Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2020, Vol. 39, No. 4, pp. 41–52

Examining Effectiveness and Validity of Accommodations for English Language Learners in Mathematics: An Evidence-Based Computer Accommodation Decision System

Jamal Abedi, Yu Zhang, and Susan E. Rowe, *University of California, Davis*, and Hansol Lee, *Korea Military Academy*

Research indicates that the performance-gap between English Language Learners (ELLs) and their non-ELL peers is partly due to ELLs' difficulty in understanding assessment language. Accommodations have been shown to narrow this performance-gap, but many accommodations studies have not used a randomized design and are based on relatively small sample sizes. Addressing such issues, we administered a standard-based mathematics assessment to approximately 3,000 Grade 9 ELL and non-ELL students under five different language-based accommodations. Results indicate that many of these accommodations did not produce significant gains for the recipients. Some even had a negative impact. We believe several factors may explain these findings. First, newer assessments, including those developed for this study, may have been linguistically modified to the point that further modification has only a limited effect. Second, the language of instruction may have not adequately prepared students for the assessment. If the language of instruction (textbook, etc.) contains unnecessary linguistic complexity, then students may not have had the opportunity to learn the assessed content. A third factor is students' unfamiliarity with these accommodations because they are seldom used in classroom instruction and teacher assessments. We discuss our findings and implications for policymakers, assessment developers, practitioners, and researchers.

Keywords: accessibility, accommodations, assessment, English learners, linguistic modification

Introduction

Improving the quality of measurement and learning for English Language Learners (ELLs) is of great importance as they are the fastest growing subpopulation of students in the United States. From 2000 to 2016, ELL enrollment increased nationally from 3.8 million to 4.9 million students (National Center of Education Statistics, 2019). According to a report by the U.S. Government Accountability Office, ELL enrollment has grown to more than 5 million (National Center of Education Statistics, 2018), representing approximately 10% of all public-school students.

However, ELL assessment performance remains substantially below their non-ELL peers (reclassified ELLs and native speaker; Abedi, 2014; Abedi & Ewers, 2013; Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Pennock-Roman & Rivera, 2011; Sireci, Li, & Scarpati, 2003). In 2017, the National Assessment of Educational Progress (NAEP) reported that "ELs (i.e., English learners) were far behind the proficiency rates of non-ELs. For example, while one-third of non-ELs were proficient in mathematics in Grade 8, just 6 percent of ELs attained this level" (NAEP, 2017, para. 3). NAEP also reported that between 2009 and 2017 nearly half of all states experienced a decrease in the percentage of Grade 4 ELLs proficient in mathematics. Although NAEP reading proficiency rates improved overall for Grade 8, ELL performance levels remain extremely low, with just 4.3% of Grade 8 ELLs in California reaching proficiency and just .7% in Hawaii (NAEP, 2017).

ELLs may also be disadvantaged by their inability to communicate effectively in English, especially when asked to complete core subject tasks that require English language proficiency (Abedi & Levine, 2013). Research suggests that assessment items with unnecessary linguistic complexity, especially in mathematics, have a potentially negative impact on the assessment outcomes of ELLs' subject mastery (Abedi, 2004; Abedi & Lord, 2001). Language-based assessment accommodations have shown promise for increasing ELL assessment accuracy, thereby reducing construct-irrelevant variance due to unnecessary linguistic complexities (Abedi, 2007; Abedi & Herman, 2010). Studies have also shown that ELLs who received appropriate accommodations significantly outperformed students receiving either inappropriate or no accommodations (Abedi & Gándara, 2006; Bailey & Kelly, 2010; Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007; Martiniello, 2009). Effective and valid accommodations should provide a fair and equitable opportunity for ELLs to

meaningfully participate in national and state assessments. This study uses a computerized assessment system and demonstrates the potential use of the system to evaluate students' needs for an accommodation. The system assigns students accommodations based on their language background measured by TIMER test scores (English and Spanish) and ELL students' language background questionnaire; thus, it is called an *evidence-based system.*

The common practice across the states is to "use" accommodations in the assessment of ELL students but there is not much information on which accommodation to use. Rivera and Collum (2005) indicated that of the 75 accommodations identified in states' assessment policies, 31 were designed explicitly to meet the needs of students with physical or cognitive disabilities and not the linguistic needs of ELLs. These accommodations were distributed across three traditional accommodation categories i.e., setting, presentation, and response accommodations for ELLs. However, there is not much information on which accommodation to use (Rivera and Collum, 2005. P. 46). The remaining 44 accommodations were relevant either exclusively to ELLs or to both ELLs and students with disabilities. For many of these accommodations there is lack of research-based evidence on their effectiveness and validity. Abedi and Ewers (2013) provided a comprehensive summary of literature on these accommodations and concluded that some of these accommodations may not be used due to the lack of research support. One of the most commonly used accommodations is extended time, however, literature on the effectiveness and validity of this accommodation is mixed. Some studies support its effectiveness while other studies report a lack of effectiveness, ultimately there is insufficient evidence on the validity of this accommodation (see Abedi & Ewers, 2013, pp. 60–62 for detailed discussion of effectiveness and validity for this accommodation).

Effectiveness and Validity of Accommodations

An accommodation is effective if it makes assessments more accessible to the recipients by controlling for construct-irrelevant factors. For example, an effective accommodation for ELL students makes assessment more linguistically accessible by reducing the level of unnecessary linguistic complexity as a construct-irrelevant factor. Accommodations—such as a glossary of uncommon or difficult vocabulary, native language assessment, or customized dictionaries—can be quite helpful and effective for ELLs (Li, Zhong, & Suen, 2012; Wolf, Kim, & Kao, 2012; Willner et al., 2009).

An accommodation is valid if it does not alter the focal construct or does not provide an unfair advantage to the recipients so that the outcomes of accommodated and nonaccommodated assessments should be comparable. That is, the assessment outcomes ideally represent a true individual's ability (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Invalid accommodations affect the validity of assessments for individual students as well as for the group in which students belong. If accommodations affect the construct, then the accommodated and nonaccommodated assessment outcomes cannot be aggregated. Studies have found that some forms of accommodations may alter the construct being measured (Abed & Ewers, 2013; Francis et al., 2006; Sireci et al., 2003). To examine the validity of accommodations, comparing test performance of non-ELL students who are randomly assigned to an accommodation (the treatment group) with non-ELLs who are tested under the standard testing condition (no accommodation provided) can be suggested. If non-ELLs who are tested under the accommodation perform significantly higher on a content-based assessment such as a mathematics test than non-ELL/nonaccommodated students, or if this difference (accommodated verses nonaccommodated) is larger for non-ELL, then the accommodation may have done more than what is supposed to do, i.e., it alters the focal construct (Kieffer, Lesaux, Rivera, & Francis, 2009).

Previous research on accommodations used for ELLs has focused on only a handful of those commonly used in practice, leaving many others unexamined (Francis et al., 2006; Sireci, et al., 2003). Among the few accommodations that have been researched, results have been mixed, calling for more rigorous research to identify accommodations that make assessments more accessible for ELL students. A meta-analysis on the effectiveness and validity of accommodations for ELLs (Kieffer et al., 2009) found that only one of the seven accommodations studied, the use of English dictionaries or glossaries, had a statistically significant (although small) effect on ELL performance. The use of this accommodation resulted in a reduction of 10–25% in the performance-gap between ELLs and native English speakers. A total of 11 studies were cited in the Kieffer et al. meta-analysis.

Another meta-analysis by Pennock-Roman and Rivera (2011) revealed that the effects of accommodations were different depending on students' level of English proficiency. For instance, based on 14 studies with random assignment they found that the modified English accommodation was more effective for ELLs with high intermediate English proficiency than for ELLs with lower levels of English proficiency, but this accommodation had very small average effect sizes ($d = .053 \sim .108$). They further noted that many studies lack the use of random assignment in their methodology, thus limiting the accuracy of inferences that can be made. Further, a fair number of studies did not include a power analysis and did not provide details of the process used for sample selection or how decisions regarding sample sizes were made. These design factors are important for making valid judgments about the adequacy of sample sizes (Ketterlin-Geller, Alonzo, Braun-Monegan, & Tindal, 2007; Kopriva et al., 2007).

Present Study

Consequently, this study seeks to fill this research gap by using a large sample size, random assignment, and students' language background variables. We also seek to expand the accommodations knowledge base, improve ELL assessment, and enhance the quality of inferences made by educators and policymakers. More specifically, the purpose of this study is to provide strong research-based evidence and recommendations on five of the most commonly used language-based accommodations including: (1) Linguistic Modifications, (2) English Read-Aloud, (3) English Glossary, (4) Spanish Math Test, and (5) Bilingual Glossary.

The extra-time accommodation was not included in this study for two main reasons: (1) schools limited our total testing time (including the student background questionnaire, TIMER tests, and the math test) to two class-periods, approximately 50 minutes each, as one of the conditions for their participation and (2) based on a pilot study results ($N =$

120), 50 minutes was sufficient for answering the 35 multiple-choice mathematics items. When we computed the percentage of mathematics test items not answered (not-reached) the results showed less than 5% of the items were at the end of the test (not-reached) were left blank, suggesting that most students had enough time to answer all mathematics items.

Due to the high number of ELL students with a Spanish language background, only English and Spanish versions of the accommodations were used in the study. In fact, the home language of the majority of ELL students in the United States (76.6%) is Spanish (National Center of Education Statistics, 2019). Because research using computer-based accommodations and assessment is limited (Abedi & Ewers, 2013), an added purpose of this study is to investigate the effectiveness and validity of accommodations using a new computer-based assessment system.

Accommodation Descriptions

Linguistic modification. The purpose of a linguistically modified accommodation is to help ELLs with their English language needs without compromising the validity of an assessment. Several studies suggest that linguistically modified versions of test items are an effective and valid accommodation for ELLs (Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, Hofstetter, & Baker, 2000; Maihoff, 2002; Rivera & Stansfield, 2001). The linguistic modification involves taking existing assessments and modifying the linguistic features in the items to simplify linguistic complexity of the test items, language that is appropriate for the general student population but could be complex for ELLs. For example, converting a phrase from passive voice to active voice is a form of linguistic modification. Some researchers refer to this as linguistic simplification; however, it is not necessarily simplifying a test but rather reducing the linguistic complexity unrelated to the construct.

English read-aloud. All questions on the original form of the test are read-aloud and audio files are available for students to play, pause, and repeat if needed. These files were associated with the assessment and added as an accommodation. For example, see studies by Ketterlin-Geller et al. (2007) and Wolf et al. (2009) regarding the effects of read-aloud/oral administration accommodations for ELLs or students with disabilities.

English glossary. The English glossary provides definitions to English terms that are not content related. In this study's English glossary accommodation, students can click on certain words to cause the definition of the word to pop up. The English glossary provides students with a better understanding of test items without disadvantaging students not receiving this accommodation. The English glossary should simply reduce unnecessary linguistic complexities without altering the focal construct. Once again, the most important condition under which glossaries are constructed is that content related terms should not be glossed (Abedi et al., 1998; Abedi et al., 2000; Wolf, Kim, Kao, & Rivera, 2009).

Spanish Math Test (Spanish Translation) is the translated version of the English test used as an accommodation. In this study, translation from English to Spanish was performed by a UC Davis doctoral student who was a Spanish-speaking content expert meeting translation guidelines (https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf). The key guidelines for Spanish translation include using familiar or accessible Spanish words or phrases for students in the translations; avoiding direct translation (i.e., word-for-word translation) especially in cases where English words or phrases cannot be easily translated into Spanish causing awkwardness or unfamiliarity for students; and reviewing the translated items with multiple reviewers from different Spanish-speaking regions to help control for dialect variation, vocabulary usage, and high frequency words (Solano-Flores, Backhoff, & Contreras-Niño, 2009).

Bilingual (English-Spanish) glossary. Another often used accommodation is a bilingual glossary, simply a translated version of the English glossary so test takers may use their home language to aid in comprehension and meaning-making. Similar to the English glossary, the bilingual glossary is presented as a pop-up glossary. Guidelines were developed from a combination of practical and empirical considerations to aid the translation process:

(1) The bilingual glossary uses the words included in the Spanish version of the test, corresponding with the English version as much as possible.
(2) The content for the bilingual glossary (i.e., single word versus descriptive phrase) is determined from the English glossary and is translated accordingly.
(3) The bilingual glossary includes a word-for-word glossary to the extent possible (to correspond to the English glossary). In cases where English words or phrases cannot be easily translated into Spanish, some adaptations have been made.
(4) The bilingual glossary frequently includes more than one Spanish translation for each word to increase the accessibility and clarity.

Research questions. Drawing from the above discussion and research literature, we seek to expand the ELL accommodations knowledge-based by answering the following research questions.

Research question 1. Do the selected accommodations help improve ELL mathematics performance (effectiveness)?

Research question 2. Do the selected accommodations have an impact on non-ELL (native English speakers and reclassified ELLs) students (validity)?

Research question 3. Do student background variables affect ELL or non-ELL mathematics performance when receiving the selected accommodations (differential impact)?

By differential impact we mean that different accommodations may work differently for students with varying academic backgrounds. For example, some accommodations may be more appropriate for students with higher-level proficiency in academic language.

Methodology

To answer the research questions above, this study was conducted in three phases: (1) develop and field-test a new expert-guided, standards-based assessment to measure Grade 8 mathematics performance; (2) create a computerized system to assess ELL and non-ELL students with and without accommodations; and (3) examine the effectiveness and validity of each of the five accommodations based on

students' demographics and academic background variables. We note that computer technology makes accommodated assessments more accessible to all students and believe that our computer-based system may be broadly useful to practitioners and researchers.

Phase 1: Develop and Field Test a New Standards-Based Assessment

To help answer our research questions, Educational Testing Service (ETS), a partner in this study, developed 65 Grade 8 Common Core mathematics assessment items and field-tested them with approximately 800 students from California and New Jersey. Based on the results (e.g., item difficulty and discrimination), ETS experts selected 35 multiple choice items for the final version of the mathematics assessment. The 35 mathematics items were scored dichotomously, correct or incorrect (correct = "1" and incorrect = "0"). The mean score (which is based on a raw score ranging from 0 to 35 points) for all participating students with complete data ($n = 2{,}965$) was 10.57 and the *SD* was 4.40. The score distribution was positively skewed indicating that the items were difficult for the targeted students possibly because the assessment was based on Common Core State Standards that had yet to be fully or properly implemented in schools. The skewness might also be due to lack of motivation because the test is not part of the student's academic grade, making this a low-stakes test.

Phase 2: Create a Computerized System to Assess ELL and Non-ELL Students with and without Accommodations

A group of content experts and computer specialists jointly developed a new technology-based assessment system to measure student mathematics performance under the study's conditions. Suggestions from the research team and advisory board members were incorporated into the system design. The new assessment system was developed to ensure functionality with both old and new school computers. The goal of the computerized system was to evaluate a student's need for an accommodation and then assign the student to an accommodation based on their language background, which is determined by TIMER test scores (English and Spanish) and the student's response to a language background questionnaire.

The team alpha and beta tested the computerized system and conducted an experimental study to examine the effectiveness of the new system. Based on the results, the team refined testing procedures as well as items, debugged the computer program, and determined a cut-off score for a test called TIMER. TIMER is a short, validated, English proficiency assessment used to measure students' English language proficiency (Abedi, Courtney, Leon, Kao, & Azzam, 2006). TIMER consists of two versions: English TIMER (ETIMER) and Spanish TIMER (STIMER).

ETIMER was developed and validated by UCLA's Center for Research on Evaluation, Standards, and Student Testing (Abedi, Courtney, Leon, Kao, & Azzam, 2006). STIMER is a Spanish translated version of the original TIMER test and was created by the UC Davis research team. Both TIMER versions consist of 10 language fluency items and 75 word-recognition items with a total score of 145 points (one point for each of the 75 word-recognition and seven points for each of the 10 language fluency items).

Phase 3: Examine the Effectiveness and Validity of the Five Language-Based Accommodations

In this phase, relevant data from schools, teachers, and students were collected in order to assign accommodations to students. Background information included student state test scores in English language arts and mathematics plus English language proficiency scores in four domains (reading, writing, speaking, and listening) for ELLs with and without disabilities. And other background variables included gender (Table 7), ethnicity (Table 8), participation in the free lunch program (Table 9), race (Table 10), ELL status (Table 11), student disability type, accommodations assigned by the school.

TIMER Assignment

Based on our prior experience with TIMER along with the experts' suggestions, a cut-off score of 89 was established. We created multiple subgroups for the comparison of both accommodated and nonaccommodated groups using specific cut-off scores, such as ETIMER > 89. Students with an ETIMER score at or above 89 were considered English proficient and were randomly assigned to the English mathematics test with no accommodations (control group) or one of the three English accommodations: Linguistic Modification, English Glossary, or English Read-Aloud. Students with an ETIMER score lower than 89 were considered non-English proficient and took the STIMER. If their score on STIMER was also lower than 89 (neither proficient in English nor in Spanish) they were randomly assigned to one of the following language-based accommodations: Linguistic Modification or Bilingual Glossary.

However, if they obtained a score higher than 89 on STIMER they were further assessed on their Spanish proficiency level by the system asking them to self-report their level of Spanish reading proficiency in a background questionnaire "Read Spanish very well." If a student had a STIMER score equal to or above 89 and answered positively to the self-reported Spanish background question, they were considered Spanish proficient and eligible to receive the Bilingual Glossary or Spanish Math Test accommodation (see Figure 1). However, since STIMER is a more valid measure of the students' proficiency in Spanish, we assigned them to the Spanish-based accommodations even if their response to the student background question revealed low Spanish reading proficiency.

Population and Sample

A priori power analysis cluster randomized trials indicated that a sample of about 2,400 students were needed for the study. Estimates of the key design parameters (ICC, a measure of the between cluster variance as a fraction of the total variance, desired effect size, Type I error rate, and level [percent] of power) were based on prior studies similar to the current study (Abedi, 2009). (For details on power combination and effect size, see Kirk, 1995, pp. 62–64; Taylor et al., 2018.) We recruited 2,965 students from 134 classrooms in California and New Jersey. Because the test was administrated in the fall, Grade 9 students were selected to ensure that they had received a full year of Grade 8 mathematics instruction.

Students with similar academic and language backgrounds were grouped into pairs within classrooms. From each pair

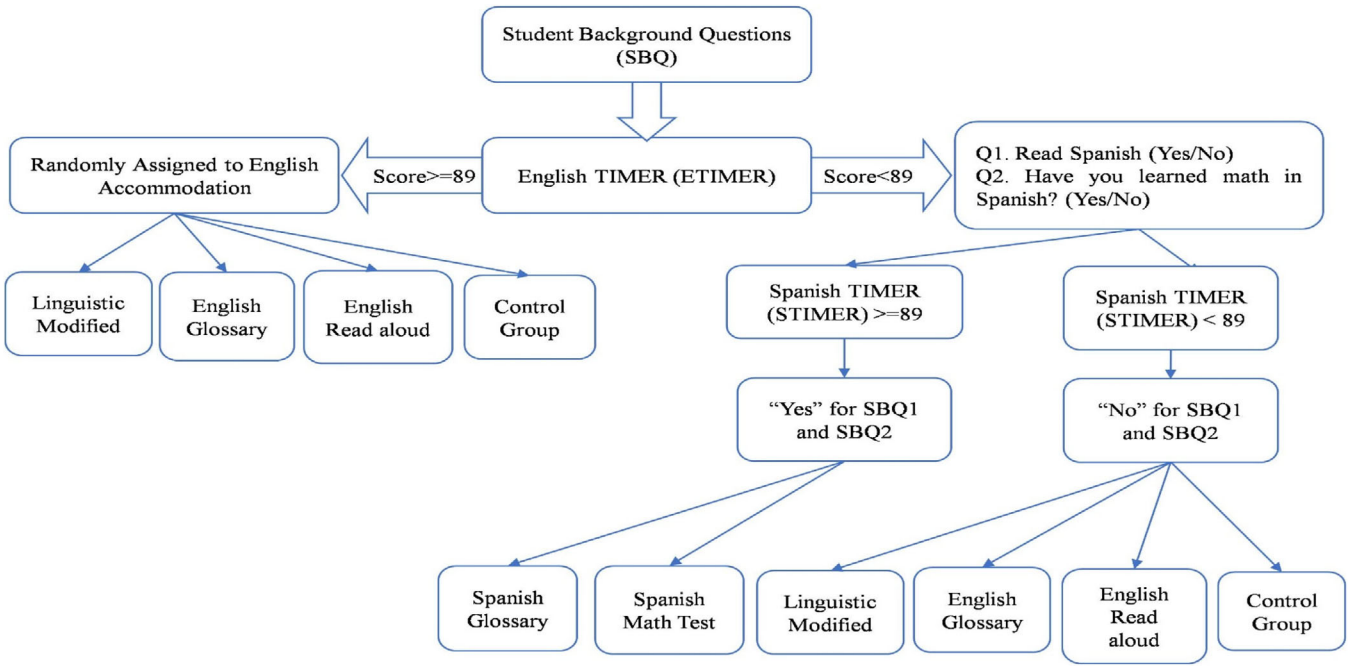

FIGURE 1. TIMER Score Accommodation Chart. [Color figure can be viewed at wileyonlinelibrary.com]
*Since STIMER is a more valid measure of the students' proficiency in Spanish than the background questionnaire, the computer system assigned the students' to one of the Spanish-based accommodations even if their responses to SBQ 1–2 were "no."

of two students, we randomly assigned one student to the treatment group (accommodation condition) and the other to the control group (no accommodation) to control for initial differences. Such random assignment of students to different accommodation conditions within a classroom controls for many of the sources of error that may threaten the internal validity of the design (see McMillan & Schumacher, 1997, pp. 183–190). Among the sources of threat to internal validity is "selection," which in this case refers to the initial differences between the accommodated (treatment) group and nonaccommodated (control) group. The most important sources contributing to initial differences between the two groups are the student's academic and language backgrounds. By random assignment of students to the treatment and control conditions within classrooms, the possibility of such initial differences was greatly reduced. Students were tested under one of the six testing conditions, five accommodation conditions and the control condition of English mathematics test with no accommodations.

Statistical Design

Because students were nested within schools (teachers), a hierarchical linear model may seem to be an appropriate method for analyzing the data. However, due to the nature of the study and research questions, other approaches could also be relevant for several reasons. First, treatment conditions were assigned to students within classrooms; therefore, teachers and schools are not linked directly to a particular accommodation condition. Second, we were interested in both effectiveness and validity of individual accommodations, not "accommodated versus nonaccommodated" in general. Therefore, we needed to conduct five comparisons for ELLs, one for each of the five accommodations (comparing each accommodation with the control group) to examine effectiveness, and another set of five comparisons for examining validity of the accommodations. Assuming we choose to examine our null hypothesis at the .05 nominal level, conducting ten individual t-tests simultaneously increases the Type-I error rate from .05 to .40 (i.e., $1 - (1 - .05)^2$) (see Kirk, 1995, p. 120).

To avoid this problem of inflation of the Type-I error rate, we used a multiple t-test procedure (Kirk, 1995) to detect possible significant differences between students receiving one of the five accommodations (treatment group) and those receiving no accommodations (control group). For computing multiple *t*-tests, we used the pooled within subject variance. First we ran the omnibus test using analyses of variance (ANOVA & ANCOVA) with the total math score as the dependent variable and the accommodation condition as the independent variable. Results indicated a significant difference between the different accommodation conditions ($F_{5,1524} = 4.39, p = .001$). This result was expected since the students' performance under accommodated and nonaccommodated conditions is anticipated to be different. The main reason for conducting this overall analysis was to calculate the pooled mean square within to be used in the computation of a set of *a priori* pairwise comparisons. The pooled within subject variance from this analysis (MS = 16.72) was used as the denominator for several comparisons. This approach was used to avoid inflation of a type-I error rate (see Kirk, 1995, p. 13).

We also conducted a two-level HLM analysis (Level 2: Teachers; Level 1: Students). For the student level, we included the Linguistic Modified, English Read-Aloud, and English Glossary accommodations. However, we did not include the Bilingual Glossary or Spanish Math Test accommodations in the ELL model ($n = 98$ and $n = 75$, respectively), or the non-ELL model ($n = 9$ and $n = 1$, respectively), due to the small sample sizes in these two accommodations. For the teacher level, we included teachers' average of their students' Grade 8 Smarter Balanced Assessment Consortium (SBAC) math score and percentage of students in the teachers' school

Table 1. Mean Test Scores by Demographic Group

	N	*Mean*	*SD*
ELL status			
ELL	1,530	9.85	4.11
Formerly ELL	645	11.58	4.22
Native English speaker	790	11.15	4.82
Total reported	2,965		
Gender			
Male	1,545	10.24	4.39
Female	1,301	11.01	4.42
Total reported	2,846		
Ethnicity			
Hispanic/Latino	2,252	10.41	4.15
Not Hispanic/Latino	527	11.55	5.24
Do you read in Spanish?			
Yes, very well	1,299	10.08	4.13
Yes, but only a little	1,037	10.99	4.48
No, I don't	612	10.97	4.70
Have you learned math in Spanish?			
Yes	931	9.67	4.07
No	2,012	11.03	4.48

participating in the National School Lunch Program (NSLP) as predictors. In total, 226 teachers participated in the study and on average each classroom had over 20 participating students.

Results

Table 1 presents mean test scores by demographic group. Of the total students, 1,545 (54.3%) were male, 1,301 (45.7%) were female. Further, 1,530 (51.6%) were ELLs, 645 (21.8%) were former ELLs, and 790 (26.6%) were Native English speakers. It must be indicated at this point that totals do not necessarily match because not all students responded to each background question.

Of the 1,530 ELLs, 430 (28.1%) took the mathematics assessment with no accommodations ($M = 10.00, SD = 4.16$); 349 (22.8%) tested using the Linguistic Modification accommodation ($M = 10.30, SD = 3.98$); 286 (18.7%) tested with the English Read-Aloud accommodation ($M = 9.63, SD = 4.59$); 292 (19.1%) tested with the English Glossary accommodation ($M = 10.02, SD = 3.85$); 75 (5%) tested using the Spanish Math Test ($M = 8.41, SD = 3.68$); and 98 (6.4%) tested with the Bilingual Glossary accommodation ($M = 8.80, SD = 3.57$) (see Table 2).

Of 1,435 non-ELLs tested, 361 (25.2%) took the original English mathematics test with no accommodation ($M = 11.29$, $SD = 4.57$); 356 (24.8%) tested with the Linguistic Modification accommodation ($M = 11.\ 90, SD = 4.50$); 334 (23.3%) tested with the English Read-Aloud accommodation ($M = 10.75, SD = 4.44$); 374 (26.1%) tested with the English Glossary accommodation ($M = 11.48, SD = 4.65$); 1 (<.1%) took the Spanish Math Test accommodation; and 9 (<.1%) used the Bilingual Glossary accommodation ($M = 7.78, SD = 4.50$). The number of students assigned to Spanish-based accommodations was limited by the assignment mechanism (see Figure 1) and resulted in a small number of non-ELLs in the Spanish Math Test and Bilingual Glossary accommodations. A visual inspection of the distribution of mathematics test scores across accommodations reveals that the accommodations used in this study do not change the score distribution.

Further, we examined distributions of the students background variables across different testing conditions (5 accommodations plus the control group). Tables 7–11 present results of these analyses. Students' background characteristics in terms of gender (Table 7), ethnicity (Table 8), free-reduced lunch program (Table 9), race (Table 10), and ELL status (Table 11) were similar across the treatment and control conditions. Therefore, it seems that the likelihood of initial differences was greatly reduced.

Research question 1: Do selected accommodations help improve ELL mathematics performance (effectiveness)?

An accommodation is effective if it helps make assessments more accessible to the intended student population (ELLs). For Research question 1, we compared performance of ELLs under a particular accommodation with those who did not receive the accommodation (control) to see whether the accommodated ELLs perform significantly higher than nonaccommodated students with the same academic backgrounds. Mean differences between the following three English accommodated groups, Linguistic Modification, English Read-Aloud, and English Glossary, were not statistically significant when compared with the control group (see Table 3); therefore, none of the three English accommodations were shown to be effective in reducing the performance-gap between ELLs and non-ELLs.

Research question 2: Do the selected accommodations have an impact on non-ELL (native English speakers and reclassified ELLs) students (validity)?

As discussed earlier, accommodations should not alter the focal construct, in this case Grade 8 mathematics. To answer this question, we examined the performance of non-ELL students (English speakers and reclassified ELLs) under accommodated and nonaccommodated conditions. Multiple group t-tests indicated that the group mean score of non-ELL students using any of the three English accommodations (Linguistic Modification, English Read-Aloud, or English

Table 2. Descriptive Statistics for the Math Score for ELL and Non-ELL by Test Accommodations

	ELL			Non-ELL		
Test Version	**N**	***Mean***	***SD***	***N***	***Mean***	***SD***
English Math Test (control group)	430	10.00	4.16	361	11.29	4.57
Linguistic Modified	349	10.30	3.98	356	11.90	4.50
English Read-Aloud	286	9.63	4.59	334	10.75	4.44
English Glossary	292	10.02	3.85	374	11.48	4.65
Spanish Math Test	75	8.41	3.68	NA	NA	NA
Bilingual Glossary	98	8.80	3.57	9	7.78	4.50
Total	1,530	9.85	4.15	1,435	11.37	4.55

Table 3. Student's Multiple *t*-Test Comparing the Mean of the Accommodated and the English Mathematics Test for the ELL Group (n = 1,530)

Test Type	*Mean*	*SD*	Student's Multiple *t*-Test	*p* Value
Linguistically Modified	10.30	3.98	1.02	.31
English Read-Aloud	9.63	4.59	–1.11	.27
English Glossary	10.02	3.85	.06	.95
Spanish Math Test	8.41	3.68	–3.10	***
Bilingual Glossary	8.80	3.57	–2.65	***
English Math Test (control group)	10.00	4.16		

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 4. Student's Multiple *t*-Test Comparing the Mean of the Accommodated and the English Mathematics Test for the Non-ELL Group (n = 1,435)

Test Type	*Mean*	*SD*	Student's Multiple *t*-Test	*p* Value
Linguistically Modified	11.90	4.50	1.83	.07
English Read-Aloud	10.75	4.44	–1.57	.17
English Glossary	11.48	4.65	.58	.56
Spanish Math Test	NA			
Bilingual Glossary	7.78	4.50	–2.28	.02*
English Math Test (control group)	11.29	4.57		

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Glossary) were not significantly different than those taking the English mathematics test without accommodations (see Table 4). This lack of significant differences supports the validity of the three English accommodations, that is, the accommodations did not have a significant effect on English speakers and reclassified ELLs; therefore, they did not alter the focal construct. Only one non-ELL student took the Spanish Math test; therefore, we excluded this result in Table 4.

Research question 3: Do student background variables affect ELL or non-ELL mathematics performance when receiving the selected accommodations (differential impact)?

To answer this question, we conducted two HLM analyses (Level 2: Teachers; Level 1: Students), one for ELL students and one for non-ELL students.

On the null model (Model 1), the mathematics total score was 9.45 points out of a possible score of 35 points for ELL students without any predictors (see Table 5). Intraclass correlations were calculated using covariance parameter estimates (Singer, 1998). The teacher covariance component was .805 (SE = .458) and the residual covariance component was 16.456 (SE = .831). The ICC indicated 5.32% of the variance in students' mathematics test scores occurred on the teacher level, leaving 94.67% of the variance on the students' level. The ICC indicates the variance of the dependent variables explained by clustering observations by a particular level. Although the ICC is low, we still want to see if any differences in variation by teachers were found between the ELL and non-ELL models. On Model 2, three accommodations were added on the student level, with random intercept. Model fit improved marginally between Models 1 and 2. There was almost no additional variance that could be explained by using the three accommodations. On Model 3, SBAC math scores had a small although significant effect on performance; model fit improved between Models 2 and 3. Because including NSLP as a predictor for Model 4 did not improve model fit, this predictor was not statistically significant. Model 3 was selected as our final model because it is the best-fitting model while being the most parsimonious.

Table 6 presents the results of a two-level hierarchical linear model for non-ELLs (n = 1,369). The math total score intercept was 11.28 for non-ELL students without adding any predictors in the null model, on average, non-ELL students received a test score of 11.28 points out of a possible score of 35 points. The ICC was calculated using covariance parameter estimates (Singer, 1998). The teacher covariance component was .695 (SE = .338) and the residual covariance component was 20.150 (SE = .775). The ICC indicated that 3.33% of students' test score variance occurred on the teacher level and about 96.67% on the student level. The ICCs for both models indicate that the variance attributed to teachers for the non-ELL model is less than the variance attributed to teachers for the ELL model. For the non-ELL model, results show no statistically significant differences between non-ELLs using English Read-Aloud and English Glossary accommodations. However, the result for using the Linguistic Modification accommodation was statistically significant. The coefficient estimate was .75, p <.05. From Model 1 (null model) to Model 2 (level 1 predictors with random intercept), there was almost no additional variance explained when adding the three accommodations based on model fit statistics. On Model 3, SBAC math scores had no significant effect on performance. Model fit improved little between Models 2 and 3. For Model 4, NSLP was a statistically significant predictor; the coefficient estimate was –5.56, $p < .05$. However, model fit was improved minimally between Models 3 and 4. Model 1 was selected as our final model because it is the best-fitting model while being the most parsimonious.

Discussion

The purpose of accommodations for ELLs is to reduce the performance-gap between ELL students and their non-ELL peers that may be due to construct-irrelevant factors such as unnecessary linguistic complexity (Abedi, 2014). For accommodations to be valid, they should not alter the focal construct, in our case, mathematics. This study (Research question 1) showed that none of the accommodations significantly improved ELL performance. We offer several possible explanations for this lack of effectiveness of the language-based accommodations.

First, given the literature on the general effectiveness of linguistic modifications, we surmise that newer assessments, including those developed for this study, have been linguistically modified to the point that further linguistic modification has only a limited effect. Second, our results also show that some ELL accommodations, including Spanish Math Test and Bilingual Glossary, had significantly lower mean scores than the control group. This may be due to the lack of alignment between language of instruction and language of assessment and issues related to translations. If students are instructed in English, then the Spanish translation and the Spanish glossary may not produce valid results no matter how proficient

Table 5. Two-Level HLM Model on the Students' Math Test Score for ELLs of 1,520 Subjects

	Unconditional Means (Model 1)	Level-1 Predictors (Model 2)	Level-2 Predictors (Model 3)	Level-2 Predictors (Model 4)
Level 1 fixed effects	Est.	Est.	Est.	Est.
Intercept	9.45***	9.43***	9.62***	13.39***
Linguistic Modified	–	.34	.26	.25
English Read-Aloud	–	–.45	–.67*	–.68*
English Glossary	–	.10	-.12	-12
Random effects	Est.	Est.	Est.	Est.
Residual	15.63***	15.56***	16.89***	16.90***
Intercept	.88**	.88**	.64*	.38*
Level 2 fixed effects	Est.	Est.	Est.	Est.
NSLP	–	–	–	–5.88*
SBAC Math	–	–	.00	.00
Model fit				
AIC	8529.5	8529.3	6893.5	6890.3
BIC	8533.2	8536.6	6900.8	6898.7

Note: *$p < .05$, **$p < .01$, ***$p < .001$.

Table 6. Two-Level HLM Model on the Students' Math Test Score for Non-ELLs of 1,369 Subjects

	Unconditional Means (Model 1)	Level-1 Predictors (Model 2)	Level-2 Predictors (Model 3)	Level-2 Predictors (Model 4)
Level 1 fixed effects	Est.	Est.	Est.	Est.
Intercept	11.28***	11.07***	11.18***	14.82***
Linguistic Modified	–	.75*	.78*	.76*
English Read-Aloud	–	–.32	–.31	–.32
English Glossary	–	.40	.43	.40
Random effects	Est.	Est.	Est.	Est.
Residual	20.15***	19.99***	20.10***	20.08***
Intercept	.70*	.69*	.65*	.47
Level 2 fixed effects	Est.	Est.	Est.	Est.
Slope NSLP	–	–	–	–5.56*
SBAC Math	–	–	.00	.00
Model fit				
AIC	8023.0	8018.2	7944.4	7942.1
BIC	8026.2	8024.8	7951.7	7950.4

Note: *$p < .05$, **$p < .01$, ***$p < .001$.

they are in Spanish. Third, we believe the lack of effectiveness of the language-based accommodations may be due to the lack of attention to the language of instruction (textbooks and teacher's instructional language). For example, in Kieffer et al.'s (2009) meta-analysis, the match between the test language and classroom instruction was highlighted to ensure the effectiveness of accommodations. If the language of instruction contains unnecessary linguistic complexity, for example, then students may not have had the opportunity to learn the assessed content. Fourth, another factor that may explain the lack of effectiveness of language-based accommodations is students' unfamiliarity with these accommodations because they are seldom used in classroom instruction and teacher assessments.

The non-ELL (Research question 2) results revealed no significant differences between the performance of accommodated and nonaccommodated non-ELL students except for the Bilingual Glossary and Spanish Math Test. While the Bilingual Glossary and the Spanish Math Test accommodations had lower mean scores, which might be due to the lack of alignment between language of instruction and language of assessment, the overall findings of nonsignificant differences between non-ELL accommodated and non-ELL nonaccommodated students confirm the validity of the language-based accommodations used in this study, i.e., that the accommodations did not alter the focal construct.

The results of the hierarchical linear model examining possible student background effects on ELL and non-ELL performance (Research question 3) indicated that all three English accommodations (Linguistically Modified, English Read-Aloud, and English Glossary) are valid, that is, none altered the focal construct. However, the HLM data showed mixed results regarding the effectiveness of the five accommodations on performance. The Linguistic Modification accommodation improved non-ELL students' performance, although models including this accommodation were not better fitting than the null model, suggesting a small effect size of the Linguistic Modification accommodation for non-ELL students. The effect of the English Glossary accommodation was similar to the Read-Aloud accommodation compared to those who did not receive the accommodation. There was no statistically significant difference between non-ELL who did

Table 7. Accommodation Conditions by Gender

	Male	Female	Total
English math test			
Count	416	344	760
% within accommodation conditions count	54.7%	45.3%	100.0%
Linguistically modified math test			
Count	366	324	690
% within accommodation conditions count	53.0%	47.0%	100.0%
Read-aloud math test			
Count	322	269	591
% within accommodation conditions count	54.5%	45.5%	100.0%
English glossary math test			
Count	352	291	643
% within accommodation conditions count	54.7%	45.3%	100.0%
Spanish math test			
Count	37	32	69
% within accommodation conditions count	53.6%	46.4%	100.0%
Spanish glossary math test			
Count	60	43	103
% within accommodation conditions count	58.3%	41.7%	100.0%
Total			
Count	1,553	1,303	2,856
% within accommodation conditions count	54.4%	45.6%	100.0%

Table 8. Accommodation Conditions by Ethnicity

	Hispanic/ Latino	Not Hispanic/ Latino	Total
English math test			
Count	607	129	736
% within accommodation conditions count	82.5%	17.5%	100.0%
Linguistically modified math test			
Count	521	138	659
% within accommodation conditions count	79.1%	20.9%	100.0%
Read-aloud math test			
Count	442	119	561
% within accommodation conditions count	78.8%	21.2%	100.0%
English glossary math test			
Count	495	133	628
% within accommodation conditions count	78.8%	21.2%	100.0%
Spanish math test			
Count	68	0	68
% within accommodation conditions count	100.0%	0.0%	100.0%
Spanish glossary math test			
Count	97	4	101
% within accommodation conditions count	96.0%	4.0%	100.0%
Total			
Count	2,230	523	2,753
% within accommodation conditions count	81.0%	19.0%	100.0%

not receive the English Glossary accommodation and those who did (mean difference = .05, effect size = .00).

Importantly, our results confirm findings of many other studies noting the continued performance-gap between ELL and non-ELL students, despite the use of large sample sizes, random assignment, matched pairs, and an expert-designed computerized assessment system. We believe that additional studies are needed that examine factors such as teacher interaction with students, students' level of experience with computer assessments and accommodations, teacher motivation and level of expertise, and deeper, focused analyses of performance by ELLs, reclassified ELLs, and non-ELLs.

Table 9. Accommodation Conditions by Free/Reduced Price Lunch

	Free	Reduced Price	Neither	Total
English math test				
Count	133	8	17	158
% within accommodation conditions count	84.2%	5.1%	10.8%	100.0%
Linguistically modified math test				
Count	120	7	13	140
% within accommodation conditions count	85.7%	5.0%	9.3%	100.0%
Read-aloud math test				
Count	105	3	12	120
% within accommodation conditions count	87.5%	2.5%	10.0%	100.0%
English glossary math test				
Count	105	7	18	130
% within accommodation conditions count	80.8%	5.4%	13.8%	100.0%
Spanish math test				
Count	4	1	0	5
% within accommodation conditions count	80.0%	20.0%	.0%	100.0%
Spanish glossary math test				
Count	16	0	5	21
% within accommodation conditions count	76.2%	.0%	23.8%	100.0%
Total				
Count	483	26	65	574
% within accommodation conditions count	84.1%	4.5%	11.3%	100.0%

Table 10. Accommodation Conditions by Race

	American Indian or Alaska Native	Asian	African American	Native Hawaiian or Other Pacific Islander	White	Two or More Races	Total
English math test							
Count	61	22	44	23	247	7	404
% within accommodation conditions count	15.1%	5.4%	10.9%	5.7%	61.1%	1.7%	100.0%
Linguistically modified math test							
Count	49	24	57	17	236	4	387
% within accommodation conditions count	12.7%	6.2%	14.7%	4.4%	61.0%	1.0%	100.0%
Read-aloud math test							
Count	41	21	46	29	165	6	308
% within accommodation conditions count	13.3%	6.8%	14.9%	9.4%	53.6%	1.9%	100.0%
English glossary math test							
Count	45	16	54	26	209	2	352
% within accommodation conditions count	12.8%	4.5%	15.3%	7.4%	59.4%	.6%	100.0%
Spanish math test							
Count	6	0	0	0	20	0	26
% within accommodation conditions count	23.1%	.0%	.0%	.0%	76.9%	.0%	100.0%
Spanish glossary math test							
Count	4	0	0	0	35	1	40
% within accommodation conditions count	10.0%	.0%	.0%	.0%	87.5%	2.5%	100.0%
Total							
Count	206	83	201	95	912	20	1,517
% within accommodation conditions count	13.6%	5.5%	13.2%	6.3%	60.1%	1.3%	100.0%

This study supports the use of computer-based systems in assessment design, accommodations applications, and delivery that would be more difficult in a traditional paper and pencil format. This computer-based assessment system is unique as it first assigns accommodations based on ELL students' related language background variables and then administers and scores the assessments. The computerized assessment system was also useful in helping assign ELLs to accommodations based on the students' language background. Consequently, it is our belief that our computer-based assessment system provides further evidence to support the efficiency of such systems to make assessments and accommodations more accessible to ELLs. Therefore, one of the main contributions of this study is the computerized assessment system which can be shared with researchers and schools along with detailed instructions on how to use the system. This study demonstrates the potential use of computer-based assessment systems in assessing the need for an accommodation and administering assessments with the accommodation.

Conclusion

What is increasingly clear from this study and many other accommodation studies from the past two decades is that while accommodations have an important role to play in the student learning and assessment process, they do not by themselves level the playing field and bring ELL performance to a comparable level to non-ELL peers. Nor, do we surmise, can they compensate for primary ELL learning factors such as teacher instruction and school quality. Researchers should continue to look at the effect of accommodations on ELL performance in a comprehensive manner, investigate why some accommodations have negative effects, and study the role that prior learning and other factors have on ELL performance. They also should continue to develop accommodations and assessment guidelines that support ELL learning. Assessment developers should continue to reduce the unnecessary linguistic complexity and other sources of bias in assessments, especially as standards, teaching, and assessments continue to shift toward deeper problem solving that tends to be more language intensive. Practitioners should adjust, or fine-tune, instruction based on individual and classroom ELL language proficiencies and needs. Finally, policymakers should support a fully integrated educational system including teacher instruction, assessment, and accommodations. Evaluation of results and systemic adjustments are imperative in order to address the persistent ELL learning gap.

This study adds to existing literature as it addresses both issues of effectiveness and validity at the same time. If the focus were solely on the effectiveness of accommodations, then recommending accommodations may not be possible. An accommodation could be highly effective and yet alter the focal construct and thus provide invalid assessment outcomes. Similarly, an accommodation may be valid but may not be effective, implying not enough justification to use that recommendation. Due to the importance of both issues, this study discusses the way in which accommodations could negatively impact the students' score on the test if accommodations are

Table 11. Accommodation Conditions by ELL Status

	ELL	Formerly ELL	Native English Speaker	Total
English math test				
Count	418	157	152	727
% within accommodation conditions count	57.5%	21.6%	20.9%	100.0%
Linguistically modified math test				
Count	347	132	182	661
% within accommodation conditions count	52.5%	20.0%	27.5%	100.0%
Read-aloud math test				
Count	281	140	146	567
% within accommodation conditions count	49.6%	24.7%	25.7%	100.0%
English glossary math test				
Count	290	145	186	621
% within accommodation conditions count	46.7%	23.3%	30.0%	100.0%
Spanish math test				
Count	70	0	0	70
% within accommodation conditions count	100.0%	.0%	.0%	100.0%
Spanish glossary math test				
Count	93	4	5	102
% within accommodation conditions count	91.2%	3.9%	4.9%	100.0%
Total				
Count	1,499	578	671	2,748
% within accommodation conditions count	54.5%	21.0%	24.4%	100.0%

not effective and valid. Furthermore, though we could not examine any impact of students' familiarity with the accommodations, we conjecture that the effectiveness and validity of the accommodations can be maximized if students have adequate experience using them in an instructional setting. Therefore, the discussion of the limitations raises awareness, specifically to teachers and educators, of the importance of increasing exposure to accommodations in the classroom setting. This should help students be more prepared in using their accommodation(s) in assessments.

Finally, it is also important for researchers to examine the effects of students' familiarity with accommodations to measure their effect on ELL performance. Another area for further investigation is to measure students' experience with computerized tests, especially for newly arrived immigrants who may have limited computer skills.

Limitations of the Study

While the research team carefully controlled for many factors affecting internal validity of this study's design, we note some possible limitations. Although the extra-time accommodation should be considered along with some of the accommodations that need additional time, it was not included in this study because schools placed limitations on the testing time allotted for the study. By restricting the time allotted for testing, students may have been rushed to effectively use the accommodations, especially the glossaries and Read-Aloud accommodations, that typically require additional time. Research on the effectiveness of extra-time accommodation is mixed, Pennock-Roman and Rivera (2011) found that accommodations were effective for ELL students when students were allowed additional time to complete the tests.

Another potential issue in this study is that student performance was generally low, possibly because the assessment was based on rigorous Common Core State Standards that had yet to be fully or properly implemented in schools at the time of test administration for this study. It is possible that one or more accommodations could have been significant if the scores were more evenly distributed.

There was also a substantial amount of missing data in the state assessment scores including SBAC mathematics, SBAC ELA and English language proficiency (ELP) assessments. In particular, we were faced with several problems regarding ELP scores since almost half of the sample (48%) had missing data on this variable. Of students with nonmissing data 44.7% had achievement level ratings and the remaining 53.3% reported scale scores. We looked at the state ELP technical manual for the conversion tables but the information was not readily available or out of date. The most serious problem with the ELP score was the missing data issue because by using ELP scores, we would have lost 50% of the sample which would have serious consequences on our analyses and interpretation of the results. Other limitations, as noted earlier, include that several accommodations did not have large enough sample sizes to provide a strong research-based judgment on their effectiveness and validity. Additionally, students may have had differing levels of experience using computerized assessments and accommodations, which may have impacted their performance. Future studies should examine this factor as well.

References

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*(1), 4–14.

Abedi, J. (Ed.). (2007). *English language proficiency assessment in the nation: Current status and future practice*. Davis, CA: University of California, Davis, School of Education.

Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment*, *14*, 195–211, 2009.

Abedi, J. (2014). Accommodations in the assessment of English language learners. In A. J. Kunnan. (Ed.), *Companion to language assessment* (pp. 1115–1129), Malden, MI: Wiley Blackwell.

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 Report 702, 2006). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Ewers, N. (2013). Accommodations for English language learners and students with disabilities: A research-based decision algorithm. LA, CA: University of California, Los Angeles, Smarter Balanced Assessment Consortium, https://www.smartbalanced.org

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*(4), 36–46. https://doi.org/10.1111/j.1745-3992.2006.00077.x

Abedi, J., & Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teachers College Record*, *112*(3), 723–746.

Abedi, J., & Levine, H. G. (2013). Fairness in assessment of English learners. *Leadership*, *42*(3), 26–38. https://eric.ed.gov/?id=EJ1004517

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

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

Abedi, J., Lord, C., Hofstetter, C. & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*(3), 16–26. https://doi.org/10.1111/j.1745-3992.2000.tb00034.x

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association and National Council on Measurement in Education.

Bailey, A. L., & Kelly, K. R. (2010). *The use and validity of home language surveys in state English language proficiency assessment systems: A review and issues perspective*. Los Angeles, CA: University of California Los Angeles, Evaluating the Validity of English Language Proficiency Assessment.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for serving adolescent newcomers*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Ketterlin-Geller, L. R., Alonzo, J., Braun-Monegan, J., & Tindal, G. (2007). Recommendations for accommodations: Implications of (in) consistency. *Remedial and Special Education*, *28*(4), 194–206. https://doi.org/10.1177/07419325070280040101

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis was on effectiveness and validity. *Review of Educational Research*, *79*(3), 1168–1201.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.

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. https://doi.org/10.1111/j.1745-3992.2007.00097.x

Li, H., Zhong, Q., & Suen, H. K. (2012). Students' perceptions of the impact of the college English test. *Language Testing in Asia*, *2*(3), 77.

Maihoff, N. A. (2002). *Using Delaware data in making decisions regarding the education of LEP students*. Presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, Palm Desert, CA.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests, *Educational Assessment*, *14*(3-4), 160–179. https://doi.org/10.1080/10627190903422906

McMillan, J. H., & Schumacher, S. (1997). *Research in education: A conceptual framework*. New York, NY: Longman.

National Assessment of Educational Progress (NAEP) (2017). *Academic performance and outcomes for English learners: Performance on national assessments and on-time graduation rates*. Washington, DC: U.S. Department of Education. https://www2.ed.gov/datastory/el-outcomes/index.html

National Center of Education Statistics (NCES) (2018). *Trends in international mathematics and science study (TIMSS)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/timss/

National Center of Education Statistics (NCES) (2019). *The condition of education*. Washington, DC: Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/programs/coe/

Pennock-Roman, M., & 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. https://doi.org/10.1111/j.1745-3992.2011.00207.x

Rivera, C. & Collum, E. (2005). *State assessment policy and practice for English language learners: A national perspective*. Washington: the George Washington University Center for Equity and Excellence in Education.

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

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

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*, 323–355.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of test translation error. *International Journal of Testing*, *9*(2), 78–91.

Taylor, J. A., Kowalski, S. M., Polanin, J. R., Askinas, K. Stuhlsatz, A. M, Wilson, K. D., Tipton, E., & Wilson, J. W (2018*).* Investigating science education effect sizes: Implications for power analyses and programmatic decisions. *AERA Open, July-September 2018*, *4*,(3), 1–19.

Willner, L. S., Rivera, C., & Acosta, B. D. (2009). Ensuring accommodations used in content assessments are responsive to English-language learners. *The Reading Teacher*, *62*(8), 696–698. https://doi.org/10.1598/RT.62.8.8

Wolf, M. K., Kim, J., & Kao, J. (2012). The effects of glossary and read-aloud accommodations on English language learners' performance on a mathematics assessment. *Applied Measurement in Education*, *25*(4), 347–374. https://doi.org/10.1080/08957347.2012.714693

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