Review of Educational Research
September 2009, Vol. 79, No. 3, pp. 1168–1201
DOI: 10.3102/0034654309332490

Accommodations for English Language Learners Taking Large-Scale Assessments: A Meta-Analysis on Effectiveness and Validity

Michael J. Kieffer and Nonie K. Lesaux
Harvard Graduate School of Education
Mabel Rivera and David J. Francis
University of Houston

Including English language learners (ELLs) in large-scale assessments raises questions about the validity of inferences based on their scores. Test accommodations for ELLs are intended to reduce the impact of limited English proficiency on the assessment of the target construct, most often mathematic or science proficiency. This meta-analysis synthesizes research on the effectiveness and validity of such accommodations for ELLs. Findings indicate that none of the seven accommodations studied threaten the validity of inferences. However, only one accommodation—providing English dictionaries or glossaries—has a statistically significant effect on ELLs' performance, and this effect equates to only a small reduction in the achievement score gap between ELLs and native English speakers. Findings suggest that accommodations to reduce the impact of limited language proficiency on academic skill assessment are not particularly effective. Given this, we posit a hypothesis about the necessary role of academic language skills in mathematics and science assessments.

KEYWORDS: achievement gap, assessment, English language learners, high stakes testing, language development.

As the standards movement in education has gained in momentum, policy makers have increasingly focused on test-based accountability systems with the goal of improving academic achievement for all children. The principles of setting high standards, assessing all students relative to those standards, and holding schools accountable for student achievement have long been central to reform movements in public education (e.g., Fuhrman, 2003). However, since the No Child Left Behind Act of 2001 (NCLB), the application of these principles to subgroups of students identified as particularly at risk for academic difficulties has become very important.

One of these subgroups consists of students who lack full proficiency in English, commonly referred to as English language learners (ELLs). ELLs represent one of the fastest-growing groups among the school-aged population in this nation (e.g., Capps et al., 2005). Speaking a wide variety of languages, this group almost doubled in size between 1980 and 2000, and the most recent estimates place the size of the population at more than 5 million (e.g., Batalova, Fix, & Murray, 2007). The results from many large-scale assessments suggest that when compared to their native English-speaking peers, ELLs lag behind in all grades and content areas. For example, on recent national assessments of reading and math, only a small minority of ELLs scored at proficient levels (4% to 11%, depending on grade and subject), compared to a third or more of native English speakers (National Center for Education Statistics, 2005).

According to many educators, NCLB has succeeded in increasing awareness of the academic needs and achievement of ELLs through new requirements to evaluate schools, districts, and states based on the English and content outcomes of this group of learners (Center on Education Policy, 2006). However, including ELLs in large-scale assessments is not a straightforward undertaking. ELLs present a unique set of challenges for educators and policy makers because of the central role played by language proficiency in the acquisition and assessment of content area knowledge. Thus, many unanswered questions remain about the inclusion of ELLs in large-scale assessments; foremost among them are questions about how valid inferences about ELLs' abilities can be made based on scores from these assessments. The purpose of this study was to determine the effectiveness and validity of test accommodations for ELLs taking large-scale assessments by using meta-analysis to quantify the impact of the specific accommodations on the performance of ELLs and native English speakers.

Including ELLs in Large-Scale Assessments

Historically, ELLs have often been excluded from large-scale assessments because limited English proficiency was thought to prevent students from understanding questions and/or result in invalid test results under standard test administration procedures (Rivera, Collum, & Shafer Willner, 2006). Exclusion of large numbers of students from participation in standards-based tests not only can result in substantial distortion of the percentage of students achieving proficiency but also, more important, can obscure important and systematic differences in student achievement between different demographic groups. Thus, one of the laudable goals of NCLB and state efforts is to increase participation of *all* learners—including those in identified subgroups—in large-scale assessments.

However, it is not enough for students to participate in state assessments; students' participation must lead to valid inferences about their achievement. Obtaining valid results is a particularly pressing issue because the stakes of mandated assessments for states, districts, and schools are high. NCLB and state accountability systems not only place considerable pressure on schools and districts to increase participation rates in large-scale assessments but also impose sanctions on schools that cannot move students in all identified subgroups toward proficiency. In addition, performance on large-scale assessments is increasingly high stakes for

students: By 2008, 28 states in the United States will require that students pass a state-administered test for high school graduation (Fuhrman, 2003).

There is reason for concern about the validity of test scores if in fact these reflect individual differences in abilities that are distinct from those that are the target of assessment (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Because language plays an integral role in most, if not all, academic learning, any test of academic achievement is also, to some degree, a test of language ability. Consequently, ELLs present a special challenge to schools and those involved in large-scale assessment; if tests are not appropriately designed or if ELLs are not tested under appropriate conditions, then language demands of the test that are not central to the target of assessment may unfairly and negatively influence their performance. Research conducted by Abedi and colleagues has demonstrated that there is indeed a substantial link between students' English language proficiency and their performance on tests of math, science, and social studies (e.g., Abedi & Leon, 1999; Bailey, 2005; Butler & Castellon-Wellington, 2005). Furthermore, although there may be substantial differences between ELLs and their peers in content knowledge, research shows that the size of this knowledge gap often depends on the language demands of the assessment. Several correlational studies have found that assessments and individual test items that have more linguistic complexity yield larger performance gaps between ELLs and non-ELLs (e.g., Abedi, Leon, & Mirocha, 2003; Abedi, Lord, Hofstetter, & Baker, 2000; Abedi, Lord, & Plummer, 1997; Martiniello, 2007).

These findings suggest that—contrary to some popular conceptions—assessments in all domains assess language skills as well as content knowledge and skills. However, such a relationship does not lead directly to the conclusion that valid inferences can never be made about the content knowledge of ELLs from large-scale assessments. Rather, the key question is to what extent the language skills measured by these assessments are essential to the construct targeted by the test and, in turn, to what extent they measure language demands that are irrelevant to the academic skills being assessed.

Use of Accommodations for ELLs Taking Large-Scale Assessments

Making specific changes to the test format or the conditions under which students are tested is one method that has been proposed to minimize the influence on content area test performance of variation in ELLs' language skills that is not central to the construct being assessed. Such test *accommodations* include any alteration to standard test administration procedures designed to provide support for students based on their special needs without changing the construct being assessed (AERA, APA, & NCME, 1999). These procedures include the presentation of the assessment items, the ways in which students respond to the items, any equipment or materials to be used, the period of time allowed to complete the test, and the environment in which students take the test. There are as many as 75 different accommodations currently in use with ELLs, although not all of them are appropriate. Moreover, their selection and implementation vary by state and district (for a review of state policies on accommodations for ELLs, see Rivera et al., 2006).

An appropriate accommodation focuses on those extraneous factors that affect the test scores of students with special needs but that are not the target of assessment. An example of an appropriate accommodation would be to provide a large-print version of a test to a student with a visual impairment. At the same time, accommodations should not provide inappropriate support or change the nature of the task such that resulting scores no longer allow valid inferences about the central construct being measured. An example of an inappropriate accommodation would be to rewrite the passages in a reading comprehension assessment in a way that alters their fundamental difficulty level. Thus, for ELLs, appropriate accommodations provide direct or indirect linguistic support to minimize the negative impact of irrelevant language demands on students' performance so that the students can demonstrate their content knowledge and academic skills to the greatest extent possible.

Evaluating Accommodations for ELLs in Large-Scale Assessments

Theoretically speaking, many accommodations that offer linguistic support, such as providing dictionaries or simplifying the English sentence structure of the test items, may indeed be appropriate for ELLs. However, because content knowledge is inextricably linked to language, the use of certain language supports for ELLs may not be as straightforward as providing a large-print version of an assessment to a student with a visual impairment; even language-based accommodations that are grounded in theory may in practice be ineffective or threaten the validity of scores. Thus, the selection of accommodations for ELLs must be based on empirical evidence for their effectiveness and validity (Abedi, Hofstetter, & Lord, 2004).

Although accommodations for ELLs can be evaluated along several dimensions, evaluating accommodations for *effectiveness* and *validity* is of paramount importance. *Effectiveness* refers to the extent to which students receiving the accommodation demonstrate improved test scores. In contrast, the *validity* of an accommodation refers, in part, to the notion that the accommodation should improve the performance of students who require it but not affect the performance of students who do not. If an accommodation affects the performance of students who do not require it, then providing the accommodation to some students but not others would threaten the validity of the resulting test scores. If an assessment is valid for use with a specific group, then students who do not require the accommodation will be neither advantaged nor disadvantaged by receiving it. A growing body of empirical research has evaluated accommodations for ELLs, but the results of these individual studies have yet to be quantitatively synthesized to produce aggregate estimates of their effectiveness and validity.

Moreover, investigation of factors that may potentially moderate the effectiveness of these accommodations (e.g., grade level, domain tested, language of instruction) is needed. It is possible that a given accommodation will be more effective for tests in some domains than for tests in other domains or that accommodations will be more effective at some grade levels than at others. Curricular content and corresponding measures of achievement change with respect to both difficulty (National Center on Education and the Economy, 1998) and the nature of the skills tested (e.g., Koenig & Bachman, 2004; RAND Mathematical Study Panel, 2003; RAND Reading Study Group, 2002) over the course of the grade span, thus potentially influencing the effectiveness of specific accommodations.

This is particularly important in the context of ELLs' test performance given the differing language demands of academic tasks over time and the language demands specific to different domains tested. For example, the fourth grade math test may emphasize and prioritize children's calculation skills, whereas the eighth grade tests in the same content area of math may emphasize complex word problems with sophisticated language. Finally, evaluating accommodations for this population must further recognize potential sources of differential effectiveness by focusing on the instructional and linguistic context in which the testing is occurring, given the differing models of instruction offered for ELLs (Abedi et al., 2004).

Present Study

The purpose of this study is to evaluate the effectiveness and validity of accommodations for ELLs participating in large-scale assessments. Two narrative reviews (Abedi et al., 2004; Sireci, Li, & Scarpati, 2003) have previously synthesized the findings of studies on test accommodations for ELLs published before 2001. The present study was designed to build on this work in two ways. First, using a meta-analytic approach, the current study quantifies the average effects of the accommodations studied. Second, the current study updates the findings of previous reviews by including the findings of several studies published since 2001 as well as those previously reviewed. Given the potential sources of differential effectiveness of accommodations discussed above, the meta-analysis also includes an examination of several moderators of effects. The analyses were guided by two specific research questions:

1. What evidence exists that specific test accommodations are effective in improving the performance of ELLs taking large-scale assessments? What evidence exists that these effects differ as a function of the grade level of students, domain tested, provision of extra time, or language of instruction?
2. What evidence exists that specific test accommodations designed for ELLs are valid in large-scale assessments?

Method

Study Inclusion Criteria

Based on our research questions, we selected four characteristics that formed the criteria for inclusion of studies that provide empirical evidence for evaluating accommodations for ELLs. We included studies in the meta-analysis that (a) examined individual accommodations or individual accommodations bundled with extra time, (b) were articles published in peer-reviewed journals or technical reports available online, (c) employed an experimental, quasi-experimental, or repeated measures design, and (d) reported sufficient data to allow for the estimation of effect sizes.

Search procedure. Studies for review were obtained through two searches conducted in July 2006 designed to include all studies available up to that time. First, we conducted a comprehensive search of online databases, including Education Resources Information Center, PsycINFO, Modern Language Association,

Education Abstracts, and Academic Search Premier (which yielded 114 entries), as well as the online database of the National Center for Research on Evaluation, Standards, and Student Testing (which yielded an additional 27 entries, many of them redundant with the 114 previously found). The abstract of each identified citation was read to determine if it was an empirical study examining the effects of one or more accommodations. Second, we collected citations of studies previously reviewed by Sireci et al. (2003) and/or by Abedi et al. (2004). Based on the list of citations of empirical studies from the two searches, we collected technical reports as well as articles. However, we did not collect presentations at academic conferences because of both practical and quality concerns. In several cases, the results of a single study were reported in multiple documents; in such cases, the documents were linked together and cross-checked for complete information and the most recent document is cited here.

Excluded studies. The search procedure above yielded 21 studies for possible inclusion in the analyses. However, several of these studies, including some cited in previous reviews, had to be excluded from the meta-analysis for reasons of data reporting or methodology. In three instances (N. E. Anderson, Jenkins, & Miller, 1996; Hafner, 2001; Lotherington-Woloszyn, 1993), the studies did not report the necessary information to quantify the effects of accommodations separately for ELLs and native English speakers. In two cases (Abedi & Hejri, 2004; Shepard, Taylor, & Betebenner, 1998), studies examined the effect of various accommodations chosen for individual students by their teachers and thus were inappropriate for examining the effect of specific accommodations. In one case, a previously cited study (Miller, Okum, Sinai, & Miller, 1999) was a conference presentation.

After excluding the studies above, a total of 15 studies remained. Of these studies, 4 (Abedi & Lord, 2001; Albus, Thurlow, Liu, & Burlinski, 2005; Castellon-Wellington, 2000; Johnson & Monroe, 2004) employed repeated measures designs in which the same group of students was tested with and without accommodations. Because the preponderance of the studies to be included employed between-groups designs and because effect sizes from repeated measures designs are not strictly comparable to those from between-groups studies, results from these studies were not included in the formal meta-analysis but were considered in our findings.

Studies Included in Meta-Analysis

In all, 11 studies were included in the meta-analysis with a total of 23,999 participants (17,445 native English speakers, 6,554 ELLs). Of these studies, 6 were conducted by Abedi and colleagues, whereas 5 others were conducted by other research teams (i.e., M. Anderson, Liu, Swierzbin, Thurlow, & Bielinski, 2000; Brown, 1999; Garcia Duncan et al., 2005; Hofstetter, 2003; Rivera & Stansfield, 2004). With respect to design, 8 were true experiments, in which students were randomly assigned to accommodated or unaccommodated conditions, whereas 3 (Abedi, Courtney, & Leon, 2003a; Abedi, Courtney, & Leon, 2003b; Brown, 1999) were classified as quasi-experiments because of factors specific to each study. In the study by Brown (1999), the mechanism of assignment is unclear in the report and could not be confirmed through communications with the study author or school personnel involved in the study. Observed pretest differences between the two groups were negligible. In the study by Abedi, Courtney, et al. (2003a),

students were originally assigned at random to a treatment condition; however, not all students randomly assigned were actually provided the accommodation because of limited space and equipment. Similarly, in the study by Abedi, Courtney, et al. (2003b), only Spanish speakers were randomly assigned to a bilingual dictionary condition, although the control group included students with native languages other than Spanish. The findings reported below were largely robust to the inclusion or exclusion of these three studies.

All but 1 of the 11 studies used multiple samples to investigate different accommodations and/or a single accommodation provided in multiple grades.[1] Thus, together the studies yielded 38 different tests of the effectiveness of specific accommodations for ELLs as well as 30 tests of the validity of accommodations. Of the 38 tests of effectiveness, 34 involved students in fourth grade ($n = 11$) or eighth ($n = 23$) grade, whereas 4 involved students in fifth grade ($n = 2$) or sixth grade ($n = 2$). Of the 38 tests of effectiveness, 17 used a math test as the outcome measure, 20 used a science test, and 1 used a reading test. Of these effects, 29 used the National Assessment of Educational Progress (NAEP) assessment or NAEP items ($n = 23$) or items drawn from the NAEP and Trends in International Mathematics and Science Study assessments ($n = 6$). Only 9 effects were based on a state accountability assessment (8 of which came from two studies using the Delaware State Test and 1 of which came from a study using the Minnesota state test). Of the 11 studies, 8 reported that students were classified as ELLs based on school records of a "limited English proficient" or "ELL" designation, whereas ELL classification was not reported in the remaining studies. Although this suggests consistency in ELL classification across studies, it is important to note that the criteria for such school-based designations can vary considerably across states and districts (Ragan & Lesaux, 2006). Appendix A provides detailed information on the design of each study and the characteristics of the participants.

In their review of state assessment policies regarding ELLs, Rivera and colleagues (2006) identified 75 accommodations that are currently made available to ELLs. Of these, they found roughly 37 that are considered potentially appropriate insofar as they are specially designed to address the linguistic needs of ELLs. In contrast to this breadth of accommodations offered to ELLs by states, the 11 studies and 38 tests of the effectiveness of specific accommodations focused on only seven different types of accommodation: simplified English ($n = 16$), English dictionary or glossary ($n = 11$), bilingual dictionary or glossary ($n = 5$), extra time ($n = 2$), Spanish language test ($n = 2$), dual language questions ($n = 1$), and dual language booklet ($n = 1$). In addition to the two effects that included extra time alone, seven estimated effects came from studies that involved extra time bundled with one of three other accommodations: simplified English ($n = 2$), English dictionary ($n = 3$), or bilingual dictionary ($n = 2$). One study (Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005) allowed extra time to participants in both the control and treatment conditions; this study was not coded as evaluating the effect of extra time. All but two of the reported effect size estimates are based on paper and pencil tests; the remaining two used computerized assessments.

Because technical reports were included in addition to published articles, there is little reason to believe that publication bias would have led to the inflation of effect sizes. Nonetheless, to investigate the possibility that the results of studies with nonsignificant results were more likely to go unreported than those with significant

results, we plotted the standard error of Hedges's g^u statistic against the value of the Hedges's g^u statistic for each study. Inspection of this plot revealed the funnel shape we would expect in the absence of substantial publication bias, with samples with more precise estimates yielding effect sizes closer to the mean and little evidence of a gap in which unreported nonsignificant effect sizes would occur.

Accommodations That Have Been Evaluated Empirically

As mentioned, in the studies reviewed, seven different types of accommodations were evaluated: simplified English, English dictionaries or glossaries, bilingual dictionaries or glossaries, tests in the native language, dual language test booklets, dual language questions for English passages, and extra time. Each of these is theoretically justifiable for ELLs insofar as they are designed to address the language needs of the ELLs by minimizing variation in scores because of construct-irrelevant language abilities. With the single exception of dual language questions, the accommodations were studied exclusively with tests of math and science.

Simplified English involves changes in the vocabulary and grammar of test items to eliminate irrelevant linguistic complexity while maintaining the same content vocabulary and level of complexity in the content task. These changes include eliminating rare vocabulary unrelated to the content, shortening or simplifying sentence structure, replacing passive voice with active voice, and replacing complex verb forms with present tense verbs (for a description, see Abedi et al., 1997). *English dictionaries or glossaries* involve providing definitional information in English in some form, including standard dictionaries, dictionaries customized to the assessment, or glossaries for specific words used in the assessment. Here again, the intent is to provide definitional information about words that are necessary to comprehend the task but do not represent key concepts of the content. Similarly, *bilingual dictionary, glossary, or marginal glosses* provide bilingual students with access to definitions or direct translations of selected noncontent words in students' native language. Another variant on this accommodation involves providing marginal glosses—explanatory notes written in the margin of the text in the students' native language.

Three other accommodations involve the use of native language in the test itself. *Native language versions* of tests involve adapting tests into the native language of students. The most common method of adapting a test to another language is to use back translation; the test is translated from the original language into the native language by a biliterate test maker. This adapted test is then translated back into the original language by an independent individual, and the two original language tests are compared for equivalence. This process not only is resource intensive but also can introduce additional threats to validity because of the difficulty in maintaining equivalence in the construct measured (American Institutes of Research, 1999). Dual language assessments involve test booklets, in which English versions and native language versions of the same item are placed on facing pages. Two types of dual language tests have been investigated—*dual language booklets* in which all items on math test are presented in two languages and *dual language questions* in which a reading passage is presented in English, followed by questions read aloud in two languages.

Finally, one of the most frequently used accommodations for ELLs is to provide *extra time* to complete the test. The theoretical rationale is that ELLs will be able to demonstrate their content knowledge and skills better if given additional time to

work through the language demands of the test. Often, extra time is provided in combination with another type of accommodation, in which case the rationale is to allow students the time required to use the accommodation (e.g., to use a dictionary to look up the meanings of unknown words).

Methods for Meta-Analysis

To evaluate the appropriateness and practical importance of test accommodations for ELLs, three sets of meta-analyses were conducted. First, a preliminary meta-analysis was conducted to compare the academic achievement test scores of ELLs in the absence of accommodations with those of native English speakers. This first analysis was undertaken in an effort to describe the magnitude of differences in test scores between ELLs and non-ELLs in the absence of accommodations. It is this set of differences that the accommodations are intended to help ameliorate, and thus it serves as a metric for judging the magnitude of the effect sizes for accommodations. The second analysis addressed the effectiveness of accommodations by estimating the degree to which each accommodation led to improved performance for ELLs. The third analysis addressed validity of the accommodations by estimating the impact of the accommodations on the performance of non-ELLs, with the assumption that a valid accommodation should have no significant effect on their performance. To compute average effect sizes, we treated each study sample as the unit of analysis, yielding 38 tests of effectiveness. We made this decision because effects of different accommodations that were derived from the same study were based on different samples of students. Although effect sizes derived from the same study cannot generally be considered independent, in the present case multiple effects from the same study were not generally involved in evaluating the effects of any particular accommodation. That is, studies contributed multiple effects across the set of accommodations but did not typically contribute multiple effect sizes for any single accommodation. Insofar as the net effect of this nonindependence is to reduce the standard error of the mean effect size, it will be seen that any failure of this strategy to fully address the issue of nonindependence would not alter the general conclusions from the analyses of mean effect sizes.

To compute average effect sizes across the entire set of samples and for all samples addressing specific accommodations, we averaged across different outcomes and grades.[2] In averaging the different effect sizes, we weighted the individual effect sizes according to their precision. As our measure of effect size, we first computed the mean difference in performance between ELLs receiving the accommodated test and ELLs taking the test without accommodations. (For analyses of validity, this difference was computed for non-ELLs taking the accommodated test with and those taking the test without accommodations.) This difference in mean performance was then standardized using the pooled within-groups estimate of the standard deviation. This measure of effect size is the common Cohen's d, which is known to be biased in small samples. We therefore corrected this measure of effect size using a transformation of d recommended by Hedges (1981) to produce estimates in Hedges's g^u. These estimates were computed directly from the means and standard deviations reported in the studies by using a programmed routine in the Comprehensive Meta-Analysis (Version 2) software (Borenstein, Hedges, Higgins, & Rothstein, 2005).

In addition to estimating the mean effect size for each accommodation, we investigated whether other aspects of the accommodation treatment moderated the effect of the accommodations; these moderators included the grade level of the students, the domain tested (math, science, or reading), whether the test was based on the NAEP or a state test, and whether the accommodation was bundled with extra time or provided alone. Using PROC MIXED in SAS (SAS Institute, 1999), a two-level hierarchical linear model (HLM) was fitted, in which Level 1 equations represented the level of the effect size for each observation and Level 2 equations represented the study level, where study characteristics (including type of accommodation as well as moderating factors) that served to explain variation in effect sizes were included (Raudenbush & Bryk, 2002). We first fitted an unconditional model, in which random effects variance at Level 1 was specified to be the variance due to sampling error within sample (which was assumed known and given by the square of the standard error of the Hedges's g^u statistic from each effect size estimate) and Level 2 variance was specified to be the variance in Hedges's g^u statistics attributable to differences between samples. Next, we fitted a set of conditional models in which dummy variables for the type of accommodation and other potential moderator variables were included at Level 2 to determine if they explained variation in the effect sizes between samples. To determine if a given variable explained statistically significant variation in effect sizes, we examined the change in goodness of fit between models using the change in –2 log likelihood statistic (Δ-2LL) and conducted a significance test by comparing this statistic to a chi-square distribution with 1 degree of freedom. In addition to investigating moderator effects because of type of accommodation, analyses were conducted to determine if the effects for specific accommodations differed as a function of a characteristic of the studies themselves (e.g., whether the study employed an experimental or quasi-experimental design, the grade level of the students, content domain measured).

Results

Preliminary Analyses: Differences in Achievement Test Scores Between ELLs and Native English Speakers

Before addressing the question of effectiveness of accommodations, we estimated the average difference in academic achievement test scores between ELLs and native English speakers that can be expected on large-scale assessments. These estimates provide a context for evaluating the practical importance of the effects of accommodations. Table 1 presents several estimates of the math and science achievement gaps between ELLs and native English speakers. The top half of Table 1 presents mean effect sizes (reported as Hedges's g^u statistics) for the differences in math and science achievement scores between ELLs and native English speakers in the unaccommodated conditions from the studies reviewed. These estimates suggest that there are large achievement score differences between the two groups across these grades and domains of knowledge, with mean effect sizes ranging from six tenths to three fourths of a standard deviation. They also suggest that the achievement gap differs by test domain to some extent, with larger gaps present in science than in math.

Although these differences between ELLs and non-ELLs are quite substantial, they are somewhat small in comparison to estimates of the achievement gap from

TABLE 1

Estimates of the achievement score differences between English language learners and native English speakers in math and science from studies reviewed (as Hedges's g[a]) and from the 2005 National Assessment of Educational Progress (as Cohen's d; National Center for Education Statistics, 2005)

	Number of studies	Mean effect size	95% confidence interval	
			Lower limit	Upper limit
By domain[a]				
Math	7	0.604	0.279	0.929
Science	11	0.748	0.581	0.914
2005 National Assessment of Educational Progress				
4th grade math		0.831	0.799	0.864
8th grade math		1.006	0.964	1.047
4th grade science		1.051	1.008	1.094
8th grade science		1.227	1.177	1.277

a. The achievement score difference in reading was not estimated because only a single study examined this domain.

national studies. For example, as another point of reference, the bottom half of Table 1 presents estimates of the achievement difference between native English speakers and ELLs from the 2005 NAEP.[3] These estimates are expressed appropriately as Cohen's *d* because of the large sample on which the estimates are based. These estimates are appreciably larger than those from the studies reviewed, with three of the four differences greater than one standard deviation. As with the studies reviewed, the gap was larger for science than for math and for eighth grade students compared to fourth grade students. The difference in magnitude between the NAEP estimates and those from the studies reviewed may be because of the confounding of concomitant predictors of achievement, such as poverty, in the national samples, which likely are better controlled by the design of the research studies of accommodations. All of the studies reviewed sampled ELL and native English-speaking students from within the same schools and/or districts, whereas the NAEP estimates are based on a nationally representative sample. The NAEP estimates may thus capture more of the variation due to differences between the schools attended by ELLs and those attended by native English speakers as well as those concomitant demographic characteristics that tend to affect achievement of at-risk populations in national samples but whose effects are masked when results are disaggregated on only a single dimension. Nevertheless, both sets of estimates indicate that there are large observed differences in achievement in both math and science between ELLs and native English speakers on large-scale assessments, suggesting that one metric by which we can judge the effectiveness of accommodations is the extent to which they reduce these apparent achievement gaps.

TABLE 2a

Average effect sizes for accommodations for English language learners used in 11 experimental and quasi-experimental studies

		Results for fixed effects analysis								
		Effect size and 95% confidence interval				Test of mean effect = 0		Test of heterogeneity in effect sizes		
Accommodation	Number of studies	Effect size *M*	*SE*	Lower limit	Upper limit	*Z*	*p*	*Q*	*df(Q)*	*p(Q)*
English dictionary or glossary	11	0.146	0.043	0.063	0.230	3.427	.001	14.804	10	.139
Simplified English	16	0.030	0.043	–0.053	0.114	0.708	.479	23.885	15	.067
Bilingual dictionary or glossary	5	–0.096	0.065	–0.223	0.031	–1.479	.139	13.53	4	.009
Spanish version	2	–0.263	0.102	–0.463	–0.062	–2.572	.010	14.465	1	<.001
Dual language booklet	1	–0.177	0.065	–0.223	0.031	–1.199	.231			
Dual language questions + read aloud in Spanish	1	0.273	0.195	–0.109	0.654	1.401	.161			
Extra time	2	0.209	0.142	–0.069	0.488	1.473	.141	0.155	1	.693
Total within								66.844	31	<.001
Total between								24.426	6	<.001
Overall mean	38	0.038	0.025	–0.012	0.087	1.481	.139	91.270	37	<.001

TABLE 2b

Average effect sizes for accommodations for English language learners used in 11 experimental and quasi-experimental studies (cont.)

		Results for Random Effects Analysis								
		Effect size and 95% confidence interval				Test of mean effect = 0		Test of heterogeneity in effect sizes		
Accommodation	Number of studies	Effect size *M*	*SE*	Lower limit	Upper limit	*Z*	*p*	*Q*	*df(Q)*	*p(Q)*
English dictionary or glossary	11	0.178	0.055	0.070	0.287	3.232	.001			
Simplified English	16	0.037	0.067	–0.093	0.168	0.557	.557			
Bilingual dictionary or glossary	5	–0.039	0.131	–0.295	0.217	–0.298	.766			
Spanish version	2	0.302	0.719	–1.107	1.711	0.420	.674			
Dual language booklet	1	–0.177	0.065	–0.223	0.031	–1.199	.231			
Dual language questions + read aloud in Spanish	1	0.273	0.195	–0.109	0.654	1.401	.161			
Extra time	2	0.209	0.142	–0.069	0.488	1.473	.141			
Total within										
Total between								9.013	6	.173
Overall mean	38	0.102	0.037	0.029	0.174	2.753	.006			

Effectiveness of Accommodations

To address the question of whether accommodations for ELLs are effective in improving the performance of these learners, we estimated an overall weighted mean effect size across all accommodations as well as mean effect sizes for each accommodation studied. Table 2 includes each of these mean effect sizes as well as the related standard error, a 95% confidence interval, and a test of the hypothesis that the average effect size is zero (see Appendix B for individual effect sizes for each test as well as the means, standard deviations, and sample sizes on which these estimates were calculated). The results in Table 2 indicate an overall lack of evidence for the effectiveness of test accommodations for ELLs. The overall mean effect size (mean $g^u = 0.04$, $SE = 0.03$) is not statistically significant ($z = 1.48$, $p < .139$) and is fairly negligible in magnitude relative to the differences between ELLs and non-ELLs in unaccommodated conditions as seen in Table 1. However, it is important to also note that the test statistic for heterogeneity in effect sizes is statistically significant and fairly large ($Q = 91.27$, $df = 37$, $p < .001$), indicating substantial variability in effect sizes. In fact, the ratio of variation between studies to variation within studies indicates that roughly 27% of the total variability in effect sizes is between accommodations (intraclass correlation = 24.43 + [24.43 + 66.84]). These latter results indicate that a focus on the mean effect size from the fixed effects analysis is not justified and that mean effect sizes may differ across accommodations.

Examining the results for individual accommodations, it can be seen that of the seven types of accommodations used, only *one*—English language dictionaries and glossaries—had an overall positive effect on ELLs' outcomes. This finding is robust to whether we include results from the three studies that employed quasi-experimental designs, and this finding is consistent with findings from three of the four studies that employed a repeated measures design. Two accommodations (bilingual dictionaries or glossaries and Spanish language assessments) demonstrated significant variability across the estimates of their effects. This finding indicates that effect sizes were not consistent across studies of these accommodations and may indicate that these accommodations are effective for some but not all ELLs. Although the number of studies that investigated these accommodations precludes a detailed investigation of the sources of variability in these effects, it is conceivable that effects of these accommodations depend on moderating factors, such as characteristics of the students and their instructional context. Below, we provide detailed findings for each accommodation studied and evidence, or lack thereof, for the existence of moderating factors.

English dictionaries and glossaries. Based on 11 effects, providing customized English language dictionaries or glossaries was the only accommodation found to have a statistically significant and positive average effect size, albeit a small one (mean $g^u = 0.15$, $p = .001$ in the fixed effects analysis; mean $g^u = 0.18$, $p = .001$ in the random effects analysis). Of these 11 effects, 7 came from randomized experiments; when the analysis was conducted without the 4 effects from quasi-experiments, the average effect size continued to be statistically significant and was comparable in magnitude to that found including these effects (mean $g^u = 0.12$, $p = .021$ in the fixed effect analysis). Moreover, the test for heterogeneity in effect sizes indicated that the effect sizes were consistent across the set of 11 effects ($Q = 14.80$, $df = 10$, $p = .139$).

Nonetheless, because the studies involving this accommodation varied along several interesting and potentially important dimensions, we fitted a set of HLM models to examine the effects of moderator variables on this effect. We found little evidence that any of the following moderator variables explained significant variation in effect sizes: providing extra time along with the dictionary (Δ-2LL = 0.40, df = 1, p = .527), providing the test and dictionary in a computerized format (Δ-2LL = 0.30, df = 1, p = .584), the grade level of the students (Δ-2LL = 0.40, df = 1, p = .527), or the domain of the test (Δ-2LL = 0.001, df = 1, p = .975). Results from the one repeated measures study investigating English dictionaries (Albus, Bielinski, Thurlow, & Liu, 2001) found that there was no significant effect for providing a standard dictionary to Hmong-speaking middle school students.

One way of evaluating the practical impact of the small effect of this accommodation on ELLs' performance is to consider it in relation to the estimates of the achievement gap between ELLs and native English speakers in Table 1. Liberal estimates of the reductions in the achievement gaps one can expect from providing English dictionaries and glossaries would be approximately 24% in science and 20% in math (based on the larger effect size estimate taken from both experimental and quasi-experimental studies and the smaller estimates of the achievement gap in Table 1). On the other hand, conservative estimates would be an 11% reduction in the gap for eighth grade math and a 9% reduction for eighth grade science (based on the smaller effect size estimate and the larger estimates of the gaps reported on the NAEP). Taken together, these estimates suggest that a modest improvement can be expected from the most effective accommodation, in the range of a 10% to 25% reduction in the achievement gap between these learners and native English speakers.

Simplified English. Based on 16 effects, simplified English was not found to have a statistically significant effect (g^u = 0.03, p = .479). Moreover, the test for heterogeneity suggests that effect sizes were consistent across the collection of effects examining this accommodation (Q = 23.89, df = 15, p = .067). Of these 16 effects, 12 were based on randomized controlled trials; 4 of the effects were based on Brown (1999), in which it is not clear whether random assignment was used. When the analysis was conducted excluding the 4 effects from this presumably quasi-experimental study, the point estimate for the average effect size was identical to that found including all 16 effects and continued to be nonsignificant (g^u = 0.03, p = .495). In addition, this effect was not found to be moderated by the grade level of students (Δ-2LL = 1.40, df = 1, p = .237), domain of the test (math compared to science and reading [Δ-2LL = 0.04, df = 1, p = .842]), or whether the test was based on the NAEP or a state test (Δ-2LL = 0.80, df = 1, p = .371).

In addition to the between-group studies investigating simplified English, two repeated measures studies were conducted using this accommodation. In one of these studies (Abedi & Lord, 2001), ELLs scored higher when taking a test composed of simplified English items than when taking a test composed of standard items. Although statistically significant, the difference between students' performance on simplified and original items was a small fraction of an item (.17) difference on a 10-item test, yielding a small effect size (g^u = 0.09). In the other study (Johnson & Monroe, 2004), the difference between ELLs' performance on simplified English items and their performance on standard items for ELLs was negligible.

Bilingual dictionaries and glossaries. In contrast to providing English dictionaries, the use of bilingual dictionaries or glossaries did not show a positive effect ($g^u = -0.04$, $p = .766$). However, despite being based on just five estimates of effect size drawn from three studies, the test for heterogeneity indicated that effect sizes were not consistent across the collection of effect size estimates ($Q = 13.53$, $df = 4$, $p = .009$). All five effects in this collection involved fourth or eighth grade science assessments, so this heterogeneity is not because of differences in the domain tested. Further analysis indicated that the effect did not differ as a function of grade level (Δ-2LL = 0.10, $df = 1$, $p = .752$). It is worth noting that the two largest effects were of opposite sign, and both came from studies with fourth grade ELLs.

Some portion of the heterogeneity in effect sizes found for bilingual dictionaries and glossaries may be because of differences in the designs of the three studies. Two of the negative effects, one of which was significantly negative, come from a quasi-experimental study (Abedi, Courtney, et al., 2003b), in which the bilingual dictionary treatment condition included only Spanish-speaking students but the comparison group included ELLs from other language backgrounds as well as Spanish speakers. Thus, it is conceivable that this negative effect may be the result of preexisting differences in achievement between the accommodated and unaccommodated groups associated with the differential distributions of language backgrounds in the two groups. When the analysis was conducted with only the three effects from true experiments, the average effect size continued to be nonsignificant, although it had a positive direction ($g^u = 0.17$, $p = .140$) and there was no longer evidence of heterogeneity among the three effect sizes ($Q = 4.87$, $df = 2$, $p = .088$). Conducting an analysis using HLM, we found that study design (quasi-experimental or experimental) did significantly moderate the effect of this accommodation (Δ-2LL = 5.10, $df = 1$, $p = .024$) such that the effect was predicted to be positive in experimental studies and negative in quasi-experimental studies. Of course, there are other characteristics of the single quasi-experimental study (e.g., characteristics of the participants, application of the treatment, and specific test items accommodated) that may also explain this association between study design and effect size.

Spanish versions of assessments. The results in the top half of Table 2 show that ELLs scored lower when provided Spanish language assessments as an accommodation than when given the original English assessment. However, the test of homogeneity of effect sizes also shows that effect sizes were not consistent across the two samples, and as a result the fixed effect mean in the top half of Table 2 should be ignored in favor of the random effects mean reported in the bottom half of Table 2. This mean is a positive 0.30 but is not statistically significantly different from zero. Both effect sizes come from the same study (Hofstetter, 2003) but from two different samples of students. One set of students consisted of Hispanic students instructed in Spanish, whereas the second set of students consisted of Hispanic students instructed in English. It is not surprising that a positive effect size for Spanish language accommodation occurred for students who had received instruction in Spanish, whereas a negative effect size occurred for students who had been instructed in English. Although it is difficult to draw inferences from just two tests originating in a single study, these effect sizes suggest that the effect of providing a native-language version of an assessment may be substantially moderated by the language or languages in which students are being taught.

Dual language assessments. Two studies evaluated the effects of providing two different types of dual language accommodations—a dual language booklet and dual language questions that are read aloud. In both studies, the effects were not statistically different from zero, but they were opposite in sign, just as with Spanish language tests. The negative effect originates from a study in which the instruction, as reported by students, was primarily in English (Garcia Duncan et al., 2005), whereas the positive effect originates from a study in which the language of instruction for participants is not reported (M. Anderson et al., 2000). These two studies also examined slightly different accommodations and used tests of different domains (reading and math), making it impossible to determine the impacts of study features or sample characteristics on outcomes.

Extra time. Two studies looked exclusively at extra time, whereas a handful of studies bundled extra time with other modifications, specifically bilingual dictionaries and glossaries ($n = 2$), English dictionaries and glossaries ($n = 3$), and simplified English ($n = 2$). In the two studies that looked exclusively at extra time, the effect was positive but not statistically significant ($g^u = 0.21$, $p = .141$). Given the relatively limited information two studies provides analytically, we also estimated the effect of extra time whether provided alone *or* in conjunction with another accommodation in nine samples using HLM in an effort to bolster this analysis. Providing extra time was found to have a nonsignificant negative effect on the overall mean effect size (effect on mean $g^u = -0.08$, $SE = 0.09$) and thus did not explain significant variation in the effect sizes across the 38 effects (pseudo-$R^2 = .05$, Δ-2LL = 0.70, $df = 1$, $p = .403$).

Extra time was also investigated in one study using a repeated measures design (Castellon-Wellington, 2000). In this study, all students took two parallel versions of a seventh grade science test, first in a standard condition and then in an accommodated condition. Students were asked to choose between extra time and having the questions read aloud to them; a third of the students were given their preference, a third were given their second choice, and a third received a randomly assigned accommodation. Extra time did not yield significantly improved test scores for any of the students, whether they chose the accommodation or not ($p = .603$). This result corroborates the findings above.

Heterogeneity in effect sizes. As shown in Table 2, the tests of heterogeneity across the collection of studies show that the effect sizes varied both within ($Q = 66.84$, $df = 31$, $p < .001$) and between accommodations ($Q = 24.43$, $df = 6$, $p < .001$). These results indicate that there is statistically significant variability in effect sizes across the collection of studies, but that much of this variability ($24.43 \div 91.27 = 26.77\%$) is because of differences in average effect sizes between the seven different types of accommodation. Fitting a set of HLM models yielded results consistent with those above; statistically significant variation in effect sizes was found across the collection of studies (Variance in $g^u = 0.03$, $p = .017$), but differences in the types of accommodations were found to explain a majority of this variation (pseudo-$R^2 = .64$). Moreover, once type of accommodation was taken into account in the HLM model, there was no significant residual variance at the level of the study left to be explained (residual variance in $g^u = 0.01$, $p = .124$). These analyses further indicated that when the type of accommodation was not taken into account, variation

across effect sizes was not significantly explained by grade level (pseudo-$R^2 = .02$, Δ-2LL = 0.30, $df = 1$, $p = .584$), domain tested (pseudo-$R^2 = .04$, Δ-2LL = 0.50, $df = 1$, $p = .480$), whether the study employed items from the NAEP or items from a state assessment (pseudo-$R^2 < .01$, Δ-2LL = 0.20, $df = 1$, $p = .655$), or whether the study employed an experimental or quasi-experimental design (pseudo-$R^2 < .01$, Δ-2LL = 0.001, $df = 1$, $p = .984$). These findings suggest that there is little evidence that these factors significantly moderated the effect of providing accommodations.

Validity of Accommodations

In addition to examining evidence for the effectiveness of accommodations in improving ELLs' performance, we examined the validity of these accommodations. A threat to validity would exist if accommodations improve the performance of those students who do not require them, native English speakers in this case. In sum, we found that there is little cause for concern that providing these accommodations will threaten the validity of inferences based on the resulting test scores.

Of the 11 studies included in the meta-analysis, 9 contributed at least one sample that addressed the validity question, yielding a total of 30 effect sizes.[4] Together, these 30 effect sizes allowed for tests of the validity of five accommodations: bilingual dictionaries and glossaries, Spanish versions of the assessment, extra time, simplified English, and English dictionaries and glossaries. The overall average effect size for all accommodations for native English speakers was not statistically significant ($g^u = -0.003$, $p = .828$). When analyzed by type, none of the accommodations had a significant effect, with the exception of providing a Spanish-language version test that had a significant negative effect in a single sample ($g^u = -0.87$, $p < .001$); this latter finding is not at all surprising given that most native English speakers could not be expected to perform well on a test in Spanish.

The test of heterogeneity indicated that there was some heterogeneity in effect sizes ($Q = 45.59$, $df = 29$, $p = .026$); however, this heterogeneity was likely because of differences between types of accommodations given that there was no significant heterogeneity within types of accommodations ($Q = 30.08$, $df = 25$, $p = .221$). The heterogeneity in effects may be the result of the significant negative effect of the Spanish version accommodation and to a lesser degree to the nonsignificant negative effect of the bilingual dictionaries ($g^u = -0.12$, $p = .096$); the three English accommodations had effects that were all much closer to zero. Thus, there is little reason to believe that providing any of these five accommodations to ELLs would give them an unfair advantage over non-ELLs and thereby threaten the validity of inferences based on the resulting scores. In particular, there is relatively robust evidence that the two accommodations studied most often—providing English dictionaries and simplified English on assessments of math and science—allow for valid inferences about student performance.

Two of the three studies that investigated the effect of accommodations on native English speakers using a repeated measures design corroborated this finding (Albus et al., 2005; Johnson & Monroe, 2004). The single exception is Abedi and Lord (2001), in which providing a version of the test in simplified English had a significant impact on the scores of native English speakers ($g^u = 0.08$). Nevertheless, given that the average weighted effect size for simplified English among native English speakers across the 13 samples from the

between-groups studies was not significantly different from zero ($g^u = 0.003$, $p = .882$), we conclude that providing this accommodation does not pose a threat to the validity of inferences about student performance.

Discussion

The growing number of ELLs in today's schools, combined with the increasing use of large-scale assessments to monitor their academic progress, raises questions about the valid assessment of their academic skills. One method frequently recommended to minimize the influence of variation in ELLs' limited English proficiency on content assessments is to provide them with appropriate test accommodations. Specifically, appropriate accommodations are those designed to provide direct or indirect linguistic support to minimize the impact of irrelevant language demands on ELLs' performance so that they can demonstrate their content knowledge and academic skills to the greatest extent possible.

The present study was designed to provide a quantitative synthesis of experimental and quasi-experimental research on the effectiveness and the validity of those accommodations for ELLs taking large-scale assessments. Using meta-analysis, the study was designed to build on the narrative reviews conducted by Abedi et al. (2004) and Sireci et al. (2003); we synthesized the results reviewed in each of these reports as well as the findings of six additional studies published between 2001 and 2006. Together, 38 samples allowed for tests of the effectiveness of seven accommodations: bilingual dictionaries and glossaries, Spanish versions of the assessment, extra time, simplified English, English dictionaries and glossaries, and dual language booklets and test items. Of these, only two were investigated in a relatively large number of samples: simplified English ($n = 16$) and English dictionaries and glossaries ($n = 11$); thus, we have more confidence in some findings concerning these two accommodations than we do concerning the others.

The findings from the meta-analysis indicated that English dictionaries and glossaries had a statistically significant—if small—impact on the performance of ELLs and that providing tests in simplified English had a negligible impact. Of the remaining five accommodations that have been studied to date, there is limited evidence that any of them are effective in improving the performance of a large and diverse group of ELLs, whether in English or in the native language (i.e., bilingual dictionaries, primary language assessments). It is important to note also that none of the accommodations studied were found to affect the performance of native English speakers, thus suggesting little reason to doubt the validity of providing these accommodations.

The findings also demonstrate little evidence of moderating effects of any of the reported characteristics of students or tests, albeit with a relatively small collection of samples. Instead, the findings from the meta-analysis suggest that any systematic differences across studies were largely because of differences in the type of accommodation used. It is also likely that the effects of accommodations vary as a function of factors that were not observed or reported in these studies. These include student characteristics (e.g., English language proficiency and primary language literacy), instructional factors (e.g., language of instruction and use of dictionaries in classroom activities), characteristics of the test, and characteristics of the study itself.

Our findings converge with and diverge from previous reviews and research on this topic in important ways. They converge with previous reviews in that they

highlight the promise of providing English dictionaries or glossaries for ELLs taking large-scale assessments. In addition, as in the narrative reviews, we identified some heterogeneity in the effects of some accommodations, most notably those involving native language support. Specifically, our findings suggest that ELLs who have received native language instruction in a content area perform better on a native language version of a test in that content area than on an English version, whereas ELLs who have received content area instruction only in English perform worse on a native language version than they would have on an English test. Thus, we agree with Abedi et al. (2004) in recommending that the language of a content assessment (or accommodations for a content assessment) must match the language in which students are receiving instruction in that domain. That said, whether this effect is because of the language of instruction for targeted content area per se or because of the literacy development of children in their native language cannot be determined from this collection of studies, and thus we would add that native language literacy development must also be taken into account.

In contrast to the conclusions of the narrative reviews (i.e., Abedi et al., 2004; Sireci et al., 2003), our meta-analytic findings suggest there is little reason to be optimistic about the potential effectiveness of simplified English as a test accommodation. Simplified English is an accommodation that involves changes in the vocabulary and grammar of test items to eliminate irrelevant linguistic complexity. Although in theory this may appear to be the most compelling accommodation for ELLs to reduce the potentially negative effects of unnecessary complexity in the language used in math and science tests, the results from this meta-analysis suggest that it is not effective in doing so.

Policy makers, researchers, and educators alike may be surprised to find that test accommodations that are specially designed to address the linguistic needs of ELLs directly have limited effectiveness in improving their performance. Such limited effectiveness is particularly intriguing given the evidence from correlational studies that score differences between native English speakers and ELLs are associated with the language demands of the assessment (e.g., Abedi et al., 1997; Abedi et al., 2000; Abedi, Leon, et al., 2003). Further research is needed to uncover the reasons behind these divergent findings; nonetheless, we assert that the most convincing current hypotheses focus on the complex relationship between language and content knowledge.

Academic Language for Academic Success

We hypothesize that the key to explaining the divergence between findings from descriptive studies using correlational designs and our findings lies in the academic language demands of the assessments and the content area itself. Three sources of variation in test scores that have received attention in the literature include "language-free" content knowledge and skills, necessary academic language skills that are essential to performing content tasks, and unnecessary language skills that are irrelevant to performing these tasks. Although psychometricians have sought to isolate these three sources of variation from one another, we argue that they overlap much more substantially than is commonly acknowledged. In describing the relationship between language proficiency and performance on content tests, many researchers, including the authors of the Standards for Educational

and Psychological Testing (AERA, APA, & NCME, 1999), have focused on unnecessary language skills as a threat to validity. All but one study in this meta-analysis examined the use of accommodations with assessments of math and science, two content areas sometimes thought to be "universal languages" characterized by symbols and numbers. If the core of math and science achievement is composed of "language-free" skills and knowledge, then accommodations to reduce the language demands of these assessments should, in theory, be effective in promoting ELLs' achievement.

We hypothesize that the differences in performance between native English speakers and ELLs may be more because of variation in the *necessary* academic language skills required by content area assessments than the *unnecessary* language skills typically highlighted by researchers. Because valid accommodations for ELLs are designed to address unnecessary language demands, they may not compensate for the limited *academic* English skills of many ELLs. This suggestion has important implications for educators seeking to improve students' test performance, particularly given that academic language and vocabulary have typically received minimal instructional attention in kindergarten through 12th grade classrooms across the nation (Durkin, 1978-1979; Roser & Juel, 1982; Scott, Jamieson-Noel, & Asselin, 2003; Watts, 1995).

A second, complementary hypothesis is that the learning of content knowledge and skills is largely mediated by language. We argue that academic language skills and content knowledge overlap with one another to a great degree, such that there are very few "language-free" content skills—virtually all sophisticated academic tasks, such as solving complex mathematical problems or reasoning with scientific information, are mediated by language and literacy skills. Although many educators are not surprised by the achievement differences in reading between ELLs and their native-English speaking peers, less attention has been paid to the similarly large gaps on assessments of math and science. As shown in Table 1, both national assessments and the studies reviewed herein indicate large and troubling achievement score differences between ELLs and native English speakers in both science and math.

Learning to perform complex tasks in math and science relies heavily on academic language skills. For example, several studies on the development of mathematical skills support the central role of academic language in mediating the learning of math (e.g., Cazden, 1986; Cuevas, 1984; Lager, 2006). The mastery of math concepts presupposes facility with the language used to characterize, express, and apply concepts. Yet in math classrooms across the United States, many ELLs struggle to understand much of the language that is used in those classrooms and in the curricular materials, and most learners are not explicitly taught to read, write, or speak mathematically (Lager, 2006). As a result, ELLs who have not had the opportunity to develop these specialized academic language skills have limited access to learning math skills and concepts; thus, they will perform poorly on assessments of mathematics, regardless of any accommodations that might be used to eliminate irrelevant language demands.

In addition to the linguistic challenges that ELLs face in learning content, ELLs also often have more limited opportunities to learn sophisticated academic content compared to their native English-speaking counterparts. Gándara,

Rumberger, Maxwell-Jolly, and Callahan (2003) argued that a substantial part of the achievement gap between ELLs and native English speakers is because of structural differences in the educational resources available to the two populations of learners. For instance, in California, Gándara et al. found evidence for differential opportunities to learn on several fronts, including access to qualified and experienced teachers, access to appropriate and challenging curriculum, and access to native English-speaking peers who can serve as models. If students do not have opportunities to learn and/or are unable to access available instruction, then accommodating the language demands of the assessment will not lead to improved performance because students fundamentally lack the language-based conceptual knowledge of the content, not simply the language skills needed to demonstrate that knowledge.

Implications

Given the increasing importance of large-scale assessments and the increasingly high stakes attached to test results for schools and students, this synthesis has valuable implications for researchers, policy makers, and educators. First, there is a need for further research to address unanswered questions in this area. Our findings primarily focus on two accommodations, simplified English and dictionary use, for which there was robust evidence and focus less so on other accommodations (e.g., extra time, native language accommodations) with fewer studies. For meta-analysis in this area, more data on these accommodations are needed. Given the wide variety of accommodations in use, future research should investigate other innovative and/or widely used methods for accommodating ELLs that have not yet been studied. For instance, a recent descriptive study found that ELLs performed better on math items that included schematic representations of problems than they did on equally difficult items that included only text (Martiniello, 2007), but no study to date has investigated the impact of including such schematic representations as a test accommodation.

Given the wide variety in how accommodations are used in practice, future studies should also investigate the factors that moderate the effect of particular accommodations. Such investigations should investigate accommodations' effectiveness for students with different levels of English and native language proficiency as well as for students from different instructional contexts. In particular, studies investigating native language accommodations could employ designs similar to that used by Hofstetter (2003) in which the effect of accommodation for students receiving content instruction in English is compared to that for students receiving instruction in their native language. These studies could also investigate the effectiveness of accommodations for ELLs from language backgrounds other than Spanish, given that the majority of studies to date have been conducted with Spanish-speaking ELLs. Moreover, given the importance of matching accommodations to students' individual needs, studies should investigate mechanisms for matching ELLs with different needs to different accommodations. For instance, in a very recent study, Kopriva, Emick, Hipolito-Delgado, and Cameron (2007) found that students who received individualized accommodations as recommended by a computerized taxonomy (based on their English language proficiency, English reading proficiency, and native language reading

proficiency) had significantly higher scores than did those who received no accommodations or nonrecommended accommodations.[5]

In addition to further research on the effectiveness of accommodations, research is needed to examine the role of necessary academic language in large-scale assessments. Our null results led us to put forth the hypothesis that content knowledge and academic language may be inextricably connected, such that necessary academic language skills play a greater role in observed achievement differences between ELLs and non-ELLs than do irrelevant language skills. However, the current studies do not allow us to evaluate this hypothesis. A program of research is needed to operationalize the notion of academic language, to investigate the role of such language in the learning and assessment of content knowledge, to describe the academic language skills of both ELLs and native English speakers, and to examine how instruction in academic language might affect students' performance on content area assessments.

Conclusion

In the future, new research may very well uncover accommodations that are more effective than those studied here, and we certainly hope that researchers and test makers continue to think deeply about the challenges of incorporating ELLs appropriately into large-scale assessment. Yet it is important to reiterate that the empirical research to date indicates that, although valid, accommodations are largely ineffective in improving the performance of the majority of ELLs on large-scale assessments. As such, there are clear implications for policy makers and educators: Accommodations are not a solution to the larger issues of promoting the academic skills of ELLs. The single accommodation with clear evidence of effectiveness—providing English dictionaries or glossaries—can be expected to result in a 10% to 25% reduction in the performance gap between ELLs and native English speakers. Although not an inconsequential improvement, this reduction leaves wide gaps in performance that must be addressed through improving instruction for these learners, especially in the content areas. Although some policy makers and educators may look to native language accommodations to improve the assessment of ELLs taking large-scale assessments, these accommodations will be effective only for students who have received instruction in the native language in the content area being assessed. Because the majority of ELLs in the United States receive instruction in English and do not have the opportunity to learn math or science in their native language at school (Crawford, 2004), such accommodations will not provide a solution to the national disparities in achievement between these learners and native English speakers.

We argue that the poor performance of many ELLs on large-scale assessments is largely because of their limited control of academic English—those academic language skills that are not irrelevant to content knowledge but rather central to performing the sophisticated tasks that serve as the goals of math, science, social studies, and language arts instruction. The key implication is that educators must not only refine how they assess ELLs but also dramatically improve how they teach these learners. To meet high standards for academic success, these learners require targeted, explicit, and intensive instruction in the complex and specialized language that lies at the heart of each content area.

APPENDIX A

Characteristics of studies included in the meta-analysis on effectiveness of accommodations

Study	Accommodation(s) investigated	Bundled with extra time	Design	Grade(s)	Domain	Test	Native language(s) of ELLs	Language(s) of instruction	Method of identifying ELLs	English proficiency of ELLs[a]
Abedi, Courtney, and Leon (2003a)	English glossary, bilingual glossary, simplified English	Yes	Experimental and quasi-experimental	4, 8	Science	NAEP and TIMSS	Spanish, Chinese, and "other Asian languages"	Not reported	School records	Mean self-reported understanding of oral English halfway between "well" and "very well"
Abedi, Courtney, and Leon (2003b)	English glossary, computerized English glossary, extra time	No	Quasi-experimental	4, 8	Math	NAEP	Spanish, Korean, Chinese, and others	Not reported	School records	75% of 4th graders and 55% of 8th graders self-reported understanding English directions "very well"
Abedi, Courtney, Mirocha, Leon, and Goldberg (2005)	English dictionary, bilingual dictionary, simplified English		Experimental	4, 8	Science	NAEP	Spanish, Korean, Filipino, Chinese, and others	Bilingual and English only English	School records	Mean self-reported understanding of oral English between "well" and "very well"
Abedi, Hofstetter, Baker, and Lord (2001)	English glossary, simplified English		Experimental	8	Math	NAEP	Spanish, Khmer, Vietnamese, Tagalog, Lao, and others		School records	Nearly half self-reported understanding and speaking English "very well"

(continued)

APPENDIX A (continued)

Study	Accommodation(s) investigated	Bundled with extra time	Design	Grade(s)	Domain	Test	Native language(s) of ELLs	Language(s) of instruction	Method of identifying ELLs	English proficiency of ELLs[a]
Abedi, Lord, Boscardin, and Miyoshi (2001)	Customized English dictionary	No	Experimental	8	Science	NAEP	Predominantly Spanish with small number of participants speaking other languages	Not reported	School records	Mean self-reported ability to understand and speak English close to "very well"
Abedi, Lord, and Hofstetter (1998)	Simplified English	No	Experimental	8	Math	NAEP	Khmer, Vietnamese, and others	English only and Spanish–English bilingual	School records	Half self-reported understanding oral English "very well"
M. Anderson, Liu, Swierzbin, Thurlow, and Bielinski (2000)	Dual language questions + read aloud in Spanish	No	Experimental	8	Reading	Minnesota Basic Standards Test	Spanish	Not reported	School records	Not reported
Brown (1999)	Simplified English	No	Ambiguous	5, 8	Math, Science	Delaware State Test	Not reported	Not reported	Not reported	Not reported

(continued)

APPENDIX A (continued)

Study	Accommodation(s) investigated	Bundled with extra time	Design	Grade(s)	Domain	Test	Native language(s) of ELLs	Language(s) of instruction	Method of identifying ELLs	English proficiency of ELLs[a]
Garcia Duncan et al. (2005)	Dual language booklet	No	Experimental	8	Math	NAEP	Spanish	Predominantly English	Not reported	Mean self-reported ratings of 15.15 (on a scale from 4 to 16)
Hofstetter (2003)	Spanish version	No	Experimental	8	Math	NAEP	Spanish	English only, Spanish and English	School records	Mean self-reported ratings between "well" and "very well"
Rivera and Stansfield (2004)	Simplified English	No	Experimental	4,6	Science	Delaware Science Test	Various	Not reported	Not reported	Not reported

Note: ELL = English language learner; NAEP = National Assessment of Educational Progress; TIMSS = Trends in International Mathematics and Science Study.
a. In addition to the self-reported ratings of English proficiency reported here, several of the studies provided data on students' English reading proficiency. However, these scores do not necessarily capture students' oral proficiency; they were also consistently reported as raw scores and thus are difficulty to interpret.

APPENDIX B

Descriptive statistics, Cohen's d, *and Hedges's* g^u *for studies included in meta-analysis on effectiveness of accommodations*

		English language learners—accommodations			English language learners—no accommodations						
Accommodation	Study name	*M*	*SD*	*n*	*M*	*SD*	*n*	Cohen's *d*	*SE*	Hedges's g^u	*SE*
Bilingual dictionary and glossary	Abedi, Courtney and Leon (2003a)	45.62	8.19	135	48.23	9.38	268	–0.290	0.106	–0.289	0.106
	Abedi, Courtney, et al. (2003a)	44.58	8.38	119	45.73	9.41	199	–0.127	0.116	–0.127	0.116
	Abedi, Courtney, Mirocha, Leon, and Goldberg (2005)	11.72	3.73	64	10.04	3.66	62	0.455	0.180	0.452	0.179
	Abedi et al. (2005)	9.38	2.69	16	10.32	3.99	22	–0.268	0.330	–0.262	0.323
	Abedi, Lord, Boscardin, and Miyoshi (2001)	8.51	4.72	70	8.36	4.4	58	0.033	0.178	0.033	0.177
Dual language booklet	Garcia Duncan et al. (2005)	30.65	11.74	74	32.92	13.36	119	–0.178	0.148	–0.177	0.148
Dual language questions + read aloud in Spanish	M. Anderson, Liu, Swierzbin, Thurlow, and Bielinski (2000)	17.70	7.31	53	15.85	6.09	52	0.275	0.196	0.273	0.195
English dictionary and glossary	Abedi, Courtney, and Leon (2003b)	13.81	6.043	64	12.27	5.242	80	0.274	0.168	0.273	0.168
	Abedi, Courtney, et al. (2003a)	9.95	3.835	86	9.41	4.005	86	0.138	0.153	0.137	0.152
	Abedi, Courtney, et al. (2003a)	14.69	5.115	35	12.27	5.242	80	0.465	0.205	0.462	0.204

(continued)

APPENDIX B (continued)

Accommodation	Study name	English language learners—accommodations			English language learners—no accommodations			Cohen's *d*	*SE*	Hedges's g^u	*SE*
		M	*SD*	*n*	*M*	*SD*	*n*				
	Abedi, Courtney, et al. (2003a)	10.17	4.361	84	9.47	4.005	86	0.167	0.154	0.167	0.153
	Abedi, Courtney, et al. (2003a)	48.37	9.75	270	48.23	9.38	268	0.015	0.086	0.015	0.086
	Abedi, Courtney, et al. (2003a)	46.68	9.00	206	45.73	9.41	199	0.103	0.099	0.103	0.099
	Abedi et al. (2005)	11.97	3.47	59	10.04	3.66	62	0.541	0.185	0.537	0.184
	Abedi et al. (2005)	11.52	3.53	23	10.32	3.99	22	0.319	0.300	0.313	0.295
	Abedi, Hofstetter, Baker, and Lord (2001)	11.84	5.94	146	12.07	5.47	144	–0.040	0.117	–0.040	0.117
	Abedi, Hofstetter, et al. (2001)	13.69	6.74	29	12.07	5.47	144	0.284	0.204	0.283	0.203
	Abedi, Lord, et al. (2001)	10.18	5.26	55	8.36	4.4	58	0.376	0.190	0.374	0.189
Extra time	Abedi, Courtney, et al. (2003a)	13.74	6.024	35	12.27	5.242	80	0.268	0.203	0.266	0.202
	Abedi, Hofstetter, et al. (2001)	12.93	5.99	30	12.07	5.47	144	0.155	0.201	0.154	0.200
Simplified English	Abedi, Courtney, et al. (2003a)	47.36	9.48	284	48.23	9.38	268	–0.092	0.085	–0.092	0.085
	Abedi, Courtney, et al. (2003a)	47.63	9.53	209	45.73	9.41	199	0.201	0.099	0.200	0.099
	Abedi et al. (2005)	10.55	3.37	20	10.04	3.66	62	0.142	0.257	0.141	0.255

(continued)

APPENDIX B (continued)

Accommodation	Study name	English language learners—accommodations			English language learners—no accommodations			Cohen's *d*	*SE*	Hedges's g^u	*SE*
		M	*SD*	*n*	*M*	*SD*	*n*				
	Abedi et al. (2005)	13.27	3.04	11	10.32	3.99	22	0.795	0.382	0.776	0.373
	Abedi, Hofstetter, et al. (2001)	12.63	5.23	124	12.07	5.47	144	0.104	0.123	0.104	0.122
	Rivera and Stansfield (2004)	4.33	2.52	18	4.67	1.91	15	–0.150	0.350	–0.146	0.342
	Rivera and Stansfield (2004)	4.38	1.71	16	3.48	1.89	23	0.495	0.330	0.485	0.324
	Rivera and Stansfield (2004)	2.11	1.27	9	4.00	1.50	9	–1.360	0.523	–1.295	0.498
	Rivera and Stansfield (2004)	2.00	1.79	6	3.23	2.45	13	–0.540	0.501	–0.516	0.479
	Hofstetter (2003)	5.50	2.73	6	5.00	3.50	9	0.155	0.528	0.146	0.497
	Hofstetter (2003)	11.49	5.40	222	11.32	4.90	229	0.033	0.094	0.033	0.094
	Abedi, Lord, and Hofstetter (1998)	12.69	6.03	117	13.41	6.14	115	–0.118	0.131	–0.118	0.131
	Brown (1999)	16.50	6.58	16	16	5.02	13	0.084	0.374	0.082	0.363
	Brown (1999)	14.00	3.37	4	10.5	6.36	2	0.811	0.897	0.649	0.718
	Brown (1999)	29.22	12.47	18	33.55	14.47	11	–0.327	0.385	–0.318	0.374
	Brown (1999)	19.25	13.87	4	9.5	2.12	2	0.809	0.897	0.647	0.718
Spanish version	Hofstetter (2003; Spanish instruction)	8.68	3.41	63	5.00	3.5	9	1.076	0.367	1.064	0.364
	Hofstetter (2003; English instruction)	9.64	3.66	147	11.32	4.90	229	–0.377	0.107	–0.376	0.106

Notes

This research was supported in part by funds from the Center on Instruction, which is operated by RMC Research Corporation under Cooperative Agreement S283B050034 with the U.S. Department of Education and in partnership with the Florida Center for Reading Research at Florida State University; RG Research Group; the Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston; and the Vaughn Gross Center for Reading and Language Arts at the University of Texas at Austin. The analyses and opinions expressed herein are those of the authors and do not necessarily represent the policy of the U.S. Department of Education, nor should one assume that they have the endorsement of the federal government. The authors would like to thank Hector Rivera for substantial contributions to this work as well as Jamal Abedi, Charlene Rivera, and Maria Pennock-Roman for their comments on earlier drafts of this article.

[1]Two studies, Abedi, Lord, and Hofstetter (1998) and Hofstetter (2003) involved partially overlapping samples. Hofstetter focused on the Hispanic students who participated in the Abedi et al. study, who composed roughly two thirds of the original study sample. Using information reported in both studies, we were able to compute means, standard deviations, and sample sizes for the non-Hispanic portion of their sample in Abedi et al. so that the statistics reported for these two studies are nonoverlapping. Additional information on how the effect size for non-Hispanic students in Abedi et al. was computed using information from Hofstetter is available from the authors on request.

[2]It is worth noting that in some studies a single control group (i.e., English language learners [ELLs] taking the test without accommodations) was compared to more than one treatment (i.e., accommodated ELL group), rendering some comparisons within a study dependent on one another. Because these different comparisons involving the control group addressed questions about different accommodations in our analysis, this dependence would serve to increase the correlation between findings across different sets of accommodations. We felt that this drawback was worth the added information about the effect of different accommodations gained by using the sample as the unit of analysis.

[3]Of course, the 2005 National Assessment of Educational Progress (NAEP) allowed for the use of some accommodations for some ELLs based on the states' current practices, so these estimates cannot be interpreted directly as unaccommodated scores. However, Abedi and Hejri (2004) found for the 2002 NAEP administration that a very small number of ELLs received accommodations on the NAEP and that there was limited evidence that accommodations raised the achievement of these learners. Thus, these estimates can be used as an approximate baseline against which we might judge the effectiveness of the use of accommodations were they to be implemented on a much more larger scale, say extended to all students classified as limited English proficient assessed on the NAEP.

[4]This analytic sample is smaller than the 38 samples used in the previous meta-analysis because three studies did not provide native English speakers with a native language accommodation and because one study disaggregated ELLs into several samples based on language of instruction but aggregated native English speakers into a single sample.

[5]Although it is a well-designed study that otherwise met the criteria for inclusion, this study was not published in time to be included in the current meta-analysis.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Abedi, J., Courtney, M., & Leon, S. (2003a). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Technical Report 608). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

*Abedi, J., Courtney, M., & Leon, S. (2003b). *Research-supported accommodation for English language learners in NAEP* (CSE Technical Report 586). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

*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 Technical Report 666). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education, 17,* 371–392.

*Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance test accommodations: Interactions with student language background* (CSE Technical Report 536). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Abedi, J., & Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CSE Technical Report 603). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

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

*Abedi, J., Lord, C., Boscardin, C. K., & Miyoshi, J. (2001, September). *The effects of accommodations on the assessment of limited English proficient students in the National Assessment of Educational Progress* (Working Paper 200113). Washington, DC: National Center for Education Statistics.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Technical Report 478). Los Angeles: 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 Measurements: Issues and Practice, 19*(3), 16–26.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Technical Report 429). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Albus, A., Bielinski, J., Thurlow, M., & Liu, K. (2001). *The effect of a simplified English language dictionary on a reading test* (LEP Project Report 1). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Albus, D., Thurlow, M., Liu, K., & Bielinski, J. (2005). Reading test performance of English-language learners using an English dictionary. *The Journal of Educational Research, 98,* 245-256.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Education Research Association.

American Institutes of Research. (1999). *Voluntary national tests in reading and math: Background paper reviewing laws and regulations, current practice, and research relevant to inclusion and accommodations for students with limited English proficiency.* Palo Alto, CA: Author.

*Anderson, M., Liu, K., Swierzbin, B., Thurlow, M., & Bielinski, J. (2000). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2* (Minnesota Report 31). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Anderson, N. E., Jenkins, F. F., & Miller, K. E. (1996). *NAEP inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics.* Washington, DC: Educational Testing Service.

Bailey, A. (2005). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Technical Report 663; pp. 79–100). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Batalova, J., Fix, M., & Murray, J. (2007). *Measures of change: The demography and literacy of adolescent English language learners.* New York: National Center on Immigrant Integration Policy.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2.* Englewood, NJ: Biostat.

*Brown, P. (1999). *Findings of the 1999 Plain Language Field Test* (Publication T99–013.1). Newark: University of Delaware, Delaware Education Research & Development Center.

Butler, F. A., & Castellon-Wellington, M. (2005). Students' concurrent performance on tests of English language proficiency and academic achievement. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Technical Report 663; pp. 47–78). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Capps, R., Fix, M., Murray, J., Ost, J., Passel, J., & Herwantoro, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act.* Washington, DC: Urban Institute.

Castellon-Wellington, M. (2000). *The impact of preference for accommodations: The performance of English language learners on large-scale academic achievement tests* (CSE Technical Report 524). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Cazden, C. (1986). Classroom discourse. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 432–463). New York: Macmillan.

Center on Education Policy. (2006). Ten big effects of the No Child Left Behind Act on public schools. *Phi Delta Kappan, 88*(2), 1110–1113.

Crawford, J. (2004). *Educating English learners: Language diversity in the classroom* (5th ed.). Los Angeles: Bilingual Education Services.

Cuevas, G. J. (1984). Mathematics learning in English as a second language. *Journal for Research in Mathematics Education, 15*(2), 134–144.

Durkin, D. (1978-1979). What classroom observations reveal about comprehension instruction. *Reading Research Quarterly, 14,* 481–533.

Fuhrman, S. (2003). Riding waves, trading horses. The twenty-year effort to reform education. In D. T. Gordon (Ed.), *A nation reformed? American education*

20 years after A Nation At Risk (pp. 7–22). Cambridge, MA: Harvard Education Press.

Gándara, P., Rumberger, R., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in California schools: Unequal resources, unequal outcomes. *Educational Policy Analysis Archives, 11*(36). Retrieved November 1, 2006, from http://epaa.asu.edu/epaa/v11n36/

*Garcia Duncan, T., del Rio Parent, L., Chen, W., Ferrara, S., Johnson, E., Oppler, S., et al. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Applied Measurement in Education, 18*(2), 129–161.

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128.

*Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English-language learners. *Applied Measurement in Education, 16*(2), 159–188.

Johnson, E., & Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assessment for Effective Intervention, 29*(3), 35–45.

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

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

Lager, C. A. (2006). Types of mathematics-language reading interactions that unnecessarily hinder algebra learning and assessment. *Reading Psychology, 27,* 165–204.

Lotherington-Woloszyn, H. (1993). *Do simplified texts simplify language comprehension for ESL learners?* (Technical report). Washington, DC: Office of Educational Research and Improvement.

Martiniello, M. (2007). *Linguistic complexity and differential item functioning (DIF) for English language learners (ELL) in math word problems.* Unpublished doctoral dissertation, Harvard Graduate School of Education, Cambridge, MA.

Miller, E. R., Okum, L, Sinai, R., & Miller, K. S. (1999, April). A study of the English language readiness of limited English proficient students to participate in New Jersey's statewide assessment system. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

National Center on Education and the Economy. (1998). *New Standards: Performance standards and assessments for the schools.* Retrieved September 4, 2008, from http://www.ncee.org/store/products/index.jsp?setProtocol=true&stSection=1

National Center for Education Statistics. (2005). *National Assessment of Educational Progress, 2005.* Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved November 1, 2006, from http://nces.ed.gov/nationsreportcard/

Ragan, A., & Lesaux, N. (2006). Federal, state, and district level English language learner program entry and exit requirements: Effects on the education of language minority learners. *Education Policy Analysis Archives, 14*(20). Retrieved November 3, 2006, from http://epaa.asu.edu/epaa/v14n20/

RAND Mathematical Study Panel. (2003). *Mathematical proficiency for all students.* Santa Monica, CA: RAND.

RAND Reading Study Group. (2002). *Reading for understanding: Toward a research and development program in reading comprehension.* Santa Monica, CA: RAND.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rivera, C., Collum, E., & Shafer Willner, L. (Eds.). (2006). *State assessment policy and practice for English language learners: A national perspective.* Mahwah, NJ: Lawrence Erlbaum.

*Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment, 9*(3–4), 79–105.

Roser, N., & Juel, C. (1982). Effects of vocabulary instruction on reading comprehension. In J. A. Niles & L. A. Harris (Eds.), *Yearbook of the National Reading Conference: Vol. 31. New inquiries in reading research and instruction* (pp. 110–118). Rochester, NY: National Reading Conference.

SAS Institute. (1999). *SAS version 8.* Cary, NC: Author.

Scott, J. A., Jamieson-Noel, D., & Asselin, M. (2003). Vocabulary instruction throughout the day in twenty-three Canadian upper-elementary classrooms. *Elementary School Journal, 103,* 269–283.

*Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's grade 4 mathematics performance assessment* (Technical Report 486). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Sireci, S., Li, S., & Scarpati, S. (2003). *The effect of test accommodation on test performance: A review of the literature* (Research Report 495). Amherst: University of Massachusetts School of Education, Center for Educational Assessment.

Watts, S. M. (1995). Vocabulary instruction during reading lessons in six classrooms. *Journal of Reading Behavior, 27,* 399–424.

Authors

MICHAEL J. KIEFFER is an advanced doctoral student in language and literacy at the Harvard Graduate School of Education, Larsen 303, 14 Appian Way, Cambridge, MA 02138; e-mail *Michael_Kieffer@post.harvard.edu*. His research interests focus on the language and literacy development of adolescent readers, especially English language learners in urban contexts. A former middle school teacher, he also aims to develop and evaluate instructional approaches to meet the needs of struggling adolescent readers.

NONIE K. LESAUX is Marie and Max Kargman Associate Professor in Human Development and Urban Education Advancement at the Harvard Graduate School of Education, Larsen 319, 14 Appian Way, Cambridge, MA 02138; e-mail: *lesauxno@gse.harvard.edu*. Her research interests focus on the reading and language development of at-risk learners, including students from linguistically diverse backgrounds, and effective instructional approaches to prevent reading difficulties.

MABEL RIVERA is a research assistant professor at the Texas Institute for Measurement, Evaluation, and Statistics in the University of Houston, 100 TLCC Annex; Houston, TX 77204–6022; e-mail: *Mabel.Rivera@times.uh.edu*. She is a former teacher of first grade students and students with special needs in the public school system. Her current research interests include the education and prevention of academic difficulties in English language learners. In addition, she engages in local and national service activities related to preparing personnel to teach English language learners and students with special needs.

DAVID J. FRANCIS is Hugh Roy and Lillie Cranz Cullen Distinguished Professor and chairman of the Department of Psychology at the University of Houston, Department of Psychology, Houston, TX 77204–5022; e-mail: *dfrancis@uh.edu*. He also directs the Texas Institute for Measurement, Evaluation, and Statistics, and the National Research and Development Center for English Language Learners, funded by IES (Institute of Educational Sciences). His research interests focus on the application of advanced statistical and psychometric methods to problems in education, especially the identification, classification, and remediation of learning and developmental disabilities, and improving educational outcomes for English language learners and other at-risk populations.