unlike those receiving instruction in English through 8th grade. However, in Robinson's study, kindergartners and first graders had one year or less to lose skills in the language they spoke at home and were tested orally so they were not required to read Spanish.

Implications for policy.

The study strongly supports the use of native language assessment for students with low fluency in English, especially in the early grades where most of the students have received instruction at home in their native language. In first grade and kindergarten, matching the language of the test to the home language appears to be much more important than matching it to the language of instruction for students at lower EP levels. However, the language of instruction can be expected to be much more salient in later school years for students who have resided in the U.S. during the course of their schooling. For recent immigrants, one would expect that the language of the test should match the language of their schooling in the home country if they are literate in that language or the home language if not literate in that language.

Implications for implementation.

The procedures for test translation and standardization of oral administrations used here were exemplary and should serve as a model for future studies.

Findings from Abedi, Courtney, Leon, Kao, & Azzam (2006).

In contrast, there was no evidence of any benefit from a written *dual language* test in the hierarchical linear modeling results by Abedi, Courtney, Leon, Kao, & Azzam (2006) after controlling for class-level variables related to opportunity to learn. However, the implementation of this accommodation was less than ideal owing in part to the limitations imposed by field conditions in the schools. In particular, it was administered to an undifferentiated group of 8th grade ELLs, who most likely had been instructed primarily in English for many years and were unlikely to be literate in Spanish or to know subject specific terms in Spanish. Although most ELLs in this study had Spanish as a home language, this characteristic varied by school, and some who received the dual language booklet may not have had Spanish as their home language. Furthermore, the dual language version, which has a booklet double in length compared to the original version, did not allow extra time.

Summary of Literature on Accommodations for ELLs

Implications for policy.

This is the first accommodation study to show unequivocally how much opportunity to learn impacts students' performance regardless of the efforts to improve test validity with test accommodations. Although test accommodations may reduce construct irrelevant variance in the measurement of achievement, they cannot be considered the panacea for closing the achievement gap between native speakers of English and language minority students in the schools. There is unequal opportunity to learn for ELLs that can account for much of the achievement gap. Closing the gap will require improving the quality of the schooling that ELLs receive. As demonstrated in one example using data from High School and Beyond in the textbook by Bryk and Raudenbush's (1992, pp. 103-113), there is much variation among schools. The authors demonstrated that a subset of effective schools was able to raise achievement for at risk students and reduce the impact of students' socioeconomic status and prior student achievement on subsequent student test scores.

Implications for implementation.

In order for the dual language accommodation test form to be effective, it has to be assigned to a particular group of ELLs—specifically students with literacy skills and instructional experience in the native language of the test. Owing to its extra length, it must be tested with sufficient time (preferably no time restriction). To evaluate the effectiveness of a Spanish-English dual language test, the accommodated and original versions need to be randomly assigned within this particular group of ELLs literate in Spanish and have both forms administered with very generous time limits.

It is no accident that the Robinson (2010) and Abedi et al. (2006) studies differ in results, not only because of differences in implementation (oral vs. written) but because of the difference in grade level (first two grades vs. 8th grade), as explained above. The native language versions of tests are more useful for students for whom the native language is still the dominant form of communication, such as young language minority children in the earliest grades. Hence, an oral administration of a native language test may be particularly appropriate in the first two grades. For recent immigrant ELLs and/or those in dual language programs in later grades, a written, translated version of the test or a dual language test format would be useful.

Findings from Young, Cho, Ling, Cline, Steinberg, and Stone (2008).

Young, Cho, Ling, Cline, Steinberg, and Stone (2008) examined the construct validity and factor structure of individual items on accommodations involving orally translated test directions or bilingual dictionaries in an operational test administration for a state's accountability purpose. They found that, "There was little evidence of differential test validity in terms of internal test structure or item functioning ...when the performance of non-ELLs and ELLs were examined and compared" (p. 190).

Implications for policy.

Scores from accommodations involving orally translated directions or bilingual glossaries for ELLs have some of the same psychometric properties as scores from the original versions of these tests. However, without an examination of the concurrent or predictive validity of accommodations in relation to

Summary of Literature on Accommodations for ELLs

external criteria these analyses are not sufficient to establish the score comparability of accommodated and original versions of tests.

Implications for implementation.

None

These analyses are quite different from, and complementary to, the contrast in mean effects carried out by Pennock-Roman and Rivera (2011). Whereas the factor analytic and DIF analyses approaches examine construct validity from the perspective of the internal factor structure or internal consistency among items, the mean effects approach considers test difficulty and the similarity of the scale for test scores near the mean. If there is a non-trivial positive improvement in the means for ELLs (mean for the accommodated test higher than the mean for the original test) and no change in means for non-ELLs, then one can claim that the accommodation has higher construct validity for ELLs. That is, the higher mean for ELLs with the accommodation is evidence that it improves access to the content of the test for ELLs without changing overall test difficulty for non-ELLs.

The mean effects approach can give a very different result than one found using DIF or factor analysis. For example, it is possible for two tests measuring the same construct to have similar factor structures and DIF results yet have unequal true score means—that is why raw scores from alternate forms of tests developed in large scale testing programs must be placed on a common scale through the process of equating. Although the mean effects analyses by Pennock-Roman and Rivera (2011) showed that, under restricted time conditions, the DL and bilingual glossaries were actually more difficult than their respective original versions, it is still possible that these tests may have had the same factor structure or the same DIF results as the original versions. Moreover, two tests with similar factor structures or similar means may have different regressions in predicting an external validity criterion. The four approaches—factor analysis, DIF analysis, mean effects analysis, and regression analysis for predicting an external criterion—are necessary for a comprehensive evaluation of the validity of test accommodations because each provides unique information.

English Language Accommodations: English Glossary

Findings from Pennock-Roman and Rivera (2011).

The results for English language dictionaries and glossaries were analogous to those found for bilingual glossaries. When administered with restricted time limits, only the pop-up version had a significantly nonzero average effect size value for ELLs (0.285 $p < .05$, based on two effect sizes) compared with other English dictionary conditions (0.085, $p > .05$, an average of six effects). The average effect size for paper and pencil versions was significantly different from zero among ELLs only when extended time was available for both accommodated and control groups (0.229 $p < .05$, an average of three effects). For non-ELLs, the average effect sizes for paper and pencil versions were close to zero regardless of time conditions (−0.004 when time was constrained, an average of six effects, and 0.018 when time limits were generous, an average of three effects). The average of two effect sizes for non-ELLs using the pop-up computer-delivered English glossary was slightly higher but still essentially zero (0.032). All of these

Summary of Literature on Accommodations for ELLs

studies used groups of ELLs that were *not* separated by level of EP; one would expect higher effect sizes among students at intermediate levels of EP as compared with lower effects for students with low EP.

Findings from Wolf, Kim, Kao, & Rivera (2009).

Wolf et al. studied an English language glossary accommodation in which the definitions of selected words were provided in the margins of the test pages. This was an experimental test administration and it is not clear to what extent ELLs had sufficient time to complete the problems. Their design included two approaches—a quantitative experimental design and a qualitative verbal protocol analysis with 50 students. Their hierarchical linear modeling analysis was truly enlightening because they included the effects of the accommodation together with interactions between test form and EP as well as interactions between test form and prior student knowledge. The effect of the accommodation was detectable as an interaction—that is, positive effects of the accommodation were seen only for students having sufficient prior knowledge. Also, the verbal protocol analyses suggested that the English glossary was hardly used owing to students’ lack of familiarity with the format. The small effects for this accommodation for the general group of ELLs are consistent with past studies having restricted time for experimentally developed tests. Perhaps the effects would have been larger under untimed conditions. This study underscores how important it is to take into account students’ prior knowledge when analyzing mean effects for accommodations.

Implications for policy.

The observed interaction effect between students’ prior knowledge and accommodation support the validity of this approach. It suggests that the effectiveness of the accommodation cannot be judged in circumstances where students have not had the opportunity to learn the test material. No matter how much construct irrelevance variance may have been reduced in the accommodation; if the students have not learned the material, one cannot expect a noticeable improvement in test performance.

Implications for implementation.

It is clear that students will profit more from a glossary accommodation if they have sufficient training and experience with the method.

English Language Accommodations: Plain English

Findings from Pennock-Roman and Rivera (2011).

This accommodation has sometimes been called *linguistically simplified English* or *linguistically modified English* or *plain English* versions of tests. The average of 11 effect size values for this accommodation administered to ELLs under constrained time limits was 0.053 in contrast to an average of 0.108 (3 values) when both the original and accommodated versions had generous time limits. Smaller effect sizes were found for non-ELLs: an average of –0.008 for 10 results under restricted time conditions vs. an average of 0.064 for two values under generous time conditions. There was statistically significant variation among effect sizes across studies that may have been associated with variation in EP among samples or in the characteristics of the particular test that underwent simplification. Aguirre-Muñoz (2000) found higher effect sizes for this accommodation among subgroups having intermediate levels of

Summary of Literature on Accommodations for ELLs

EP as compared with low EP. Kiplinger, Haug, and Abedi (2000), and Albus, Thurlow, Liu, and Bielinski (2005) found a significantly higher test performance for the accommodated groups vs. unaccommodated groups at an intermediate level of EP, whereas no effects were observed at low levels of EP. Pennock-Roman and Rivera stated that:

It is also possible that the quality of the implementations of the accommodation varied by study. Alternatively, the original test booklets may have varied in grammatical complexity, and some original items may already have had a reduced language load thereby benefitting little from more simplification (p. 20).

That is, effect sizes based on original tests that were already reduced in language load would be smaller than those based on those with greater grammatical complexity.

Findings from Abedi et al. (2006) and Sato, Rabinowitz, Gallagher, and Huang (2010).

The authors examined the effects of linguistic modification on test performance. Whereas Abedi et al. found no significant effects after controlling for class-level opportunity-to-learn variables, Sato et al. reported one of the largest effects (0.16) found in the literature so far for this accommodation type in a general group of ELLs. Abedi et al. found no interaction between students' reading level and the effectiveness of the technique, whereas Sato et al. did find a pattern of differential effects by reading level for non-ELLs. The discrepancy between the two studies is most likely due to the percentage of items that underwent modification (28% for Abedi et al. vs. 100% for Sato et al.). By design, the original items for Sato et al.'s analyses were deliberately selected for being language-intensive. Although the largest individual effect sizes for linguistic simplification have occurred for ELL groups having an intermediate level of EP (e.g., 0.57 by Aguirre-Muñoz in Pennock-Roman and Rivera's meta-analysis), neither Abedi et al. or Sato et al. reported the effects of the accommodation specifically for ELLs having an intermediate level of EP. Hence, both studies may have underestimated the effects of linguistic simplification because they implemented the technique with a heterogeneous group of ELLs; consequently, the method is more likely to be effective for students with high or intermediate EP.

Implications of the Sato et al. study for policy.

The results of the study suggest that linguistic modification of test items does have potential as a viable accommodation, contrary to the conclusion of Kieffer et al (2009). The improvement in the reduction of construct irrelevant variance owing to linguistic simplification can be substantial if there are many language-intensive items in the original test.

Implications of the Sato et al. study for Implementation. The process of development and refinement of the linguistically modified items here was exemplary and should provide a model for future studies and test development. Owing to the very small number of studies that examined the effects of linguistic modification for ELLs of intermediate EP, the hypothesized interaction between EP and the effects of linguistic modification needs to be confirmed in future research. The trend in past data suggests that

Summary of Literature on Accommodations for ELLs

effects for linguistic modification will be larger if (1) ELLs have an intermediate level of EP and (2) the original version of the test has a large proportion of linguistically demanding items (as did the original test here).

Findings on Other Accommodations

Wolf et al. (2009) examined the read aloud condition that showed more positive effects than did the English glossary condition, perhaps because students were more familiar with it or perhaps because this condition was administered with more generous time limits. There was only one instance of the read aloud accommodation (Kopriva et al., 2007) among studies reviewed by Pennock-Roman and Rivera (2011) and the effect size for it was essentially zero.

Mann, Emick, Cho, and Kopriva (2006) evaluated the concurrent validity of language liaison and read aloud accommodations together with other techniques in terms of predicting teacher's ratings of student knowledge. Their findings showed inconsistent trends, but in general, both the accommodated and unaccommodated versions did not show particularly high validity. Nevertheless, the language liaison approach was novel and intriguing (see Primary Variables spreadsheet for description). The importance of this study was its use of concurrent validity criteria—in this case, teacher's ratings of student knowledge. More research on accommodations should take into account validity criteria external to the test.

Summary of Literature on Accommodations for ELLs

Conclusions

Although there are still many gaps in the research on accommodations for ELLs, the emerging evidence supports the view that practitioners must tailor the choice of accommodations for ELLs to the students' proficiency skills in English and their native language together with their instructional history. Results so far suggest that native language accommodations are more suitable for students for whom the native language is still the dominant form of communication and for whom subject-specific terms are more familiar in the native language. On the other hand, English language accommodations may be more suitable for students with an intermediate level of EP and knowledge of subject-specific terms in English.

The effectiveness of accommodations interacts not only with language background but also with students' prior knowledge and opportunity to learn. Even if all sources of construct irrelevant variance due to language are removed from a measure of content knowledge, students who have not been instructed in the concepts of the test will still demonstrate a large achievement gap, and the accommodation will appear ineffective.

Students are likely to benefit more from accommodations with which they are familiar. Training in the use of particular types of accommodations may improve the effectiveness of test accommodations.

Methodological approaches.

Recent studies show an encouraging trend towards a greater variety and sophistication of approaches for examining the validity and effectiveness of accommodations and state assessments of ELLs. Several studies applied confirmatory factor analysis and DIF analyses (Sato et al., 2010; Steinberg, Cline, Ling, Cook, & Tognatta, 2009; Young et al. 2008; Young, Holtzman, & Steinberg, 2011;); some examined concurrent validity with teachers' ratings of student knowledge (Mann et al., 2006; Robinson, 2010) or another mathematics test (Sato et al., 2010); two applied hierarchical linear modeling (Abedi et al., 2006; Wolf et al., 2009); and two studies applied a mixed methods approach including qualitative analyses of student performance that improved test development or clarified the findings (Sato et al., 2010; Wolf et al., 2009). The control for prior student knowledge and opportunity-to-learn variables and the inclusion of interaction terms in the hierarchical linear models were very enlightening in terms of explaining why accommodation effects are frequently so small in relation to large, pre-existing achievement gaps. The discontinuous regression approach applied by Robinson (2010) appears to be a technique that is ideally suited for situations where it would be unfeasible to randomly assign students to alternate language versions of a test.

The confirmatory factor analytic approach has mostly been used to compare the dimensionality of items within one particular test across groups. While this information is valuable, none have included English proficiency or reading level together with math or science content in a factor analysis. Such an analysis would be valuable to see if the accommodated vs. unaccommodated versions of math and science tests have a smaller relationship to construct irrelevant sources (e.g., reading or EP).

Summary of Literature on Accommodations for ELLs

How the Literature Search Was Conducted: Detailed Methodological Notes

Scope

Whereas the focus of the meta-analytic review was on the contrast between mean effect sizes for accommodations vs. no accommodations in experimental studies for ELLs, the current summary was expanded to include other data-based approaches to evaluate the construct validity of accommodations. Specifically, quantitative studies that considered differential item functioning (DIF) and correlational approaches such as hierarchical linear models (HLM) and confirmatory or exploratory factor analysis (CFA and EFA, respectively) in both experimental and non-experimental studies were included. The 2011 meta-analysis was restricted to studies for ELLs and non-ELLs among U.S. students in grades K-12; that restriction remains in effect for the quantitative studies selected for detailed summary in the current update. However, abstracts from other studies on topics

related to validity and fairness in assessment for ELLs that emerged in the search but did not meet the criteria for detailed summary were separated from the excluded abstracts and classified by topic, as explained below.

Search Terms by Subject

In the initial passes through the ERIC EBSCO database to find new citations, we experimented with several combinations of search terms using all available years to see which combination would identify all or nearly all the papers already included in the meta-analysis. Experimentation with various terms yielded the following results:

- *Test accommodations* produced a search that was too narrow; in contrast, changing it to the singular *accommodation* dramatically increased the number of target articles.
- *State assessment* and variations such as *accommodating* or *accommodate* were terms that did not completely overlap with *accommodation* and were valuable in identifying additional related studies.
- *Test* or *assessment* without the words *state* or *large-scale* or *high-stakes* produced too many irrelevant papers not associated with the measurement of standardized achievement tests. For example, it produced too many experimental studies of teaching and learning not concerned with fairness in assessment per se. Also, in connection with *translation* there were more than 900 articles, most of them reporting research outside the U.S.
- Although many researchers have replaced the term *Limited English Proficient* (LEP) students with the term *English language learners* (ELL or ELLs) in the literature, the two terms do not completely overlap. To include the studies referring to this group, one must include both search terms or risk overlooking a large proportion of the studies. In addition, some researchers use *English learners*. Hence the most comprehensive search was applied using the following Boolean combination of subject terms:
 - (English language learners or English learners or limited English) AND
 - (accommodat* or state assessment or large scale assessment or high stakes test*
 - or large scale test*)

Summary of Literature on Accommodations for ELLs

The asterisk after accommodate or test enabled related words to also be searched (e.g., accommodated, testing). In addition, searches were included for particular types of accommodations such as dual language, glossaries, and linguistic modification as shown in Table 1. While these yielded some related studies outside the U.S., they did not produce any additional target articles that met the criteria specified above.

Other Restrictions

Once the comprehensive combination of search terms was identified, the search was restricted by year (2005-2012) and by document type—Eric documents or academic journals (ED or EJ, respectively)—thereby excluding magazines such as *Education Week*. The category of Eric documents includes books, educational reports, technical reports, and papers at conferences. Restricting the years to 2005 and later largely omitted articles that we had cited in Pennock-Roman and Rivera (2011) or had screened previously as having little relevance to the topic of that article.

A total of 257 papers (129 Eric documents and 128 academic journal articles) were identified with the initial search criteria and another 80 were identified by the specific accommodation search terms. The latter search contributed some new abstracts for Category B (studies outside the U.S.) but the rest were either repeats of the larger search or contained references not relevant to validity or effectiveness of accommodations. ERIC results were classified into two groups based on their degree of relevancy to issues of validity or effectiveness of accommodations for ELLs as judged from their titles and abstracts. The excluded, less relevant category comprised papers such as NAEP report cards, descriptive surveys of available accommodations for ELLs, etc. The more relevant category was further subdivided into the following groups:

- A. Quantitative studies directly evaluating issues of validity or effectiveness of test scores on subject content tests with accommodations in grades K-12 in the U.S.
 - A1. Accommodations using translation, glossaries, or linguistic simplification (target articles)
 - A2. Other accommodations or state assessments that may include accommodations but were not specified
- B. Quantitative studies with the dependent variable being test scores in locations outside the U.S. or outside the context of students in grades K-12, or with other types of tests such as measures of English language proficiency.
- C. Qualitative studies on students' read aloud processing of test items or attitudes/perceptions about accommodations; also teachers' perceptions and practices concerning tests with accommodations.
- D. Recommendations for validity framework, improvement of test design, assessment policies or practices.

Summary of Literature on Accommodations for ELLs

The papers in Groups B-D¹ were judged to have some value in providing background on the construct validity of tests using accommodations in content area tests with students in grades K-12 in the U.S. However, only studies in Group A were included in the detailed summary that contains entries in the spreadsheet on primary variables using the fields suggested by ETS. None of the studies used computer based testing or computer delivery of testing; thus, none were entered into the secondary variables spreadsheet. The annotated bibliography is restricted to the target articles that applied accommodations using glossaries, translation, or linguistic simplification (A1 group).

Abstracts for papers in groups A1 and A2 were carefully inspected to check whether the data reported had not been included in the Pennock-Roman and Rivera (2011) meta-analysis. For example, the 2009 paper by Abedi, "Computer testing as a form of accommodation for English language learners," (*Educational Assessment*, 14(3-4), 195- 211) presents the same tables of data as the Abedi, Courtney, & Leon (2003) CRESST technical report #586 that was included in the meta-analysis. Any papers that were "repeats," were not included in the annotated bibliography or spreadsheet for primary variables.

The search yielded five papers in group A1 and another four papers in group A2. All nine papers were included in the primary variables spreadsheet and the subset of five on glossaries, translation, and linguistic simplification were included in the annotated bibliography And highlighted in Executive Summary. None of the studies were computer-delivered and thus were not included in a spreadsheet for secondary variables.

¹ These papers were not required by the work specifications but a set of electronic files in which detailed Eric abstracts/citations are classified into folders by topic can be provided if requested.

Summary of Literature on Accommodations for ELLs

Table 1. Search Terms for Specific Accommodation Types

Accommodation Type	Search Terms Related to:	
	Names for Accommodation	Testing or English Language Learners
<i>Simplification/ Plain Language</i>	(linguistically modified or linguistic modification or simplified English or modified English or simpler English or plain English)	AND (state assessment or large scale assessment or high stakes test* or large scale test* or accommodat* or limited English or English learners or English language learners)
<i>English or bilingual glossaries</i>	(word to word dictionary or word to word glossary)	AND (state assessment or large scale assessment or high stakes test* or large scale test* or accommodat* or limited English or English learners or English language learners)
<i>Bilingual glossaries</i>	(native language glossary or native language dictionary or bilingual glossary or bilingual dictionary or Spanish English glossary or Spanish-English dictionary)	AND (state assessment or large scale assessment or high stakes test* or large scale test* or accommodat* or limited English or English learners or English language learners)
<i>English glossaries</i>	(English dictionary or customized English dictionary or English glossary)	AND (state assessment or large scale assessment or high stakes test* or large scale test* or accommodat* or limited English or English learners or English language learners)

Summary of Literature on Accommodations for ELLs

<i>Translations</i>	(translat*)	AND	(state assessment or large scale assessment or high stakes test* or large scale test* or accommodat* or limited English or English learners or English language learners)
<i>Dual language</i>	(Dual language)	AND	(state assessment or large scale assessment or high stakes test* or large scale test* or accommodat* or limited English or English learners or English language learners)

References

- Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment*, 14(3-4), 195-211.
- Abedi, J., Courtney, M., & Leon, S. (2003a). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Technical Report No. 608). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from <http://www.cse.ucla.edu/products/Reports/R608.pdf>.
- Abedi, J., Courtney, M., & Leon, S. (2003b). *Research-supported accommodation for English language learners in NAEP* (CSE Tech. Report No. 586). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from <http://www.cse.ucla.edu/products/Reports/TR586.pdf>.
- ²Abedi, 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 Technical Report No.702) Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). 92 pp. Retrieved 2/4/12 from <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED520528&site=ehost-live>.
- Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CSE Report 666). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from <http://www.cse.ucla.edu/products/reports/r666.pdf>.
- Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and tests accommodations: Interactions with student language background* (CSE Technical Report No. 536). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from <http://www.cse.ucla.edu/products/Reports/newTR536.pdf>.
- Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance*. (CSE Technical Report No. 478) Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from <http://www.cse.ucla.edu/products/Reports/TECH478.pdf>.

² All references listed here are cited in the text of this summary or in Appendices 1 and 2; only those with the footnote are new references not included in the Pennock-Roman & Rivera (2011) article

Summary of Literature on Accommodations for ELLs

- Abedi, J., Lord, C., Kim, C., & Miyoshi, J. (2001). *The effects of accommodations on the assessment of limited English proficient students in the national assessment of educational progress*. (Publication No. NCES 2001-13) Washington, DC: National Center for Education Statistics. Retrieved from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200113>.
- Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from <http://www.cse.ucla.edu/products/Reports/TECH429.pdf>.
- Aguirre-Muñoz, Z. (2000). *The impact of language proficiency on complex performance assessments: Examining linguistic accommodation strategies for English language learners* (Doctoral dissertation, University of California at Los Angeles). Retrieved from Proquest Dissertations and Theses Full Text. (Publication no. AAT 9973171).
- Albus, D., Thurlow, M., Lui, K., & Bielinski, J. (2005). Reading test performance of English-language learners using an English dictionary. *Journal of Educational Research*, 98(4), 245-254.
- Anderson, M., Liu, K., Swierzbis, B., Thurlow, M., & Bielinski, J. (2000). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2*. (Minnesota Rep. No. 31). Minneapolis, MN: National Center for Educational Outcomes. Retrieved from <http://education.umn.edu/NCEO/OnlinePubs/MnReport31.html>
- Cormier, D. C., Altman, J., Shyyan, V., & Thurlow M. L. (2010). *A Summary of the Research on the Effects of Test Accommodations: 2007-2008* (NCEO Technical Report 56). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Duncan, T. G., del Rio Parent, L., Chen, W.-H., Ferrara, S., Johnson, E., Oppler, S., & Shieh, Y.-Y. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Applied Measurement in Education*, 18(2), 129-161.
- Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English language learners. *Applied Measurement in Education*, 16(2), 159-188.
- Kiplinger, A., Haug, C. A., & Abedi, J. (2000). *A math assessment should test math not reading: One State's Approach to the Problem*. Presented at the 30th Annual National Conference on Large-Scale Assessment, Snowbird, UT.
- Kopriva, R. J., Emick, J. E., Hipolito-Delgado, 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
- ²Mann, H., Emick, J., Cho, M., & Kopriva, R. (2006). *Addressing the validity of test score inferences for English language learners with limited proficiency using language liaisons and other accommodations*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Summary of Literature on Accommodations for ELLs

- Pennock-Roman, M. and Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice* 30(3), 10-28 .
- Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment*, 9(3), 79-105.
- ²Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. *Educational Researcher*, 39, 582-590.
- ²Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C. (2010). *Accommodations for English language learner students: The effect of linguistic modification of math test item sets* (NCEE Report 2009-4079). National Center for Education Evaluation and Regional Assistance. Downloaded 7/7/2010 from <http://www.eric.ed.gov/PDFS/ED510556.pdf>.
- ²Steinberg, J., Cline, F., Ling, G., Cook, L. Tognatta, N. (2009). Examining the validity and fairness of a state standards-based assessment of English-language arts for deaf or hard of hearing students. *Journal of Applied Testing Technology*, 10(2). (no page numbers--online--33 pp.).
- Thompson, S., Blount, A., Thurlow, M. (2002). *A Summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- ²Wolf, M. K., Kim, J., Kao, J. C., R., & N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a math assessment* (CRESST Report 766). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). 69 pp. Retrieved 2/4/12 from <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED507754&site=ehost-live>.
- ²Wolf, M. K. & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, 14(3-4), 139-159
- ²Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment*, 13(2-3), 170-192.
- ²Young, J. W., Holtzman, S., Steinberg, J. (2011). *Score comparability for language minority students on the content assessments used by two states* (Research Report. ETS RR-11-27). Princeton, NJ: Educational Testing Service.



Summary of Literature on Accommodations for ELLs

Appendix A

Empirical Studies of the Validity and Effectiveness of Test Accommodations:

Glossaries or Dictionaries, Linguistic Modification or Translation³

Annotated Bibliography 2005-2011

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 Technical Report No.702) Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). 92 pp. Retrieved 2/4/12 from <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED520528&site=ehost-live>.

In this experimental study, the authors evaluated the dual-language (English-Spanish) accommodation and a linguistically simplified English language accommodation for a Grade 8 paper-and-pencil algebra test taking into account opportunity to learn. The unaccommodated test booklet was included among the three test forms randomly distributed to students in a class. All students within a class responded to test items under the same limited time constraints regardless of test form.

The main focus of the analysis using hierarchical linear modeling was on the degree to which test performance was affected by opportunity to learn and test accommodation type after controlling for prior math ability of the students. Opportunity to learn was represented by three class-level variables (class average of student perceptions of content coverage, teacher's knowledge, and class average of students' prior math performance). All three opportunity-to-learn variables were significantly related to math performance, after controlling for prior math ability at the individual student level. Results also indicated that the two language accommodations did not impact students' math scores after controlling for the other variables. In general, ELL students reported less content coverage than their non-ELL peers, and they were in classes of overall lower math ability than their non-ELL peers. However, there was no relationship between observed class interactions between students and teachers according to ELL status or students' English reading level proficiency.

There are several factors here that may have lowered test performance on the dual-language accommodation type owing to the limitations of carrying out an experimental study with the constraints of a real school setting. First, the dual language test form contained approximately twice the number of pages as compared to the other test forms; however, it was necessary to test all members of the class within particular class periods. Consequently, when the allowed time limits were the same for all forms, the effectiveness of this accommodation may have been reduced – see Pennock-Roman and Rivera (2011). Furthermore, the random assignment of the form to ELLs in general did not guarantee that the students receiving it had Spanish as their

³ Studies here involve quantitative analyses of data designed to explore validity and effectiveness of the accommodation types listed above. It is an update to the meta-analytic review by Pennock-Roman and Rivera (2011). Articles published during 2005-2011 with data already analyzed in technical reports included in aforementioned meta-analyses were excluded from the annotated bibliography. For example, the 2009 paper by Abedi, Computer testing as a form of accommodation for English language learners, (Educational Assessment, **14**(3-4),195-211) presents the same tables of data as the Abedi, Courtney, & Leon (2003) CRESST technical report #586.

home language or that they were literate in Spanish, or that they had received prior recent instruction in mathematics in Spanish. The majority of 8th grade students most likely had been receiving instruction in English for 5-8 school years. If there were students who had recently migrated to the U.S. after receiving math instruction in Spanish for most previous grades this group may have been too small in relation to other ELLs to make the accommodation effect detectable.

Implications for Policy: This is the first accommodation study to show unequivocally how much opportunity to learn impacts students' performance regardless of the efforts to improve test validity with test accommodations. Although test accommodations may reduce construct irrelevant variance in the measurement of achievement, they cannot be considered the panacea for closing the achievement gap between native speakers of English and language minority students in the schools. There is unequal opportunity to learn for ELLs that can account for much of the achievement gap. Closing the gap will require improving the quality of the schooling that ELLs receive. As demonstrated in one example using data from *High School and Beyond* in the textbook by Bryk and Raudenbush's (1992, pp. 103-113), there is much variation among schools. The authors demonstrated that a subset of effective schools was able to raise achievement for at risk students and reduce the impact of students' socioeconomic status and prior student achievement on subsequent student test scores.

Implications for Implementation: In order for the dual language accommodation test form to be effective, it has to be assigned to a particular group of ELLs—specifically students with literacy skills and instructional experience in the native language of the test. Owing to its extra length, it must be tested with sufficient time (preferably no time restriction). To evaluate the effectiveness of a Spanish-English dual language test, the accommodated and original versions need to be randomly assigned within this particular group of ELLs literate in Spanish and have both forms administered with very generous time limits.

Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. *Educational Researcher*, 39, 582-590.

This study is highly unusual in its focus on the kindergarten and first grade student performance using individually administered oral assessments of math knowledge (in Spanish or English) and oral English proficiency (EP) to students with Spanish as a home language. Assignment to the Spanish or English version varied according to whether students reached a cut-off score value (37) on the test of EP. Although quasi-experimental in design, they applied a very elegant and sophisticated discontinuous regression approach to evaluate the effectiveness of the native language version the test as compared with the English language version for students at low or intermediate levels of English proficiency. Math test scores were predicted by EP scores for students taking the Spanish version (EP scores below 37) and for students taking the English version (EP scores of 37 or higher). Specifically, the procedure examined the degree of discontinuity in the two regression lines at the values of 36 and 37 on the EP score horizontal axis. One of the predictor variables entered into the regression models (in addition to EP) was the match between the language of classroom instruction to the language of the test. Testing and analyses were done at three points in time: fall of the kindergarten year, spring of the kindergarten year, and spring of the first grade. The dependent, outcome variable was total math

scores in some analyses; in other analyses, the dependent variable was one of several math subscores derived from item clusters according to level of math.

The regressions were found to be essentially the same for Spanish and English versions in the prediction of the math subscore based on items involving counting. These items had low language processing demands. For more advanced math concepts having more linguistically taxing items (e.g., those asking for relative size, sequence, addition/subtraction, and multiplication/division), the regressions for test versions differed. The Spanish version was superior to the English version in revealing students' knowledge for the more linguistically demanding items.. The regressions for English and Spanish versions were nearly identical in the fall semester of kindergarten where most of the items students could answer were in the simpler counting section; the patterns of the regressions diverged more in the spring semester with the greatest difference at the end of first grade. In the later two time periods, students' performance in Spanish was substantially better and less related to EP as compared with the English version. The later divergence in regressions reflected the increasing language demands of the items in going from early kindergarten to the spring semester of first grade. Although ratings by teachers on mathematical and other academic skills for students at the threshold (36-37) in KB and 1B were nearly equal, those who took the English version scored lower than those who took the Spanish version. The author stated that "assessing their mathematical skills in English prevented those ELLs from demonstrating their skills." (p. 586)

These results were clear evidence of greater validity of the Spanish versions owing to their lesser relationship to the oral proficiency in English. The match between the *language of classroom instruction* to the language of the test was found to have *no significant* relationship to the test scores after controlling for EP. For kindergarten and first grade students, matching the *home* language to the language of the test appears to be the more valid approach.

Implications for Policy: The study strongly supports the use of native language assessment for students with low fluency in English, especially in the early grades where most of the students have received instruction at home in their native language. In first grade and kindergarten, matching the language of the test to the home language appears to be much more important than matching it to the language of instruction for students at lower EP levels. However, the language of instruction can be expected to be much more salient in later school years.

Implications for Implementation: The procedures for test translation and standardization of oral administrations used here were exemplary and should serve as a model for future studies.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment*, 13(2-3), 170-192.

The authors of this correlational study used data from a state's operational testing for NCLB accountability to evaluate the psychometric properties of accommodated versions of tests as compared with the original versions. The analyses included exploratory factor analysis (EFA) multi-group confirmatory factor analyses (CFA), comparisons of internal consistency reliabilities, and group contrasts in differential item functioning (DIF). The key accommodations studied were

(1) orally translated test directions and (2) bilingual glossaries of paper-and-pencil tests of math, algebra, and science in Grades 5 and 8.

The authors concluded that:

There was little evidence of differential test validity in terms of internal test structure or item functioning for these examinee groups, when the performance of non-ELLs and ELLs were examined and compared... The item-level and item parcel EFA and CFA results showed that the tests are essentially unidimensional for ELLs, with or without accommodations, and [for] non-ELLs, which is reassuring for the purposes of ascertaining construct validity. However, for ELLs, there appears to be more construct irrelevant noise possibly affecting the magnitude of the first eigenvalue, as this is generally smaller for them when compared to non-ELLs. Further, the factor analysis results indicate that the use of one of the ELL testing accommodations, access to translation glossaries/word lists, was effective for supporting the unidimensionality of some of these assessments. In addition, the use of translation glossaries/word lists appears to have a more beneficial effect for ELL examinees on the eighth-grade assessments than on the fifth-grade assessments, by making a stronger case for unidimensionality....The DIF analyses... showed that group differences in performance on the test items, after matching on total test score, are small. This indicates that almost all of the items functioned appropriately, in that significant DIF was rarely observed for ELLs, with or without accommodations (pp.189- 190).

Implications for Policy: Scores from accommodations involving orally translated directions or bilingual glossaries for ELLs have some of the same psychometric properties as scores from the original versions of these tests. However, without an examination of the concurrent or predictive validity of accommodations in relation to external criteria these analyses are not sufficient to establish the score comparability of accommodated and original versions of tests.

Implications for Implementation: None.

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C. (2010). *Accommodations for English language learner students: The effect of linguistic modification of math test item sets* (NCEE Report 2009-4079). National Center for Education Evaluation and Regional Assistance. Downloaded 7/7/2010 from <http://www.eric.ed.gov/PDFS/ED510556.pdf>.

This experimental study constitutes the largest, most comprehensive evaluation to date of the validity and effectiveness of linguistic modification of test items as an accommodation for ELLs. The items that were taken from a pool of 256 NAEP items in math appropriate for 7th and 8th grade. They were specifically selected only if they contained "sufficient language to linguistically modify (number sense/operations and measurement content strands)." These items presumably had unfamiliar words, or complicated sentence structure, or complex verb tenses, or could benefit from added graphics or tables to increase clarity. The final set of items after review by experts and pre-testing included 25 matched pairs, one modified, the other original (see p.23). This selection process enhanced the contrast between the original and modified versions of the items. Compared to previous studies of linguistically modified items, the assessment here included a larger number of items for which modification could make a big difference. The development and

refinement of the modified items included reviews by experts, examinee think aloud methods, and other good psychometric practices.

Sets of modified and original items were distributed randomly within a class. Results were analyzed in three subgroups: (1) ELLs; (2) non-ELLs less proficient in English language Arts (ELA); and (3) non-ELLs more proficient in ELA. There were 606 ELLs in the original items group and 608 ELLs in the linguistically modified group. Sample sizes for each of the non-ELL groups were even larger. The authors also compared the two versions of the items with respect to reliabilities, correlations with another math test (concurrent validity), factor structure, and DIF. The DIF analyses contrasted each measure across groups 1 and 2 and groups 2 and 3.

The authors found that both versions were essentially unidimensional in their factor structure and their concurrent validity correlations with the state math assessment score were comparable. The authors concluded that "As implemented in the current study, linguistic modification did not alter the targeted math construct assessed" (p. 2).

The magnitude of the difference in mean scores found between the original item set and the linguistically modified item set for EL students was 0.16 standard deviation units (raw score metric), a larger effect than averages found in previous meta-analyses of this technique (Kieffer, et al., 2009; Pennock-Roman & Rivera, 2011). However, the authors did not examine whether the effect was equally large for groups ELLs of at low vs. intermediate EP as suggested in the Pennock-Roman and Rivera paper where an individual effect size of 0.57 was found for a group having high intermediate EP. Given the differences found between groups of non-ELLs varying in ELA scores, one would expect the effect to be smallest for beginning ELL students with such low EP that even a simplified English version would be inaccessible. Hence, the effect size with these sets of items could be potential larger if the mean had been based only on ELLs with at least an intermediate level of EP.

Implications for Policy: The results of this study suggest that linguistic modification of test items does have potential as a viable accommodation, contrary to the conclusion of Kieffer et al. (2009). The improvement in the reduction of construct irrelevant variance owing to linguistic simplification can be substantial if there are many language-intensive items in the original test.

Implications for Implementation: The process of development and refinement of the linguistically modified items here was exemplary and should provide a model for future studies and test development. Owing to the very small number of studies that examined the effects of linguistic modification for ELLs of intermediate EP, the hypothesized interaction between EP and the effects of linguistic modification needs to be confirmed in future research. The trend in past data suggests that effects for linguistic modification will be larger if (1) ELLs have an intermediate level of EP and (2) the original version of the test has a large proportion of linguistically demanding items (as did the original test here).

Wolf, M. K., Kim, J., Kao, J. C., & Rivera, N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a math assessment* (CRESST Report 766). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). 69 pp. Retrieved 2/4/12 from <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED507754&site=ehost-live>.

In this experimental study, ELL and non-ELL participants were assigned randomly to three testing conditions (English glossary, read-aloud, and no accommodations) in two states. Whereas the read-aloud condition was separately tested and allowed more time, the English glossary condition had the same time restrictions as the original test version because they were administered simultaneously within the same class. A selected group of participants was analyzed in a qualitative study involving verbal protocols and think aloud techniques. The authors state that:

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.....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 (p. 47).

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...[the results suggested] 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 (p. 48).

In both states' samples, no significant interaction effect was found between the given accommodation and students' ELP 49 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 (pp. 48-49).

The authors did not provide effect sizes such as Glass's index for mean effects; given the low statistical power, some non-trivially large effect sizes for the accommodations at the higher levels of English proficiency may not have reached statistical significance.

Implications for Policy: The observed interaction effect between students' prior knowledge and accommodation support the validity of this approach. It suggests that the effectiveness of the accommodation cannot be judged in circumstances where students have not had the opportunity to learn the test material. No matter how much construct irrelevance variance may have been reduced in the accommodation, if the students have not learned the material, one cannot expect a noticeable improvement in test performance.

Implications for Implementation: It is clear that students will profit more from a glossary accommodation if they have sufficient training and experience with the method.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Abedi et al. (2006, Report #702) Study
A	Study Citation	Abedi, J, Courtney, M., Leon, S., Kao, J., Azzam, T. (2006). <i>English language learners and math achievement: A study of opportunity to learn and language accommodation</i> . (CSE Technical Report No.702) Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). 92 pp. Retrieved 2/4/12 from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED520528&site=ehost-live .
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	There were two main research goals: (1) To evaluate the role of classroom content coverage, test accommodation type, class participation on ELL and non-ELL student performance in algebra after controlling for previous student knowledge in math; (2) To examine whether ELL students or students less proficient in reading were exposed to less complete content coverage in the classroom and had less active interactions with teachers.
B	Description of accommodation(s)	A) dual-language (English and Spanish) test version accommodation with side by side presentation of items: B) linguistically modified test version accommodation (English language).
C	Computer delivery of accommodation (yes/no)	No
D	Test Content	algebra
E	Age or grade	8
F	Disability	It is unclear whether students with disabilities were excluded or not in the selection of students for the study (pp.21-23) .There is no mention of the percentage of students with disabilities by school in the descriptive section (pp. 27-
G	CBT? (yes/no)	No

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Abedi et al. (2006, Report #702) Study
H	Research design	For goal (1) design was a correlation/regression approach; the random assignment to one of three test forms containing one of three accommodation conditions was done by spiraling booklets. The authors state: "The test versions had similar appearances, were pre-collated for almost equal distribution in each of the classes, and were distributed randomly by the test administrators who often were assisted by teacher and/or student volunteers" (p.39). For goal (2): the design involved contrasting means for groups differing in ELL status or reading level proficiency according. The means were derived from types of variables. One group of variables were classroom-level characteristics related to opportunity to learn. The second group were variables on teacher-initiated and student initiated interactions derived from classroom observation.
I	Data analysis technique	hierarchical linear modeling; ANOVA.
K	Sample size category	Large (N= 2,321) for hierarchical linear models.
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	No
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations	Yes
NONE	How restricted were time limits for the original test?	Not a power test. The test was constructed specifically for the study and was not stated to be a power test. Administrators allowed 40 minutes for 32 items in an algebra test, but there was no indication of how frequently students in each language proficiency group were able to complete the original test.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Abedi et al. (2006, Report #702) Study
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	Yes, same time limits for all versions. All three versions were included within booklets distributed to a class and administered during the same time period, so that the design did not permit separate time limits for different versions. There was no mention of any variation in allowed time limits among those receiving the three accommodation type conditions. Hence, students receiving the dual-language booklet had to page a much larger test within the same constraints as the other two conditions.
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	Yes, with student reports about content covered in the classroom and measures of teachers' content knowledge related to the test.
NONE	Was English language proficiency measured?	Strictly speaking, no, but they included two measures of reading proficiency in English in the analyses which they use as a proxy for "language proficiency." Also, they collected data on student English language development information, and home language from school records. In addition, they collected data on student language background characteristics from students: "An 8-item questionnaire was used to collect data pertaining to students' language background, such as country of origin, length of time in the U.S., and language other than English spoken in the home." (p.36)
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	Yes. Reading level was used to subdivide students into groups and these groups were used to address the following research question (#4): "Do the dual-language test version and linguistic modification accommodations differentially impact the math performance of students with varying language proficiency?" (p.41) . The authors state that: All main effects and two-way interactions between class-level OTL, type of test accommodation, and student language proficiency were included in each of the three [HM] models' (p. 46).
NONE	Did the language of the accommodation (English or native) match the language of instruction for participants?	One can infer yes, although this issue was not specifically addressed in the report. Instruction can be assumed to be in English for students in California owing to state policies discouraging use of the native language for instruction. Both accommodations included items in English (although the dual language version also had the same items in Spanish).

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Abedi et al. (2006, Report #702) Study
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' home language?	For the majority, yes, but not for all. There is no mention of leaving out ELLs with a non-Spanish background from the analyses involving the dual-language test version in the method or results. Hence, it is possible that the analyses of the Spanish-English dual language test version included at least a few students who were not Spanish speakers. . Across the 21 schools, 62.5% of the students (ELL and non-ELL) chose Spanish as one of their home languages. The percentage of students of Hispanic origin (ELLs and non-ELLs) ranged from 34.4% to 97.1 across the 21 schools (see pp. 21-23).
NONE	Was native language proficiency measured?	It was not specifically mentioned in the description of student reports: "An 8-item questionnaire was used to collect data pertaining to students' language background, such as country of origin, length of time in the U.S., and language other than English spoken in the home." (p.36).
NONE	Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	No. This issue was not specifically mentioned in research questions 1-9 (pp. 41-42), nor in the results on HLM models (pp. 46-58). For the ELL Hispanic students who could not read Spanish or for ELLs with a non-Spanish background, the effect was to take an English-language test with double as many pages as compared with the unaccommodated test version.
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	"9 of the 32 math items were noticeably modified in the linguistically modified test version" (p.46). That is, only 28% of the items were linguistically simplified in a noticeable way.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Abedi et al. (2006, Report #702) Study
L	Findings	<p>Results indicated that all three class-level components (student perceptions of content coverage, teacher's knowledge, and class prior math performance) were significantly related to math performance, after controlling for prior math ability at the individual student level.</p> <p>Class prior math ability had the strongest effect on math performance. Results also indicated that teacher content knowledge had a significant differential effect on the math performance of students grouped by a quick reading proficiency measure, but not by students' ELL status or by their reading achievement test percentile ranking. Results also indicated that the two language accommodations did not impact students' math. Additionally, results suggested that, in general, ELL students reported less content coverage than their non-ELL peers, and they were in classes of overall lower math ability than their non-ELL peers. However, there was no relationship between observed class interactions between students and teachers according to ELL status or English reading level proficiency.</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Mann et al. 2006 Study
A	Study Citation	Mann, H., Emick, J., Cho, M., & Kopriva, R. (2006). <i>Addressing the validity of test score inferences for English language learners with limited proficiency using language liaisons and other accommodations</i> . Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	1) To compare the effectiveness of language liaisons vs. read aloud accommodations for ELL students and 2) to examine the concurrent validity of accommodated test scores for predicting teacher ratings of student knowledge in mathematics
B	Description of accommodation(s)	<p>Read aloud (oral administration of the test in English), a word-picture list in English, a Spanish-English glossary, use of manipulatives, oral administration in English, small-group administration, and language liaison. A language liaison is an adult proficient in Spanish who reads test instructions in Spanish and is available to answer certain defined questions in Spanish.</p> <p>The language liaison was trained to address certain language tasks and was instructed not to provide an on-the-fly translation. They answered questions about phrases, concepts, and other item tasks that are difficult to define in a glossary and not connected to the targeted constructs being measured on the tests. Also, some accommodations were bundled in an individualized way for students according to the challenges they faced. All test administrations involving language liaisons were tape-recorded, as were a sample of the oral administrations. All participants received linguistically simplified items.</p>
C	Computer delivery of accommodation (yes/no)	No
D	Test Content	Math (included 11 multiple choice items and 8 constructed-response items). These items were rewritten versions of state released mathematics items, designed to measure the same mathematics constructs but provide more access for students with less proficiency (e.g., shorter sentences, modified vocabulary, more accessible problem contexts, clearer formatting, use of pictures/graphic organizers, etc.).

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Mann et al. 2006 Study
E	Age or grade	Grades 3 and 5.
F	Disability	Unknown if students with disabilities were included.
G	CBT? (yes/no)	No
H	Research design	Analyses A. <i>Not an experimental study.</i> Accommodations were assigned to students on the basis of school records data and teacher reports of student challenges. Language liaisons were provided for identified Spanish speaking students only. For purposes of analyses, five groups were identified: three groups of ELL students (beginning, intermediate, and advanced) and two additional groups (exited ELL and non-ELL). Analyses B. <i>Experimental.</i> Students in the intermediate ELL group with low reading proficiency in English were randomly assigned to the oral administration in English vs. the language liaison condition.
I	Data analysis technique	Multiple regressions were used to investigate the relationship between test scores and teacher ratings of student knowledge. Also, means of the experimental groups were compared with <i>t</i> -tests..
K	Sample size category	ELL groups at various levels of proficiency within each grade had sample sizes in the small to medium range (37 to 197); former ELL groups had medium sample sizes (245-255); non-ELL groups had large sample sizes (711-719).
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	Yes.
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations?	In the experimental group analyses, yes. In the regression analyses, no. In the latter, accommodations were administered in individualized packets for some students and could be thought of being the combination most suited to a particular student among available choices.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Mann et al. 2006 Study
NONE	How restricted were time limits for the original test?	unknown
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	unknown
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	Test items met state standards but there was no attempt to quantify how well test content was covered in ELL classrooms.
NONE	Was English language proficiency measured?	Yes, to define ELL groups and categorize them into beginning, intermediate, and advanced groups.
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	No
NONE	Did the language of the accommodation (English or native) match the language of <i>instruction</i> for participants?	unknown

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Mann et al. 2006 Study
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' <i>home</i> language?	Yes for language liaison; unknown for bilingual glossary but the latter was not analyzed separately.
NONE	Was native language proficiency measured?	No
NONE	Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	No
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	N/A

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Mann et al. 2006 Study
L	Findings	<p>During the process of training the language liaison personnel, it was found that more than one training session was needed to prepare language liaisons. Although the language liaisons were screened for oral proficiency in Spanish, our results indicated that it was also necessary to screen for their literacy in English. Finally, it was also clear that this accommodation should be practiced in the classroom prior to the assessment. This provides the student with the opportunity to understand the role of the language liaison and fully utilize their presence on the day of the test.</p> <p>In the experimental portion of the study, which was restricted to poor readers, the group receiving language liaisons had lower means than the group receiving the read aloud accommodation, but the difference was statistically significant only for the Grade 3 multiple choice portion of the test..</p> <p>Regression results varied by grade level, multiple-choice vs. constructed response, data from analyses A vs. B, ELL status, and type of accommodation (language liaison vs. read aloud).</p> <p>"On both the multiple choice and constructed response tests, validity is lower for all three ELL groups relative to exited and non-ELLs in Grade 3. In Grade 5, validity is the same for advanced ELLs and exited and non-ELLs on both multiple choice and constructed response. Surprisingly, validity is not different for beginning ELLs compared to exited and non-ELLs on the constructed response test in Grade 5 " (p. 26)</p> <p>"The analyses in this study suggest low validity for early ELLs, even when proper accommodations are given." (p. 25)</p> <p>"It is possible that early ELLs do not get as full a curriculum as non-ELLs and have less of an opportunity to learn. Also, because of their language problems, ELL teachers often will teach math concepts using just algorithms rather than using word problems—of which the entire test is comprised." (p. 26).</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Robinson2010:
A	Study Citation	Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. <i>Educational Researcher</i> , 39, 582-590.
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	To evaluate the degree of relationship between English proficiency (EP) and math content scores for English and Spanish versions of the test, taking into account regression intercept differences (which reflect mean differences between the language versions).
B	Description of accommodation(s)	Administration of either an English version of the test or a Spanish version of the test, both versions in a read aloud format. Items were read according to a standardized script by the test administrator. Assignment to Spanish or English version varied according to whether students reached a cut-off score value on the test of EP.
C	Computer delivery of accommodation (yes/no)	No
D	Test Content	Math
E	Age or grade	K and 1st grade
F	Disability	not mentioned
G	CBT? (yes/no)	Partially yes. Scoring recorded and calculated by computer although items individually administered orally by a test administrator..
H	Research design	Rigorous, well designed quasi experimental study using a sophisticated regression approach. Students with Spanish as a home language were not categorized into groups by ELL status in the schools per se, but rather into Fluent or Not Fluent (F or NF) groups according to a measure of oral English proficiency (EP) at three points in time. The three points in time were fall semester and spring semester kindergarten--KA and KB-- and spring semester of first grade--1B. The groups in the analyses were: (1) students who were sufficiently fluent in English during the first semester of Kindergarten (KA) to be administered the English version of the math test in KA; (2) students who were NF in KA and who were administered the Spanish version of the math test in KA; (3) students who were NF in KA and NF in KB who received the Spanish version in KB; (4) students who

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Robinson2010:
		were NF in KA and F in KB who received the English version in KB; (5) students who were NF in KB and NF in 1B who were given the Spanish version in 1B; (6) students who were NF in KB but F in 1B and given the English version in 1B.
I	Data analysis technique	Discontinuous regression analyses; analyses were done for total math scores and also by subscores derived from item clusters according to level of math.
K	Sample size category	Large in all analyses ($N \geq 576$)
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	In a sense, yes. The groups receiving the English version of the test in the KB and 1B semesters could be considered analogous to former ELLs because, at the previous testing, they scored below threshold (classified not fluent) on the EP test.
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations	Yes
NONE	How restricted were time limits for the original test?	No time limits; individually administered instrument.
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	No time limits; individually administered instrument.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Robinson2010:
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	No. Items were derived from comparable published measures for early school years.
NONE	Was English language proficiency measured?	Yes. EP was measured continuously and analyzed as a predictor variable in the regressions.
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	Yes, in a sense; the regression discontinuity for students near the cut-off of scores for the F/NF dichotomy could be considered a type of interaction.
NONE	Did the language of the accommodation (English or native) match the language of <i>instruction</i> for participants?	In some cases yes, but not systematically. However, for these very early grades, the language of instruction in the classroom was less relevant to the proper choice of language in the assessment as shown by the results. Language of classroom instruction (whether English or Spanish) was included in the analyses but had no effect on outcomes in these early grades. All participants had Spanish as the home language and for kindergarteners most of the "prior instruction" would have been in the home, presumably in Spanish.
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' <i>home</i> language?	Yes. Those who received the Spanish version had Spanish as a home language.
NONE	Was native language proficiency measured?	Yes
NONE	Was level of literacy/proficiency in the native language among ELLs	Not mentioned in results or method.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Robinson2010:
	included as a variable that could interact with the accommodation?	
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	N/A

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Robinson2010:
L	Findings	<p>Students with scores of 36 and 37 on the EP test had essentially equal math skills, literacy, or general knowledge as rated by teachers who were unaware of students' EP scores.</p> <p>The relationship between EP and math score on the English and Spanish versions varied by math content and year of testing.</p> <p>The regressions were essentially the same for Spanish and English versions for the subscore based on items involving counting. These items had low language processing demands. For more advanced test items having more linguistically taxing items, the regressions for test versions differed. The Spanish version was superior to revealing students' knowledge for the more linguistically demanding items (e.g., those asking for relative size, sequence, addition/subtraction, and multiplication/division).</p> <p>The regressions for English and Spanish versions were nearly identical in KA where most of the items students could answer were in the simpler counting section; the patterns of the regressions diverged more in KB with the greatest difference in 1B. In the later 2 semesters, students' performance in Spanish was substantially better and less related to EP as compared with the English version. The later divergence in regressions reflected the increasing language demands of the items in going from KA to 1B.</p> <p>Although ratings by teachers on mathematical and other academic skills for students at the threshold (36-37) in KB and 1B were nearly equal, those who took the English version scored lower than those who took the Spanish version. The author stated that "assessing their mathematical skills in English prevented those ELLs from demonstrating their skills." (p. 586)</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Sato et al. (2010)
A	Study Citation	Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C.(2010). <i>Accommodations for English language learner students: The effect of linguistic modification of math test item sets (NCEE Report 2009-4079)</i> .National Center for Education Evaluation and Regional Assistance. Downloaded 7/7/2010 from http://www.eric.ed.gov/PDFS/ED510556.pdf .
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	To evaluate the efficacy of linguistic modification of math test items for ELLs and non-ELLs who were less proficient in English language arts.
B	Description of accommodation(s)	Linguistically modified English language items
C	Computer delivery of accommodation (yes/no)	No
D	Test Content	math
E	Age or grade	7 and 8
F	Disability	not mentioned in main body of results (check appendix)
G	CBT? (yes/no)	No
H	Research design	Experimental study. Sets of modified and original items were distributed randomly within a class. Results were analyzed in three subgroups: (1) ELLs; (2) non-ELLs less proficient in English language Arts (ELA); and (3) non-ELLs more proficient in ELA.
I	Data analysis technique	Mean differences between groups; DIF; factor analyses; reliabilities; correlations with another math test (concurrent validity). The DIF analyses contrasted each measure across groups 1 and 2, and groups 2 and 3. The development and refinement of the items included expert review, think aloud methods, and other good psychometric practices.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Sato et al. (2010)
K	Sample size category	Large (sample sizes per group > 600). See Table A1 p. 67
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	No. The non-ELL group was subdivided into Groups 2 and 3 by their performance on the state assessment in English language arts (ELA), not former ELL status.
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations	Yes
NONE	How restricted were time limits for the original test?	50 min. to answer 25 math items and a survey. Initial pilot testing reduced the number of items from 30 to 25: "Team members concurred that five items should be removed to ensure that students had adequate time to answer both the math questions and the Student Language Background Survey questions." (p. 29)
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	Same as above.
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	No quantitative assessment of classroom curricular representation. However, items were selected from the pool of released NAEP items after expert review in which curricula and state standards were part of the review process.
NONE	Was English language proficiency measured?	Yes, using the California English Language Development Test (CELDT). The target ELL sample was students whose first language was Spanish and who demonstrated early intermediate to advanced levels of ELP on this test.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Sato et al. (2010)
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	Not EP per se but ELA proficiency status can be considered a proxy for EP and it was included in the analyses as a variable that could interact.
NONE	Did the language of the accommodation (English or native) match the language of <i>instruction</i> for participants?	Language of instruction not specifically mentioned or analyzed.
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' home language?	N/A
NONE	Was native language proficiency measured?	No
NONE	Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	No
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	100%. That is, 81 of the items selected for study out of a total of 256 items were those that contained "sufficient language to linguistically modify (number sense/operations and measurement content strands)." Of these, 51 pairs of matched items were derived, one modified, the other original. These 51 items presumably had unfamiliar word or complicated sentence structure or complex verb tenses, or that could benefit from added graphics or tables to increase clarity. The final set of items after review by experts and pre-testing included 25 matched pairs (see p.23) .

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Sato et al. (2010)
L	Findings	"The effect size, or magnitude of the difference in mean scores between the original item set and the linguistically modified item set for EL students, was 0.16 standard deviation units using a raw score metric and 0.17, 0.12, and 0.09 standard deviation units when the scores were derived using the 1-, 2-, and 3-[parameter logistic] models, respectively. (p.2)
		"For each student subgroup, the mean difference in performance on the two item sets was greatest for EL students [Group 1], followed by...[Group 2, non-ELLs with low ELA scores]" (p. 2)
		The DIF results showed few items with DIF. The original items had one item with DIF in the Group 1 vs. Group 3 contrast and no DIF items in the Group 2 vs. Group 3 contrast. The linguistically modified items showed two items with DIF in the Group 1 vs. 3 contrast and no DIF items in the Group 2 vs. Group 3 contrast. "Subsequent review of these items by content, population, and assessment experts did not find evidence of bias in either item set." (p. 2)
		"As implemented in the current study, linguistic modification did not alter the targeted math construct assessed [as measured by concurrent validity correlations with the state math assessment scores] " (p. 2) For all three student subgroups, one dominant factor (math understanding) was found to underlie both item sets; however, the measurement structure between the underlying factor and the items differed across student subgroups." (p. 2)

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study by Steinberg et al., 2009
A	Study Citation	Steinberg, J., Cline, F., Ling, G., Cook, L. Tognatta, N. (2009). Examining the validity and fairness of a state standards-based assessment of English-language arts for deaf or hard of hearing students. <i>Journal of Applied Testing Technology</i> , 10(2). (no page numbers--online--33 pp.)
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	(1) Evaluate test structure and item comparability in unaccommodated tests across groups classified by hearing status and ELL status; (2) Evaluate item comparability across accommodated vs. unaccommodated versions among groups of Deaf and/or Hard of Hearing Students (DAOHH); (3) Evaluate item comparability for DAOHH students receiving accommodations vs. non-disabled students receiving no accommodations.
B	Description of accommodation(s)	Not specified. Only deaf and hard of hearing students received accommodations; non-disabled ELLs received no accommodations. There was no mention of direct or indirect linguistic accommodations.
C	Computer delivery of accommodation (yes/no)	Not specified, but we can presume that computer delivery was not involved, given that the prevalent accommodations for deaf and hard of hearing students involve live interpreters.
D	Test Content	English language arts
E	Age or grade	Grades 4 and 8
F	Disability	Deaf or hard of hearing
G	CBT? (yes/no)	No
H	Research design	No mention of how accommodations were assigned to deaf and hard of hearing participants. Presumably, there was no random assignment.
I	Data analysis technique	parcel-level exploratory and confirmatory factor analyses and differential item functioning (DIF).
K	Sample size category	For research purpose (1) above, sample sizes were large (500). For research purposes (2 and 3) Sample sizes were medium (above 50) to large. Overall, sample sizes ranged from 104 to 30,225 among 11 groups studied.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study by Steinberg et al., 2009
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	No
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations	Not specified in article
NONE	How restricted were time limits for the original test?	Not specified in article
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	Not specified in article
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	No. The assessment was based on state standards but the correspondence between classroom instruction and the standards was not a focus of the study.
NONE	Was English language proficiency measured?	No

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study by Steinberg et al., 2009
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	No
NONE	Did the language of the accommodation (English or native) match the language of instruction for participants?	Language of instruction for ELLs not specified. ELLs that were analyzed received only the original test under standard conditions.
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' <i>home</i> language?	No native language accommodations were specified—accommodations received by participants were not described.
NONE	Was native language proficiency measured?	No
NONE	Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	No

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study by Steinberg et al., 2009
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	N/A
L	Findings	<p>Factor Analyses, Test Under Standard Conditions. The unaccommodated test had similar, but not identical, factor structure across four groups classified by hearing and ELL status in each grade. All were consistent with a one-factor solution, but loadings, error variances, and/or residuals were sometimes unequal across groups defined by hearing status. Also, ELLs and disabled students differed substantially from non-disabled non-ELLs in means.</p> <p>DIF Analyses. Among Grade 4 ELLs who took the test without accommodations, items were comparable for Deaf or Hard of Hearing Students as compared with non-disabled students (there were no items showing large DIF). An analogous result was found for the same contrast among Grade 8 ELLs. Among Grade 4 Deaf or Hard of Hearing students who took the test without accommodations, items were comparable for ELLs and non-ELLs. An analogous result was found for the same contrast for Grade 8 students.</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Wolf & Leon (2009) Study
A	Study Citation	Wolf, M. K. & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. <i>Educational Assessment</i> , 14(3-4), 139-159
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	To examine differential item functioning in 11 assessments from three states and investigate whether the language demands of the items are associated with the degree of DIF for English language learner (ELL) students.
B	Description of accommodation(s)	None specified. Not clear if any ELLs were measured with accommodations.
C	Computer delivery of accommodation (yes/no)	No.
D	Test Content	math and science
E	Age or grade	Grades 4, 5, 7, and 8
F	Disability	Specifically excluded students with disabilities.
G	CBT? (yes/no)	No.
H	Research design	Correlational study, not experimental. As stated by the authors: "The focal group of interest included all ELLs, low English proficient ELLs (Low ELLs), and high English proficient ELLs (High ELLs). We categorized high and low English proficient ELLs based on the English language proficiency levels available in the data for each state. We categorized ELLs 'High' when the state label was either 'advanced,' 'proficient,' or 'superior.' We categorized ELLs as 'Low' for this study when the state label was 'intermediate' or below." (p.142) The reference group for each analysis was non-ELLs. To examine the language characteristics of test items, a linguistic analysis protocol and rater guidelines were developed.
I	Data analysis technique	Differential item functioning (DIF);

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Wolf & Leon (2009) Study
K	Sample size category	High ELL groups were medium to large (N =192- 2667); Low ELL groups were medium to large (N= 291-2871); total ELL groups and non-ELL groups all large.
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	No.
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations?	N/A
NONE	How restricted were time limits for the original test?	Unknown
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	N/A
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	No
NONE	Was English language proficiency measured?	Yes, and used to define groups.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Wolf & Leon (2009) Study
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	As a categorical variable. ELL group subdivided into high and low EP.
NONE	Did the language of the accommodation (English or native) match the language of <i>instruction</i> for participants?	N/A
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' <i>home</i> language?	N/A
NONE	Was native language proficiency measured?	No
NONE	Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	N/A

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Wolf & Leon (2009) Study
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	N/A
L	Findings	<p>"One notable finding was that the math tests, which are typically assumed to possess lower language demands than science, contained a wide variety of general academic vocabulary [general vocabulary terms are not technical terms specific to the context but words that are infrequent outside a formal academic setting]...In general, the science tests were more linguistically demanding than the math tests and .. higher grade tests were more linguistically demanding than lower grades.... This finding suggests that a math test can also be linguistically demanding. Among the academic language features included in the analyses, there were very few occurrences of academic grammar and discourse features in either of the content tests." (p. 155)</p> <p>"The results yielded a stronger association between the linguistic rating and DIF statistics for ELL students in the "relatively easy" items than in the "not easy" items. Particularly, general academic vocabulary and the amount of language in an item were found to have the strongest association with the degrees of DIF, particularly for ELL students with low English language proficiency. Furthermore, the items were grouped into four bundles to closely look at the relationship between the varying degrees of language demands and ELL students' performance. Differential bundling functioning (DBF) results indicated that the exhibited DBF was more substantial as the language demands increased. By disentangling linguistic difficulty from content difficulty, the results of the study provide strong evidence of the impact of linguistic complexity on ELL students' performance on tests." (p.139)</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Study Wolf et al. (2009), #766
A	Study Citation	Wolf, M. K., Kim, J., Kao, J. C., & Rivera, N. M. (2009). <i>Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a math assessment</i> (CRESST Report 766). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). 69 pp. Retrieved 2/4/12 from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED507754&site=ehost-live .
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	To examine the effectiveness and validity of glossary and read-aloud accommodations for English language learners.
B	Description of accommodation(s)	(1) English glossary: same items as the standard version except for the "addition of an English-to-English glossary appearing in the right margin. Only non-content (i.e., non-math) terms [and some phrases] were glossed, and glossed words appeared next to their corresponding test item in the order of appearance within the item" (p.12); (2) read-aloud: administrators read a script prepared specifically to meet each state's standards concerning numbers and symbols and students received the standard written version of the test (see p.13).
C	Computer delivery of accommodation (yes/no)	No
D	Test Content	math
E	Age or grade	8
F	Disability	none specified.
G	CBT? (yes/no)	No

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study Wolf et al. (2009), #766
H	Research design	(1) For the quantitative analyses, participants were classified by ELL status. Individual students from ELL and non-ELL groups were randomly assigned to one of three testing conditions. Former ELLs were included in the test administration but there were too few in number for analyses and were excluded from the regression analyses. (2) For the qualitative analyses, students' think-aloud responses were elicited for a sample of items using a retrospective interview protocol.
I	Data analysis technique	(1) For the quantitative portion, they used a combination of regression analyses and hierarchical linear models having dummy variables for each of the accommodations. ELL and non-ELL groups were analyzed separately per state. (2) For the student verbal protocol analyses, they coded each interview and reported descriptive statistics.
K	Sample size category	(1) Regression analyses. The number of cases per accommodation per ELL status group and per state were typically less than 50 (small) but the non-ELL groups for one state tended to be slightly larger than 50 (medium). The sample sizes per state per ELL status group used in the regression analyses were in the range 115-134 (medium). (2) Verbal protocol analysis. Close to 50 for the ELL group and 15 for the former ELL group (small to medium).
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	Separately identified and excluded from both the ELL and non-ELL groups in the regression analyses although means and SDs were reported for this group by state for all three experimental conditions and the verbal protocol analyses.
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations	Yes
NONE	How restricted were time limits	Classified as restricted because it was an experimental test with an unknown degree of speededness.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Study Wolf et al. (2009), #766
	for the original test?	Specifically there were 35 multiple choice and 2 open-ended items for 45 minutes of administration.
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	We can infer that the same restrictions in time were applied for the glossary condition because "Standard and Glossary conditions were sometimes administered together in the same room" (p 15). Authors that that the "Read Aloud was always administered in a separate room" and that "Test administration was completed in one to two class periods (approximately 50–90 minutes), depending on the condition." (p. 15) It's not clear, however, if the glossary condition was always limited to 45 min. They imply that at least the read aloud condition was typically given more time.
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	Not curriculum in classrooms per se but they ensured that the items proportionately reflected the Grade 8 math standards for the two states in the study. Also, they ranked the linguistic complexity of the items and found that "results were comparable in terms of the range of the rating scores as well as the mean rating scores " [for the corresponding state assessment] (p.12).
NONE	Was English language proficiency measured?	Yes
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	Yes
NONE	Did the language of the accommodation (English or native) match the language of instruction for participants?	Language of instruction was not stated and cannot be inferred because the name of the states was anonymous.

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Study Wolf et al. (2009), #766
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' <i>home</i> language?	N/A
NONE	Was native language proficiency measured?	No
NONE	Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	No
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	N/A

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Study Wolf et al. (2009), #766
L	Findings	<p>"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.....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." (p.47)</p> <p>"As for the read-aloud accommodation, the statistical analysis yielded mixed results on its effect on the students' performance on a math test [non-significant for State X but positive and significant for State Y]....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." (pp. 47-48).</p> <p>"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...[the results suggested] 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." (p. 48)</p> <p>"In both states' samples, no significant interaction effect was found between the given accommodation and students' ELP 49 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." (pp. 48-49)</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study Wolf et al. (2009), #766
		<p>The authors did not provide effect sizes such as Glass's index for mean effects; given the low statistical power, some non-trivially large effect sizes for the accommodations may not have reached statistical significance</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study by Young et al., 2008
A	Study Citation	Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. <i>Educational Assessment</i> , 13 (2-3), 170-192.
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	Evaluate the psychometric properties of accommodated versions of tests as compared with the original versions. These analyses included the test structure, reliability, group differences in item functioning,
B	Description of accommodation(s)	Orally translated test directions and bilingual glossaries in a state's operational testing for NCLB accountability
C	Computer delivery of accommodation (yes/no)	No
D	Test Content	Math, algebra, and science (state's standards-based assessment)
E	Age or grade	5 and 8
F	Disability	No
G	CBT? (yes/no)	No
H	Research design	Correlational study. Not an experimental study. Students were assigned to accommodation conditions according to need as perceived by school personnel.
I	Data analysis technique	Confirmatory and exploratory factor analysis; differential item functioning, reliabilities, descriptive statistics on test performance
K	Sample size category	13 of 16 groups had large sample sizes. The other 3 groups had medium sample sizes (183-310).

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study by Young et al., 2008
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	No
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations	Yes
NONE	How restricted were time limits for the original test?	Not mentioned; however, most state assessments are power tests.
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	Same as above.
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	Assessments were constructed according to state standards but no quantitative data was provided on alignment to curriculum in the classrooms for ELLs.
NONE	Was English language proficiency measured?	Not reported

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study by Young et al., 2008
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	Other than ELL status, no.
NONE	Did the language of the accommodation (English or native) match the language of <i>instruction</i> for participants?	Not reported.
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' <i>home</i> language?	Not reported
NONE	Was native language proficiency measured?	No
NONE	Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	No
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	N/A

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Study by Young et al., 2008
L	Findings	<p>"There was little evidence of differential test validity in terms of internal test structure or item functioning for these examinee groups, when the performance of non-ELLs and ELLs were examined and compared" (p. 190).</p> <p>"The item-level and item parcel EFA and CFA results showed that the tests are essentially unidimensional for ELLs, with or without accommodations, and non-ELLs, which is reassuring for the purposes of ascertaining construct validity. However, for ELLs, there appears to be more construct irrelevant noise possibly affecting the magnitude of the first eigenvalue, as this is generally smaller for them when compared to non-ELLs. Further, the factor analysis results indicate that the use of one of the ELL testing accommodations, access to translation glossaries/word lists, was effective for supporting the unidimensionality of some of these assessments. In addition, the use of translation glossaries/word lists appears to have a more beneficial effect for ELL examinees on the eighth-grade assessments than on the fifth-grade assessments, by making a stronger case for unidimensionality" (p. 189).</p> <p>"The DIF analyses showed that group differences in performance on the test items, after matching on total test score, are small. This indicates that almost all of the items functioned appropriately, in that significant DIF was rarely observed for ELLs, with or without accommodations" (p.189).</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Young et al. (2011) Study
A	Study Citation	Young, J. W., Holtzman, S., Steinberg, J. (2011). <i>Score comparability for language minority students on the content assessments used by two states</i> (Research Report. ETS RR-11-27). Princeton, NJ: Educational Testing Service
J	Research purpose (compare scores, evaluate test structure, predictive validity, item comparability)	To examine score comparability (reliability, internal test structure, and differential item functioning) for selected 4 NCLB-mandated content assessments in two states
B	Description of accommodation(s)	No description or identification of accommodations by type. There is no mention of excluding ELLs receiving accommodations.
C	Computer delivery of accommodation (yes/no)	Unknown
D	Test Content	math and English and language arts (ELA)
E	Age or grade	Grades 4 & 8
F	Disability	Explicitly excluded.
G	CBT? (yes/no)	No
H	Research design	Correlational study.
I	Data analysis technique	Confirmatory and exploratory factor analysis, DIF analyses
K	Sample size category	All eight groups had large sample sizes (smallest one had 577 cases).

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Young et al. (2011) Study
NONE	Were former ELLs separately identified and excluded from the non-ELL group?	Yes
NONE	Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations?	Unknown
NONE	How restricted were time limits for the original test?	State assessments are usually power tests
NONE	How restricted were the time limits for the accommodated test as compared with the original test?	Same as above
NONE	Was there any assessment of the degree to which the test matched the curriculum received by participants?	No quantitative analysis as to whether curriculum in ELL classrooms was well represented in the test, but the items were based on state standards.
NONE	Was English language proficiency measured?	No

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Young et al. (2011) Study
NONE	Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	No
NONE	Did the language of the accommodation (English or native) match the language of <i>instruction</i> for participants?	N/A
NONE	For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' <i>home</i> language?	N/A
NONE	Was native language proficiency measured?	No

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Young et al. (2011) Study
NONE	Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	N/A
NONE	For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	N/A

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spread-sheet	Field/Variable	Characteristics of Young et al. (2011) Study
L	Findings	<p>"This investigation showed that, for the assessments we studied, there is a high degree of score comparability for language minority students on the content assessments used by these two states"</p> <p>(p. 17).</p> <p>"For the State A assessments, the internal consistency reliability values of the assessments were comparable across all of the language proficiency groups... For State B, the reliabilities for the assessments were generally comparable across all three language proficiency groups, with the only exception being that the reliability estimate for the ELLs on the Grade 8 Mathematics test (0.77) was lower than for the native English speakers (.84) or the former ELLs (.86)" (pp. 15-16).</p> <p>"The factor analysis results provided evidence that one-factor models fit well with data from each of the assessments...Using the invariance criterion of a change in the CFI of less than 0.01, we found invariance of the [basic structure,] factor loadings and of the factor errors of measurement [across ELLs, former ELLs, and non-ELL groups] for all seven assessments [four in State A and three in State B]... In summary, these results provide compelling evidence that similar factor structures exist for students in the different language proficiency groups" (p. 13).</p> <p>In terms of DIF, the Grade 4 assessment in State A showed fewer discrepant items as compared with Grade 8 in the same state. The Grade 4 math assessment showed no DIF among 37 items in all three contrast pairs, and the Grade 4 ELA assessment showed only 1 out of 14 items with discrepancy in the ELLs vs. native speakers contrast.</p> <p>"For the assessments from State B, the only item that exhibited C-level DIF was an item on the Grade 8 English-language arts assessment [which had 37 items]. This item exhibited DIF in favor of the native English speakers when compared with the ELLs and also in favor of the native English speakers when compared with the former ELLs"</p>

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Young et al. (2011) Study
		(p.15).

Primary Variables for Accommodation Studies

Corresponding Column in ETS Spreadsheet	Field/Variable	Characteristics of Young et al. (2011) Study
		<p>The Grade 8 math assessment in State A (44 items) showed 3 discrepant items in the contrast between ELLs vs. native speaker, 1 discrepant item in the contrast of former ELLs vs. native speakers, and 0 discrepant items in the contrast of ELLs vs. former ELLs. The number of discrepant items in the ELA Grade 8 assessment (42 items) for these groups were 4, 3, and 1, respectively. Overall, the number of discrepant items favoring native speakers vs. were approximately half of the items with DIF; i.e., there was approximately the same number of items that favored ELLs or former ELLs among DIF items.</p>

Notes

The Primary Variables worksheet includes the same titles for variables as the worksheet original worksheet provided by ETS. However, columns and rows have been transposed to allow greater legibility for the entries such as the citation, purpose, and findings that require many words and characters. To show the correspondence between original columns and transposed rows, the row number corresponding to each column is indicated in the Primary Variables worksheet.

A second difference is that the variables related to research design/sample size have been grouped all together, preceding findings. Findings begin in row 25 and may take up several rows.

A third difference is that additional variables have been added that are necessary to understand the effect of accommodations on ELLs according to past research (see Abedi et al., 2004; Pennock-Roman & Rivera, 2011). These include the following research design variables:

	Row number
Were former ELLs separately identified and excluded from the non-ELL group?	13
Did all participants in the accommodated group receive exactly the same accommodation or the same combination of accommodations?	14
How restricted were the time limits for the original test? Power= essentially no restrictions for all students; Not power=restricted time limits	15
How restricted were the time limits for the accommodated test as compared with the original test?	16
Was there any assessment of the degree to which the test matched the curriculum received by participants?	17
Was English language proficiency measured?	18
Was level of English language proficiency among ELLs included as a variable that could interact with the accommodation?	19
Did the language of the accommodation (English or native) match the language of instruction for participants?	20
For accommodations involving a native language version or a bilingual dictionary, did the language of the accommodation match the participants' language ?	<i>home</i> 21
Was native language proficiency measured?	22
Was level of literacy/proficiency in the native language among ELLs included as a variable that could interact with the accommodation?	23
For accommodations involving linguistic simplification, how many (or what percent of) items underwent substantial modification?	24

For category options for sample size see:

<http://apps.cehd.umn.edu/nceo/accommodations/AdvancedSearch.aspx>

For pdf files, page numbers for quotes reflect the page number included in the manuscript itself (with the first page of the introduction being p. 1), not the Adobe file page number that begins at 1 with the cover page..