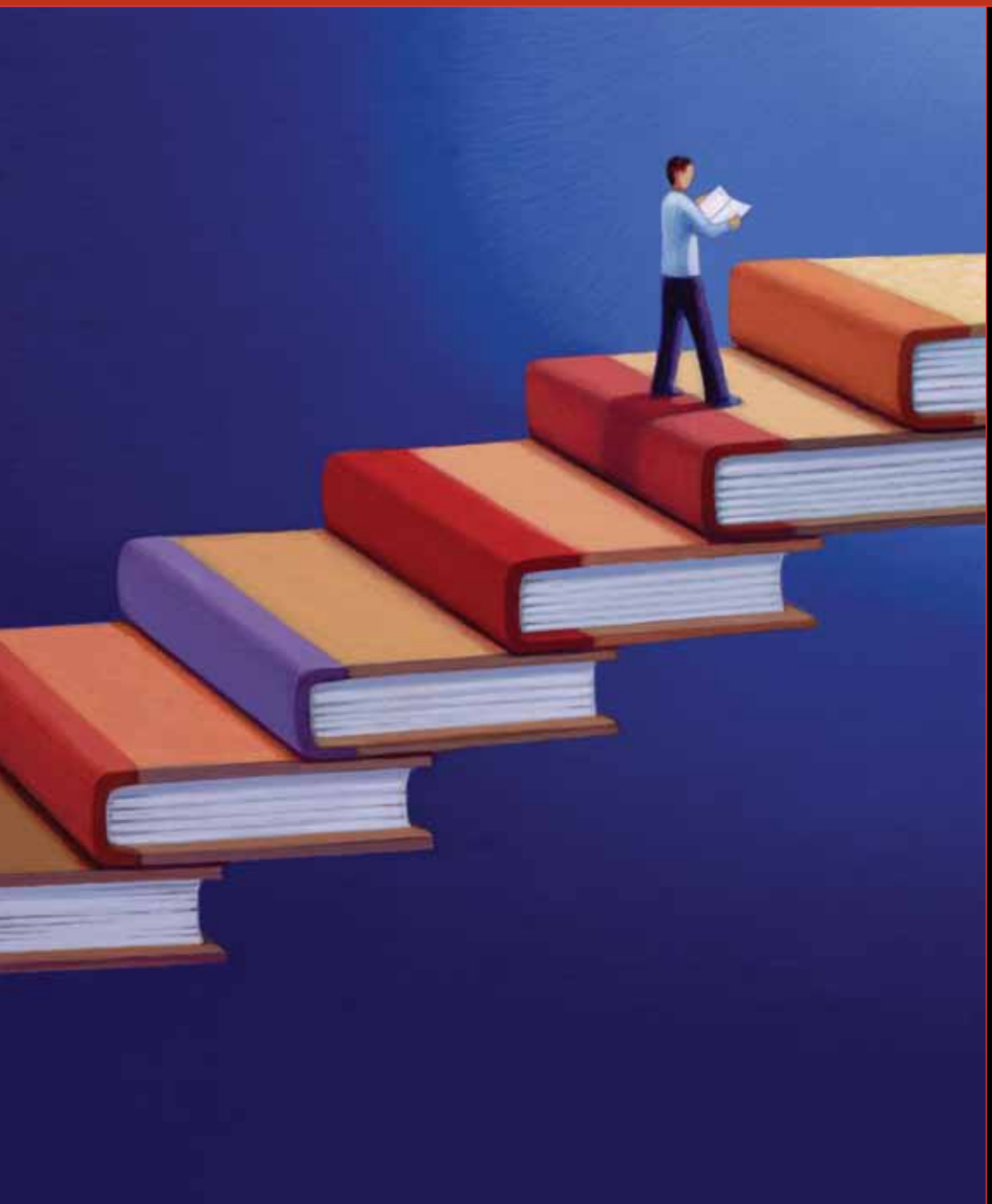
Book 4

Practical Guidelines for the Education of English Language Learners

RESEARCH-BASED RECOMMENDATIONS FOR THE USE OF ACCOMMODATIONS IN LARGE-SCALE ASSESSMENTS

2012 Update



CENTER ON INSTRUCTION

Book 4

Practical Guidelines for the Education of English Language Learners

RESEARCH-BASED RECOMMENDATIONS FOR THE USE OF ACCOMMODATIONS IN LARGE-SCALE ASSESSMENTS

2012 Update

Michael J. Kieffer
New York University

Mabel Rivera
University of Houston

David J. Francis
University of Houston

This is Book 4 in the series
Practical Guidelines for the Education of English Language Learners:
Book 1: Research-based Recommendations for Instruction and Academic Interventions
Book 2: Research-based Recommendations for Serving Adolescent Newcomers
Book 3: Research-based Recommendations for the Use of Accommodations in Large-scale Assessments
Book 4: Research-based Recommendations for the Use of Accommodations in Large-scale Assessments. 2012 Update

CENTER ON INSTRUCTION

This publication was created by the Center on Instruction, which is operated by RMC Research Corporation in partnership with the Florida Center for Reading Research at Florida State University; Instructional Research Group; Lawrence Hall of Science at the University of California, Berkeley; Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston; and The Meadows Center for Preventing Educational Risk at The University of Texas at Austin.

The authors acknowledge the editorial and production support provided by Angela Penfold, C. Ralph Adler, and Robert Kozman of RMC Research Corporation.

The development of this document was supported by the U.S. Department of Education, Office of Elementary and Secondary Education and Office of Special Education Programs, under cooperative agreement S283B050034. However, these contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

Preferred citation:
Kieffer, M.J., Rivera, M., and Francis, D.J. (2012). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments. 2012 update.* Portsmouth, NH: RMC Research Corporation, Center on Instruction.

CONTENTS

1 EXECUTIVE SUMMARY

3 OVERVIEW

4 English language learners in U.S. schools
5 Use of test accommodations
6 Summary of recent meta-analytic reviews of test accommodations for ELLs
8 Method for the new meta-analysis
8 Study inclusion criteria
9 Search procedure
10 Studies included in the meta-analysis
10 Test accommodations investigated
12 Data analysis
14 Results and recommendations
16 Concluding thoughts

19 REFERENCES

25 APPENDICES

26 Appendix A. Characteristics of studies included in the meta-analysis on effectiveness of accommodations
31 Appendix B. Comparisons of included studies to those included in Francis et al. (2006) and Pennock-Roman (2011) with reasons for inclusion for newly added studies since Francis et al. (2006)
32 Appendix C. Discussion of the choice of unit of analysis
33 Appendix D. Reporting of technical results

EXECUTIVE SUMMARY

This report presents results from a new quantitative synthesis of research on the effectiveness and validity of test accommodations for English language learners (ELLs) taking large-scale assessments. In 2006, the Center on Instruction published a review of the literature on test accommodations for ELLs titled *Practical Guidelines for the Education of English Language Learners: Research-based Recommendations for the Use of Accommodations in Large-Scale Assessments* (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006). This new publication provides an update to the 2006 report, incorporating evidence from nine studies not previously included and providing updated recommendations for educators and policy-makers. Results drawn from 20 studies (including, in total, more than 33,000 students, of whom more than 9,400 were ELLs) were aggregated using meta-analysis. The studies were primarily conducted using researcher-created tests of mathematics and science with items drawn from the National Assessment of Educational Progress (NAEP) in grades 4 and 8. Drawing on the existing evidence, we suggest the following recommendations, ordered by the strength of the available evidence:

1. Use simplified English in test design, eliminating irrelevant language demands for all students. Our results indicated that simplified English had a small but statistically significant effect on the test performance of ELLs. This effect was equivalent to a 9 percent to 19 percent reduction in the observed achievement difference between ELLs and non-ELLs. Consistent with the recommendations of experts in this area (Abedi et al., 2004), we recommend that this accommodation should be integrated into test design for all students, making sure to eliminate irrelevant language demands while retaining relevant academic language demands central to the tested constructs. Relatively strong evidence exists for this recommendation in that it was based on the results from 12 studies with appropriate experimental, quasi-experimental, or counter-balanced repeated-measures designs.

2. Provide English dictionaries/glossaries to ELLs. Our findings indicated that providing English dictionaries or specialized glossaries also had a small but statistically significant effect on the test performance of ELLs. This effect was equivalent to an 11 percent to 21 percent reduction of the observed achievement difference between ELLs and non-ELLs. Relatively strong

evidence exists for this recommendation in that it was based on the results from nine studies with appropriate designs.

3. Match the language of tests and accommodations to the language of instruction. Our results indicated that there was no average effect of any of four native language accommodations investigated. However, we found considerable variation across the effects of this accommodation. These results suggest that the effectiveness of native language accommodations will differ based on a variety of factors; the most important may be the language of instruction. The evidence for these accommodations is limited. We found only four qualified studies for the most studied accommodation in this group (providing bilingual dictionaries) while three or fewer studies were investigated for the other accommodations. Consistent with experts in this area (e.g., Abedi et al., 2004), we therefore recommend that the language of tests and accommodations should match the language of instruction.

4. Provide extended time to ELLs or use untimed tests for all students. Our results also indicated that providing extended time to ELLs to complete tests yielded a small but statistically significant effect. This effect equaled a 15 percent to 31 percent reduction in the observed achievement difference between ELLs and non-ELLs. However, the evidence for this recommendation is weaker than for recommendations 1–3 with only three qualified studies found for our meta-analysis. Moreover, a movement toward untimed tests in large-scale assessments may obviate the need for this accommodation.

Despite this evidence of effectiveness for some accommodations, educators and policy-makers should be aware that accommodations alone cannot eliminate the achievement gaps between ELLs and non-ELL students. Our results suggest that while accommodations may help to reduce irrelevant language demands that depress ELLs' test scores, real achievement differences remain, requiring concerted efforts to improve instruction to teach ELLs the academic language and knowledge they will need to succeed in the content areas.

OVERVIEW

Assessment of content knowledge for English language learners (ELLs) is a challenging task given the strong relationship between language proficiency and content learning. Language plays an integral role in academic learning, so any test of academic achievement also assesses, to some degree, language ability. This confounding of language and content area knowledge raises serious concerns about whether ELLs' test scores reflect English language abilities not relevant to the target of the assessment. Indeed, correlational research suggests a substantial link between ELLs' English language proficiency and their performance on math, science, and social studies tests (e.g., Abedi & Leon, 1999; Bailey, 2005; Butler & Castellon-Wellington, 2005) and demonstrates that linguistically complex assessments and individual test items yield larger performance gaps between ELLs and non-ELLs (e.g., Abedi, Lord, Hofstetter, & Baker, 2000; Abedi, Leon, & Mirocha, 2003; Abedi, Lord, & Plummer, 1997; Martiniello, 2007). These relationships do not necessarily imply that we cannot make valid inferences about the content knowledge of ELLs with traditional content area assessments. Rather, they underscore the importance of distinguishing between language abilities central to the academic skills being measured and language demands of the test that are not relevant to the skills and abilities being measured. In this document, we refer to the former as *essential language skills* and to the latter as *irrelevant language skills,* because of their lack of relevance to the academic constructs being measured.

Test accommodations can reduce the influence of irrelevant language skills on test performance for ELLs. Test accommodations for ELLs are changes to the test format or the conditions under which students are tested. They are designed to minimize language-related obstacles that ELLs encounter during testing. The Center on Instruction's report on test accommodations for ELLs (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006) included a review of the literature and the results of a meta-analysis of studies that investigated the effects of test accommodations on the performance of ELLs in large-scale assessments. Research based on that report was later published (Kieffer, et al., 2009) in a form better suited to the expectations and demands of education researchers. Since the publication of those documents, new studies of test accommodations and another review of the accommodations literature (Pennock-Roman & Rivera, 2011) have emerged. In this new document, we

provide updated recommendations on the use of test accommodations for ELLs based on a newer meta-analysis that incorporates evidence published since the development of the Practical Guidelines series and the completion of the Pennock-Roman and Rivera (2011) review.

English language learners in U.S. schools

Success in today's competitive world requires a well-rounded education and the ability to apply knowledge of content and language skills effectively across domains. Students must develop mastery of listening, speaking, reading, and writing skills to process, understand, and apply the meaning of content. Success in school represents a dual challenge for ELLs who must become proficient in English as they acquire content knowledge.

The continuous growth in the population of ELLs in U.S. schools has raised awareness and driven a search for practices that would effectively meet their linguistic needs. The National Clearinghouse for English Language Acquisition reports that ELL enrollment grew by approximately 64 percent between 1994 and 2010, while the total school enrollment grew by only 4 percent (NCELA, 2011). The 4.6 million ELLs in U.S. schools now make up 9 percent of the total student population (U.S. Department of Education, Office Elementary and Secondary Education [OESE], 2012).

TEN MOST COMMONLY USED TEST ACCOMMODATIONS FOR ELLS AS RECOMMENDED BY STATE POLICIES

- Use of dual language dictionary
- Extended time
- Reading items aloud
- Translating directions orally into native language
- Clarifying/explaining directions in English
- Repeating directions
- Reading directions aloud
- Allowing student to respond orally in English and describing responses
- Clarifying/explaining directions in the native language
- Simplifying directions

(Shafer Willner, Rivera, & Acosta, 2008)

Use of test accommodations

Along with effective instructional strategies, the education of ELLs must also include practices that facilitate accurate assessment. The use of test accommodations can improve the accuracy of assessing ELLs' learning and progress. These accommodations serve to minimize the language barriers that negatively impact ELL students' ability to demonstrate what they have learned when taking content area assessments administered in English. Test accommodations are changes to the test or testing environment intended to minimize obstacles that students may encounter during the assessment process. Accommodations for ELLs should respond to their language-related needs and remove language barriers, either directly, for example, with changes to the language of the test or indirectly by changing the testing environment (Acosta, Rivera, & Shafer Willner, 2008). (See the sidebar on page 4 for a list of the accommodations most frequently used for ELLs in large-scale assessments, as recommended by state policies.)

While accommodations for ELLs are used during assessment, they have value during instruction as well. Experts recommend that accommodations used during assessment should also be offered during instruction to increase familiarity with the procedures and to improve ELLs' access to English language instruction, i.e., to allow ELLs to profit more fully from content area instruction.

Accommodations for ELLs must be *effective*, *valid*, and *feasible*. *Effective* accommodations should lead to improved scores for content knowledge. *Valid* accommodations can be used during testing without altering the construct being assessed. Proof of an accommodation's validity comes from evidence that the accommodation is differentially effective: the accommodation should improve test performance if the student needs the accommodation but it should not improve or hinder performance if the student does not need the accommodation. Such evidence of differential effectiveness represents the first essential condition of a test accommodation's validity, but is not the only condition. The validity of a test accommodation can never be proved, but can only be argued for or against on the basis of available evidence. Finally, accommodations must also be *feasible* to implement if they are to be useful. High cost, complexity, and burden of implementation can render useless an otherwise effective accommodation. An effective and valid accommodation that is too costly or too complex to implement accurately may not provide the intended benefit.

With these qualities of accommodations in mind, this document has two purposes: to provide an updated synthesis of the body of research on the effectiveness and validity of test accommodations for ELLs since 2006 and to provide updated recommendations for policy and practice based on these findings. We begin by summarizing the work from our team's prior review and the recently published review by Pennock-Roman and Rivera (2011), who reached somewhat different conclusions than we did. Because we expect that readers may be familiar with these earlier reports, we offer a brief summary of the important differences in the studies' conclusions and approaches. We do not intend to critique the earlier work or extensively compare and contrast the different reviews. Instead, we offer context for the current review and the differing conclusions. We then describe the current work in greater detail, reviewing the methods of our investigation and updating findings and recommendations.

Summary of recent meta-analytic reviews of test accommodations for ELLs

In their 2006 work, Francis et al. found evidence supporting the efficacy and validity of certain test accommodations for ELLs. The meta-analysis, drawing on data from 11 studies conducted before July 2006, yielded a total of 37 estimates of effect size (i.e., an estimate of the effectiveness of an accommodation), each providing information about one of seven possible accommodations: simplified English, English dictionaries or glossaries, bilingual dictionaries or glossaries, extra time, dual language booklets, dual language booklets read aloud, and native language tests. The results of their earlier analysis demonstrated that the use of English language dictionaries or glossaries was the only accommodation of the seven found to have a statistically significant and positive average effect size (mean effect size=.15; p=.001), an effect equal to a decrease in the achievement gap between ELLs and non-ELLs of between 8 percent and 24 percent, depending on the estimates of the achievement gap used (Kieffer et al., 2009). Although we found little evidence for the effectiveness of other accommodations, the data were limited: many of the accommodations had been examined in only a small number of studies.

In their 2011 study, Pennock-Roman and Rivera reached somewhat different conclusions. Their review covered a larger set of studies (14) conducted between 1990 and 2007 and used a slightly different classification

of accommodations for investigating the effects of specific accommodations. In addition, these researchers asked a slightly different set of questions in examining the effectiveness and validity of linguistic accommodations. Specifically, Pennock-Roman and Rivera attempted to investigate whether the format and conditions under which an accommodation was provided and the characteristics of the ELLs to whom accommodations were given led to differences in the effectiveness of the accommodation. For instance, by interpreting and comparing the effect sizes from individual studies, they concluded that computer-administered (pop-up) glossaries were effective even when time limits were restricted. They similarly concluded that, overall, accommodations provided with less restricted time were more effective than accommodations provided with more restricted time.

In addition, Pennock-Roman and Rivera attempted to provide a more nuanced examination of the role of simplified English (also known as plain English) by examining within-study variation in effect sizes for students with different levels of language proficiency. Although the simplified English accommodation had very small average effect sizes, Pennock-Roman and Rivera concluded that this accommodation's effectiveness varied based on the language proficiency of the tested students: specifically, they found that the accommodation may be much more effective for ELLs at intermediate levels of English language proficiency. Finally, in contrast to Francis et al, who concluded that native language assessments showed variability in their effect but did not speculate on the conditions under which they might be effective because of the limited number of studies, Pennock-Roman and Rivera concluded that Spanish language assessments had the greatest effect for Spanish-speaking students with low proficiency in English. They reached this conclusion on the basis of the effect sizes reported for different samples from a single study.

Several differences exist in the methodological approaches undertaken by Pennock-Roman and by Francis et al., resulting in different conclusions. Although a point-by-point comparison of the approaches exceeds the scope of this report, the studies differed substantially in their decisions to emphasize either mean effect sizes (and tests of statistical significance and heterogeneity) aggregated across studies (Francis et al.) or to interpret the magnitudes of the effects from individual studies (Pennock-Roman & Rivera).

The decision in our prior and current meta-analyses to emphasize mean effect sizes aggregated across studies sought to minimize the influence of

idiosyncrasies specific to individual studies. In so doing, we acknowledged that we know more about the effectiveness of accommodations that have been studied more often (i.e., simplified English and English dictionaries and glossaries). In contrast, Pennock-Roman and Rivera interpreted the magnitudes of effect sizes for individual studies, basing some conclusions on the effects found in one or two studies or in one or two samples within a single study. Given the limited research on some accommodations in some contexts and the challenges facing schools in educating and testing ELL students, such differences in emphasis across reviews would be expected.

Method for the new meta-analysis

Study inclusion criteria. Based on the established research questions, we selected four characteristics that determined the criteria for the inclusion of studies in our new review. We included studies in the meta-analysis that (a) examined individual accommodations or individual accommodations bundled with extra time; (b) were published in peer-reviewed journals or in technical reports available online, or were unpublished in dissertations available online; (c) employed an experimental, quasi-experimental, or counter-balanced, repeated-measures design (see the sidebar on this page for definitions of terms); and (d) reported sufficient data to allow for the estimation of effect sizes. Criteria (a) and (d) were identical to the criteria we used in Francis et al. (2006) and were similar to those used by Pennock-Roman and Rivera (2011).

DEFINITIONS OF TECHNICAL TERMS

Meta-analysis. A method for statistically summarizing results from different research studies in order to understand why results differ across studies.

Repeated-measures design. In studies that use this design, the research participants experience more than one treatment. A counterbalanced, repeated-measures design includes features to prevent the order in which the treatments are introduced to participants from influencing inferences about treatment effects.

Between-subjects design. In studies that use this design, independent groups of subjects experience different conditions and their performance is compared. The simplest between-groups design occurs with two groups, where one group receives the accommodation and the other does not.

Criterion (b) differs from our previous criteria because we did not target dissertation databases (not because we specifically chose to exclude such studies). However, in the interest of including as many relevant studies as possible and given that Pennock-Roman and Rivera included dissertations, we decided to include them in this update. Nonetheless, we tested whether the results changed if we included the one dissertation study. Criterion (c) differs from our earlier criteria in that we did not previously include studies with repeated-measures designs based on concerns about the potential statistical comparability of results from these studies with the results from the majority of studies, which used between-groups designs. We did, however, review these studies narratively. In this update, after weighing statistical concerns against the added information provided by these studies, we decided to include them, consistent with the criteria used by Pennock-Roman and Rivera (2011), while also investigating whether their effects differ systematically from the effects of studies with between-groups designs.

TEST ACCOMMODATIONS INVESTIGATED AND NUMBER OF SAMPLES INVOLVED

Simplified English: **24**

English dictionaries or glossaries: **18**

Bilingual dictionaries or glossaries: **6**

Native language tests: **5**

Dual language booklets: **5**

Extended time: **3**

Dual language booklets read aloud: **1**

Reading the test aloud: **2**

Small group administration: **1**

Search procedure. In this new analysis, we used two steps to identify and select studies for inclusion. First, we searched online databases including ERIC, PsychInfo, MLA, Education Abstracts, Academic Search Premier, and Dissertation Abstracts International as well as the online database for the National Center for Research on Evaluation, Standards, and Student Testing at the University of California, Los Angeles. Second, we collected studies previously reviewed by Sireci, Li, and Scarpati (2003); Abedi, Hofstetter, and Lord (2004); and/or Pennock-Roman and Rivera (2011). We read the abstract to determine if a study met our inclusion criteria. The previous search (Francis et al., 2006) included studies conducted before July 2006, while the updated search included studies conducted before May 2012.

Studies included in the meta-analysis. The updated meta-analysis examined 20 studies. These included 11 studies previously employed in Francis et al. (2006), five new studies conducted since July 2006, three studies conducted before July 2006 using a counter-balanced, repeated-measures design, and one dissertation study conducted before July 2006. Appendix A provides details on the included studies. Appendix B lists the included studies along with an indication of whether the study was included in Francis et al. (2006) and/or Pennock-Roman and Rivera (2011) and an explanation of why the study was newly included since Francis et al. (2006). As shown in Appendix B, all studies included in Francis et al. and Pennock-Roman and Rivera have also been included in the current meta-analysis. We do not elaborate here on why individual studies were excluded from the meta-analysis, except to say that they did not meet one of the four criteria described above (details are available from the first author).

Test accommodations investigated. The new meta-analysis investigated the effect sizes of nine test accommodations (see the sidebar on page 9) in 20 studies (see Appendix A). These accommodations are designed to support the linguistic needs of ELLs and to remove or minimize language barriers during assessment. However, as shown below, not all of the accommodations are appropriate on these grounds. We note that the investigated test accommodations do not necessarily match those most commonly used in state testing programs, though there is considerable overlap.

Some test accommodations for ELLs belong to the category of direct linguistic supports. For example, the simplified English accommodation involves linguistic changes in the vocabulary and syntax of test items to eliminate irrelevant complexity while keeping the content the same. Some of these changes may be accomplished by eliminating non-content-related vocabulary, shortening sentences, and using simple sentence structures where possible, using familiar or frequently used words, active instead of passive voice, and using present verb tense where possible.

English, English pop-up, and bilingual dictionaries or glossaries. The use of *English dictionaries or glossaries* involves the use of supplemental English materials in the form of dictionaries or adding definitions or simple paraphrases for potentially unfamiliar or difficult words in test booklets and materials (usually in the margins). Dictionaries usually provide pronunciation guides and multiple meanings of words, where glossaries provide the meaning of a word that applies to the referenced work. Another variation on this

accommodation provides computerized tests or instructional materials with built-in English glossaries, which we will refer to as *English pop-up dictionaries or glossaries*. In this latter variation, a computer program provides a simple and item-appropriate synonym for each difficult non-content word in a test. Similarly, *bilingual dictionaries or glossaries* define words both in English and the student's native language.

Dual language test booklets or instructional materials and native language tests or instructional materials. Some types of direct linguistic accommodations involve the use of native language. For example, *dual language test booklets or instructional materials* encompass format changes. Most booklets or materials include English items on one side and the corresponding items translated into the learner's first language on facing pages. *Native language tests or instructional materials* are translated to the student's primary language. Typically, these materials have been adapted to preserve the meaning of the text rather than providing a simple translation. Back translation, the most highly preferred method of adapting a test to another language, first translates the test from its original language to the native language version by a proficient speaker, reader, and writer of both languages. An independent, bilingually proficient individual then translates the adapted test back into the original language and the two original language tests are compared for equivalence. If the two original language versions are deemed to be different, the process repeats, focusing on correcting unsuccessfully adapted areas of the test.

Read aloud or oral administration. In this direct linguistic accommodation, directions, items, or both may be read aloud or orally presented to the student in English or in the native language. This accommodation is used when the language of test items or directions is complicated enough to interfere with students' ability to access difficult content.

Extended or extra time. Providing more time than usual to complete test sections and instructional tasks may be considered an indirect linguistic accommodation. This accommodation, known as *extended* or *extra time*, makes changes to the testing conditions, not to the test itself. Extended time may be provided in combination with other types of accommodations or offered alone. In either case, the student is awarded extra time to process the language of the test. When bundled with another accommodation such as an English language dictionary, the extra time is intended to offset the time required to use the other accommodation.

Small group administration. Other accommodations typically provided for students with disabilities have not been considered as responsive to the needs of ELLs unless presented in a bundle with other accommodations more closely related to the students' linguistic needs. For example, *small group administration* involves testing or providing instruction to a small number of students under adult supervision in a quiet resource room, such as the school library. We included this accommodation in our meta-analysis because it has been studied for ELLs, though we do not necessarily recommend it, given that it is not considered responsive to the specific needs of ELLs. That said, there may be a subset of ELLs, such as those with diagnosed attention problems, who, like their non-ELL counterparts with such problems, might benefit from such an accommodation.

Data analysis.[1] To determine the effectiveness of specific test accommodations for ELLs, we estimated the difference in test scores between accommodated and un-accommodated conditions for ELLs using Hedges' g^u as our effect size index, i.e., our metric for evaluating the effectiveness of an accommodation for improving student performance. Each study included in the meta-analysis yielded one or more effect sizes depending on the number of unique accommodations investigated, grade groups, and content areas included in the study. For instance, a single study that investigated the accommodation of simplified English and the accommodation of providing English dictionaries (provided to different treatment groups) for grade 4 mathematics would yield two effect sizes. Similarly, a single study that focused on one type of test accommodation for math and science outcomes for students in grades 4 and 8 would yield a total of four effect sizes. In our meta-analysis, we used these individual effect sizes from samples as the unit of analysis. The complicated issue of unit of analysis merits some discussion and we have provided Appendix C to clarify this issue. To shed light on the impact of accommodations on the validity of inferences based on test scores, we also estimated the effects of accommodations on non-ELLs' test scores. Non-ELLs in these studies primarily included native English speakers, but sometimes also included students formerly classified as ELLs and/or students who were bilingual but initial classified as English proficient when they entered school. For each accommodation, we averaged across all studies investigating that

[1] In this section and its accompanying sidebar (page 13), we provide a moderately technical introduction to the methods of meta-analysis used in this report. Although we have attempted to shape this section for readers with little or no experience with meta-analysis, readers who are not interested in these details can skip to the next section without loss of continuity.

accommodation, averaging across different outcomes and grades and weighting individual effect sizes according to their precision. We averaged across different outcomes and grades due to our interest in the overall effects of accommodations and a concern about the limited number of studies for a given accommodation, but we also conducted additional analyses to investigate whether effects appeared to be larger at some grade levels or for some content areas, when made possible by a sufficient number of samples. We also conducted additional analyses to examine the possibility that effect sizes for studies with between-groups designs and studies with repeated-measures designs differed systematically.

DEFINITIONS OF TECHNICAL TERMS

Effect size. A standardized metric for evaluating the effectiveness of a treatment (in this case an accommodation) in improving student performance. Effect sizes near .20 are considered small, near .50 are considered medium, and near .80 are considered large. We can also interpret effect sizes by comparing them to other benchmark; in our study, we compared the effects of accommodations to the achievement gaps between ELLs and non-ELLs.

Moderator analysis. Correlational tests to determine if the size of the effect size is related to features of the sample or studies included. For instance, the effect of a given accommodation might hypothetically be more effective at grade 4 than at grade 8, in which case moderator analyses can shed light on whether the sizes of accommodation effects were related to whether the samples were in grade 4 or grade 8. It is important to note, however, that such analyses cannot support causal conclusions because one moderating variable may be related to other, unobserved moderating variables. For instance, if the studies with grade 4 samples were all conducted in one state while the studies with grade 8 samples were conducted in another state, we could not determine whether grade level or state was the cause of the differences in effectiveness.

Results and recommendations

Our meta-analysis yielded several results summarized in Table 1 (page 18) and Table 2 (page 19) and presented in greater detail in Appendix D, which reports results with a moderate level of technical detail for readers with little or no background in meta-analysis. For readers with less interest in those technical details, we summarize the results briefly in this section and then present our recommendations. For a detailed presentation of the results and the complete basis for the recommendations, please see Appendix D.

We found, as we did in our previous meta-analysis, that the observed achievement gaps in mathematics and science between ELLs and non-ELLs were large when accommodations were not provided. These large gaps provide a metric for understanding the effectiveness of specific accommodations, namely by describing the extent to which use of an accommodation would reduce the achievement gap. We express this impact as a percentage of the observed achievement gap.

More importantly, we found that three of the accommodations studied had statistically significant, small average effects on ELLs' test performance: *simplified English*, *providing English dictionaries or glossaries*, and *providing extra time*. The first two of these accommodations have been studied more frequently, giving us more confidence in their effectiveness than in the effectiveness of providing extra time. In contrast, the following accommodations yielded non-significant average effects for ELLs: *providing dual language booklets, providing dual language questions read aloud in the first language, reading tests aloud in English*, and *small group administration*. Note, however, that none of these latter accommodations has been studied extensively. In addition, there was heterogeneity in the effects of two native language accommodations—*providing bilingual dictionaries or glossaries* and *translating assessments*—suggesting that these accommodations may be effective for some students but not for others, or under some circumstances but not under others. Finally, none of the accommodations was found to significantly improve the performance of non-ELLs, providing supporting evidence for the conclusion that these accommodations do not alter the construct being measured for ELL students on these assessments.

Based on our findings, we offer the following recommendations:

1. Use simplified English in test design, eliminating irrelevant language demands for all students.

Our previous findings (Francis et al., 2006) indicated that simplifying the English of assessments did not yield a statistically significant effect on the performance of ELLs. However, the current findings, which benefitted from eight additional samples drawn from five new studies, indicated that this accommodation had a small but significant effect, equal to a 9 percent to 19 percent reduction in observed achievement gaps. We therefore now recommend this accommodation for use in large-scale assessments, with a reminder of the complicated nature of altering test materials and encouragement to follow the thoughtful recommendations of Abedi and colleagues (e.g., Abedi, Lord, & Hofstetter, 1998). We also concur with Abedi and colleagues (2004) that this accommodation can be provided to all students by thoughtfully minimizing the irrelevant language demands of tests during the test design process.

2. Provide English dictionaries/glossaries to ELLs.

Results from the meta-analysis supported the recommendation from our earlier study that providing English dictionaries can produce a small but statistically significant effect on ELLs' performance on large-scale assessments, an effect equal to an 11 percent to 21 percent reduction in observed achievement gaps.

3. Match the language of tests and accommodations to the language of instruction.

Results supported the findings from our earlier study that the effects of native language accommodations are heterogeneous, with little evidence of a significant average effect. The finding of wide heterogeneity suggests that the effectiveness of native language accommodations will likely vary by a variety of student and instructional factors, such as native language proficiency, native language literacy, and language of instruction. We therefore continue the recommendation made by Francis et al. (2006), as well as Abedi et al. (2004) that the language of accommodations should match the language of instruction whenever possible.

4. Provide extended time to ELLs or use untimed tests for all students.

Our original findings indicated that providing extra time alone as an accommodation did not have a statistically significant effect on ELLs'

performance, while our current findings, based on three samples drawn from three studies, indicate that this accommodation had a small but significant effect on ELLs' performance, an effect equal to a 15 percent to 31 percent reduction in achievement gaps. Although this result for providing extra time alone shows promise, we reiterate that the weight of evidence for this accommodation comes from a very small number of studies in need of further replication. Moreover, we did not find evidence that bundling extra time with other accommodations yielded greater effects than providing those accommodations without additional time. Consequently, we argue that when time is not central to the construct being measured, using untimed tests with all students obviates the need to provide extra time as an accommodation for ELL students and would not be expected to differentially benefit ELL students using other accommodations, such as dictionaries or glossaries.

Concluding thoughts

Although these findings provide some basis for optimism concerning specific accommodations, we reiterate the conclusions from Francis et al. (2006) and discussed in more detail in Kieffer et al. (2009) that providing test accommodations alone cannot be expected to eliminate content-area achievement gaps between ELLs and non-ELLs. In light of the large achievement differences observed and the relatively small reductions yielded by the accommodations studied, test accommodations should be considered to be only a small part of a much larger effort to improve instruction and assessment for ELLs. Specifically, we reiterate our hypothesis that a much larger proportion of the math and science achievement differences observed between ELLs and non-ELLs result from real differences in relevant academic language between these groups, compared to the proportion due to the irrelevant language demands that can be minimized through accommodations. To the extent that this hypothesis holds up, a substantial need remains to improve the opportunities to learn academic English offered to ELLs in American schools.

Table 1

Average effect sizes, tests of mean effects, and tests of heterogeneity for FIXED EFFECTS ANALYSIS for effectiveness of accommodations for English language learners

Accommodation		Results for fixed effects analysis									
		Number of samples	Effect size and 95% confidence interval				Test of mean effect = 0		Test of heterogeneity in effect sizes		
			Mean effect size	s.e.	Lower limit	Upper limit	*Z*	*p*	*Q*	*df(Q)*	*p(Q)*
Accommodations with significant effects	Simplified English	24	0.14	0.03	0.09	0.19	5.06	<.001	49.27	23	.001
	English dictionary-glossary	18	0.14	0.04	0.07	0.20	3.81	<.001	29.30	17	.032
	Extra time	3	0.23	0.10	0.03	0.44	2.22	.026	0.21	2	.900
Accommodations with non-significant effects	Bilingual dictionary-glossary	6	-0.09	0.06	-0.21	0.04	-1.35	.176	13.97	5	.016
	Spanish version	5	0.09	0.08	-0.07	0.25	1.13	.260	55.47	4	<.001
	Dual language booklet	5	-0.01	0.07	-0.14	0.13	-0.11	.916	2.94	4	.568
	Dual language questions and read aloud in Spanish	1	0.27	0.20	-0.11	0.65	1.40	.161			
	Read aloud	2	0.09	0.16	-0.23	0.41	0.55	.582	0.26	1	.609
	Small group	1	-0.54	0.32	-1.17	0.09	-1.67	.095			
	TOTAL WITHIN								151.42	56	<.001
*	TOTAL BETWEEN								20.49	8	.009
	OVERALL MEAN	65	0.11	0.02	0.07	0.14	5.72	<.001	171.91	64	<.001

Table 2

Average effect sizes, tests of mean effects, and tests of heterogeneity for RANDOM EFFECTS ANALYSIS for effectiveness of accommodations for English language learners

Accommodation		Results for random effects analysis								
		Number of samples	Effect size and 95% confidence interval				Test of mean effect = 0		Test of heterogeneity in effect sizes		
			Mean effect size	s.e.	Lower limit	Upper limit	*Z*	*p*	*Q*	*df(Q)*	*p(Q)*
Accommodations with significant effects	Simplified English	24	0.14	0.05	0.05	0.24	2.90	.004			
	English dictionary-glossary	18	0.16	0.05	0.06	0.25	3.22	.001			
	Extra time	3	0.23	0.10	0.03	0.44	2.22	.026			
Accommodations with non-significant effects	Bilingual dictionary-glossary	6	-0.03	0.12	-0.25	0.20	-0.24	.814			
	Spanish version	5	0.46	0.34	-0.20	1.13	1.37	.170			
	Dual language booklet	5	-0.01	0.07	-0.14	0.13	-0.11	.916			
	Dual language questions and read aloud in Spanish	1	0.27	0.20	-0.11	0.65	1.40	.161			
	Read aloud	2	0.09	0.16	-0.23	0.41	0.55	.582			
	Small group	1	-0.54	0.32	-1.17	0.09	-1.67	.10			
	TOTAL BETWEEN								13.00	8	.112
	OVERALL MEAN	65	0.12	0.03	0.06	0.17	4.21	<.001			

REFERENCES

Asterisks indicate studies included in the current meta-analysis.

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

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

* Abedi, J., Courtney, M., Leon, S., Kao, J., & Azzam, T. (2006). *English language learners and math achievement: A study of opportunity to learn and language accommodation* (CSE Technical Report 702). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

* Abedi, J., Courtney, M., Mirocha, J., Leon, S., and Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CSE Technical Report 666). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

* 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, CA: 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, CA: 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,* 219–234.

* Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Technical Report 478). Los Angeles, CA: 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., Kim, 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., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Technical Report 429). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Acosta, B., Rivera, C., & Shafer Willner, L. (2008, October). *Best practices in the state assessment policies for accomodating English language learners: A Delphi study.* Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

* Aguirre-Muñoz, Z. (2000). *The impact of language proficiency on complex performance assessments: Examining linguistic accommodation strategies for English language learners.* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses (AAT 9973171)

Albus, A., Bielinski, J., Thurlow, M., and Liu, K. (2001). *The effect of a simplified English language dictionary on a reading test* (LEP project report 1). Minneapolis, MN: 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*(4), 245–255.

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

Aud, S., Hussar, W., Johnson, F., Kena, G., Roth, E., Manning, E., Wang, X., and Zhang, J. (2012). *The Condition of Education 2012* (NCES 2012-045). U.S. Department of Education, National Center for Education Statistics. Washington, DC. Retrieved from http://nces.ed.gov/pubsearch

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, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Bailey, A. L., & Huang, B. H. (2011). Do current English language development/ proficiency standards reflect the English needed for success in school? *Language Testing, 28*(3), 343–365.

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). 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, CA: 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: The Urban Institute.

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

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

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

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

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

* 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.

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

National Center for Education Statistics [NCES]. (2009). *National Assessment of Educational Progress, 2009.* Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from http://nces.ed.gov/nationsreportcard/.

National Center for Education Statistics [NCES]. (2011). *National Assessment of Educational Progress, 2011.* Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from http://nces.ed.gov/nationsreportcard/.

National Clearinghouse for English Language Acquisition [NCELA]. (2011). The growing numbers of English learner students. Washington, DC: George Washington University. Retrieved from http://www.ncela.gwu.edu/files/uploads/9/growingLEP_0809.pdf.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30,* 10–28.

* 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.

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

* Sato, E., Rabinowitz, S., Gallagher, C. Huang, C.W. (2010). *Accommodations for English language learner students: the effect of linguistic modification of math test item sets.* (NCEE 2009-4079). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shafer Willner, L. S., Rivera, C., & Acosta, B. D. (2008). Descriptive study of state assessment policies for accommodating English Language learners. Arlington, VA; The George Washington University Center for Equity and Excellence in Education. Retrieved from http://ceee.gwu.edu

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, MA: University of Massachusetts School of Education, Center for Educational Assessment.

U.S. Department of Education, Office of Elementary and Secondary Education [OESE]. (2012). Report to Congress on the Elementary and Secondary Education Act: State-Reported Data for School Year 2009–10. Washington, D.C. Retrieved from http://www2.ed.gov/about/reports/annual/esea-report-2009-2010.doc

* Wolf, M. K., Kim, J., Kao, J. C., & Rivera, N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners* (CSE Technical Report 766). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

APPENDICES

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†
Abedi (2009)	Computerized English glossary, English dictionary, small-group testing, extra time	Yes (for glossary and dictionary)	Experimental	4, 8	Math	NAEP & TIMSS	Not reported	Not reported	Not reported	Not reported
Abedi, Courtney, & Leon (2003a)	English glossary, bilingual glossary, simplified English	Yes	Experimental and quasi-experimental	4, 8	Science	NAEP & TIMSS	Spanish, Chinese, & "other Asian languages"	Not reported	School records	Mean self-reported understanding of oral English halfway between "well" and "very well"
Abedi, Courtney, & Leon (2003b)	English glossary, computerized English glossary, extra time	No	Quasi-experimental	4, 8	Math	NAEP	Spanish, Korean, Chinese, & others	Not reported	School records	75% of 4th graders and 55% of 8th graders self-reported understanding English directions "very well"

† Note: 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 difficult to interpret.

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†
Abedi, Courtney, Leon, Kao, & Azzam (2006)[1]	Dual language booklet, simplified English	No	Experimental	8	Math	California Standards Test, NAEP, TIMSS	Spanish, others not specified	Not reported	School records	Not reported
Abedi, Courtney, Mirocha, Leon, & Goldberg (2005)	English dictionary, bilingual dictionary, simplified English	No—given to both groups	Experimental	4, 8	Science	NAEP	Spanish, Korean, Filipino, Chinese, and others	Bilingual and English-only	School records	Mean self-reported understanding of oral English between "well" and "very well"
Abedi, Hofstetter, Baker, & Lord (2001)	English glossary, simplified English	No	Experimental	8	Math	NAEP	Spanish, Khmer, Vietnamese, Tagalog, Lao, & others	English	School records	Nearly half self-reported understanding and speaking English "very well"

† Note: 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 difficult to interpret.

1 This study only provided appropriate data for meta-analysis from a subset of nine items (from a 32-item test) judged by the authors to require the most extensive simplification. The authors note that this subset of items had low reliability (alpha = .599). Although these differences might yield a different effect size for this individual study, sensitivity analyses, which excluded this study, indicated that the study had little effect on the overall mean effect for this accommodation.

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†
Abedi, Lord, & Hofstetter (1998)	Simplified English	No	Experimental	8	Math	NAEP	Khmer, Vietnamese, & others	English only & Spanish-English bilingual	School records	Half self-reported understanding oral English "very well"
Abedi, Lord, Kim, & 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, & Plummer (1997)[2]	Simplified English	No	Counterbalanced Repeated Measures	8	Math	NAEP	Spanish, Korean, Chinese, Farsi, Filipino, & others	Not reported	School records	Mean self-reported listening & speaking of "very well"
Aguirre-Muñoz (2000)[3]	Simplified English, dual language (with simplified English), Spanish translation	No	Experimental	7	History	Researcher-designed performance assessment	Spanish, others not specified	Not reported	Researcher-created Writing Proficiency Test	43% limited English proficient, 27% intermediate fluent English proficient, 30% redesignated fluent English proficient, based on Language Assessment Scales

† Note: 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 difficult to interpret.

2 We selected this technical report rather than the peer-reviewed version, Abedi & Lord (2001), because it provided more detailed data necessary for the meta-analysis, similar to Pennock-Roman & Rivera (2011).

3 This study provided multiple outcome measures for each sample (i.e., different subtest scores for a history assessment), which produced inappropriately dependent effect sizes. We therefore averaged effects within sample for this study prior to aggregating its effects with the effects from the other studies.

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†
Albus et al. (2005)[4]	English dictionary	No	Counterbalanced repeated measures	8	Reading	Minnesota Basic Standards Test	Hmong	English only	School records	4% beginning, 13% low intermediate, 28% intermediate, 48% high intermediate, 8% high
Anderson et al. (2000)	Dual language questions and read aloud in Spanish	No	Experimental	8	Reading	Minnesota State 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
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"
Johnson & Monroe (2004)	Simplified English	No	Counterbalanced repeated measures	7	Math	Washington State Test	Not reported	Not reported	Not reported	Not reported

† Note: 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 difficult to interpret.

4 We selected this peer-reviewed article rather than the technical report version, Albus, Belinski, Thurlow, & Liu (2001), which reports on the same study, similar to Pennock-Roman & Rivera (2011).

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†
Kopriva et al. (2007)[5]	Picture dictionary, bilingual glossary, read aloud	No	Experimental	3, 4	Math	Researcher developed based on South Carolina standards	Spanish	Not reported	School records	31% low ELL, 30% intermediate ELL, 28% high ELL, 11% on grade level
Rivera & Stansfield (2003)	Simplified English	No	Experimental	4,6	Science	Delaware Science Test	Various	Not reported	Not reported	Not reported
Sato et al. (2010)	Simplified English	No	Experimental	7, 8	Math	Grade 8 NAEP, grade 7 California state test	Spanish	Not reported	School records for ELL status and self-report of Spanish as native language	Early intermediate to advanced on the California English Language Development Test
Wolf et al. (2009)	English glossary, read aloud	No	Experimental	8	Math	NAEP, TIMSS, California state test, State X instructional materials	Spanish, several others	Not reported	School records	

† Note: 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 difficult to interpret.

5 This study investigated the effects of accommodations both when randomly assigned and when individually-recommended for students. We used the data for accommodations that were randomly assigned. Because standard deviations were not provided, we estimated them based on the standard errors for the means and the sample sizes provided.

continued

APPENDIX B

Comparisons of included studies to those included in Francis et al. (2006) and Pennock-Roman (2011) with reasons for inclusion for newly added studies since Francis et al. (2006)

Study	Included in Francis et al. (2006) meta-analysis	Included in Pennock-Roman & Rivera (2011) meta-analysis	Reason for addition to the current meta-analysis
Abedi (2009)	No	No	Published after July, 2006
Abedi, Courtney, & Leon (2003a)	Yes	Yes	
Abedi, Courtney, & Leon (2003b)	Yes	Yes	
Abedi, Courtney, Leon, Kao, & Azzam (2006)	No	No	Published after July, 2006
Abedi, Courtney, Mirocha, Leon, & Goldberg (2005)	Yes	Yes	
Abedi, Hofstetter, Baker, & Lord (2001)	Yes	Yes	
Abedi, Lord, & Hofstetter (1998)	Yes	Yes	
Abedi, Lord, Kim, & Miyoshi (2001)	Yes	Yes	
Abedi, Lord, & Plummer (1997)	No	Yes	Criteria changed to include repeated-measures studies
Aguirre-Muñoz (2000)	No	Yes	Criteria changed to include dissertation studies
Albus et al. (2005)	No	Yes	Criteria changed to include repeated-measures studies
Anderson et al. (2000)	Yes	Yes	
Brown (1999)	Yes	No	
Garcia Duncan et al. (2005)	Yes	Yes	
Hofstetter (2003)	Yes	Yes	
Johnson & Monroe (2004)	No	No	Criteria changed to include repeated-measures studies
Kopriva et al. (2007)	No	Yes	Published after July, 2006
Rivera & Stansfield (2003)	Yes	Yes	
Sato et al. (2010)	No	No	Published after July, 2006
Wolf et al. (2009)	No	No	Published after July, 2006

APPENDIX C

Discussion of the choice of unit of analysis

Like both previous meta-analyses discussed in this document (Francis et al., 2006; Pennock-Roman & Rivera, 2011), we used samples as the unit of analysis. We acknowledge that this approach has both strengths and drawbacks. Given the limited number of studies available, we made this decision to preserve the maximum amount of information in the collection of studies about different accommodations, in different grades, and for different content areas. The alternate strategy of treating the study as the unit of analysis would have required averaging across the effects of different accommodations (as well as across grades and content areas) even though the samples were independent, at least to an extent. In the one case where multiple outcomes (i.e., different subtest scores within a single domain) were reported for each sample (Aguirre-Muñoz, 2000), we averaged the effects within sample prior to aggregating these effects with effects from other samples.

Nonetheless, we note that in some studies, a single control group (i.e., ELLs taking the test without accommodations) was compared to more than one treatment (i.e., more than one accommodated ELL group), rendering some comparisons within a study dependent on one another. Because these different accommodations involving the control group addressed questions about the accommodations in our analysis, this dependence would increase the correlation between findings across different sets of accommodations. Nevertheless, on balance, we felt that this drawback was worth the added information gained by using the sample as the unit of analysis.

APPENDIX D

Reporting of technical results

Preliminary analyses: Describing the achievement gap

To provide a metric for judging the effectiveness of accommodations, we estimated the average observed differences in academic achievement test scores between ELLs and non-ELLs that can be expected on large-scale assessments of math and science in the absence of accommodations. These estimates provide a context for evaluating the practical importance of the effects of accommodations. The top half of Table 3 (see page 30) presents mean effect sizes (reported as Hedges' g^u statistics) for the differences in math and science achievement scores between ELLs and non-ELLs in the un-accommodated conditions from the studies reviewed. These estimates suggest that there are large achievement score differences between the two groups in both math (mean effect size = 0.86) and science (mean effect size = 0.75). In other words, the scores of ELLs and non-ELLs, in the absence of accommodations, differed greatly in both math and science tests used in the studies. To provide another point of reference, we have provided estimates of the achievement difference between non-ELLs and ELLs from the most recent National Assessment of Educational Progress (NAEP), i.e., 2011 for math and grade 8 science and 2009 for grade 4 science. Table 3 shows that these estimates are somewhat larger than the mean estimates from the reviewed studies. These differences in magnitudes may be due to the confounding of concomitant predictors of achievement, such as socioeconomic status, in the national samples, which are likely better controlled in the research studies of accommodations. All of the studies reviewed sampled ELL and

DEFINITIONS OF EFFECT SIZE

Effect size is a standardized metric for evaluating the effectiveness of a treatment (in the case of the present meta-analysis, an accommodation) in improving student performance. Effect sizes near .20 are considered small, near .50 medium, and near .80 large. We also can interpret effect sizes by comparing them to other benchmarks; in our study, we compared the effects of accommodations to the achievement gaps between ELLs and non-ELLs.

non-ELL students from within the same schools and/or districts, whereas the NAEP estimates reflect a nationally representative sample. Nevertheless, both sets of estimates reveal large differences in achievement in both math and science between ELLs and non-ELLs on large-scale assessments, suggesting that we can judge the effectiveness of accommodations based on the extent to which they reduce apparent achievement gaps. We do not suggest that these observed achievement gaps can be removed entirely or only by the use of appropriate accommodations. Rather, the size of the achievement gap simply helps us to interpret the practical importance of providing appropriate test accommodations.

Table 3

Estimates of the achievement score differences between English language learners and non-English language learners speakers in math and science in the absence of accommodations from studies reviewed (as Hedges' g^u) based on a random effects model and from the 2009 and 2011 National Assessment of Educational Progress (as Cohen's d)

Research studies By domain†	Number of samples (Number of studies)	Mean effect size
Math	15	0.86
Science	11	0.75
NAEP		
Grade 4 math (NCES, 2011)		0.86
Grade 8 math (NCES, 2011)		1.20
Grade 4 science (NCES, 2009)		1.21
Grade 8 science (NCES, 2011)		1.50

† The achievement score differences in reading and history were not estimated because only one or two studies examined each of these domains.

Effectiveness of accommodations for ELLs in the reviewed studies

We found evidence for the effectiveness of three test accommodations for ELLs: simplified English, providing English dictionaries or glossaries, and extended time. Below, we first describe the results for these three accommodations, followed by the results for the accommodations with limited evidence of effectiveness. Within these groupings, we organize our discussion based on the weight of evidence, starting with the accommodations that have been studied more frequently, followed by accommodations that have been studied less frequently. Tables 1 and 2 in the main text of this document (see pages 18 and 19) provide the results of the meta-analysis for the effectiveness of individual accommodations for ELLs, including mean effect sizes, standard errors, tests of the mean effect, and tests for heterogeneity by accommodation and overall. (See the sidebar for definitions of terms.) Table 1 provides the results of the fixed effects analysis, while Table 2 provides the results of the random effects analysis. Below we emphasize the results appropriate for the given accommodation (i.e., results from the random effects model when the test for heterogeneity indicated that

DEFINITIONS OF TECHNICAL TERMS

Mean effect size. Effect size for a given treatment (i.e., accommodation in the current study) averaged across the available samples.

Standard error. A statistic for the precision of the effect size. Larger standard errors indicate less precision, i.e., that the effect size may be much larger or smaller than the point estimate.

Test of mean effect. A test of whether the mean effect size is statistically different from 0, i.e., whether the accommodation was effective, on average in the population of ELLs.

Test of heterogeneity. A test of whether the individual effect sizes provided by the different samples different significantly from one another, more than we would expect based on normal sampling error.

Fixed effects model. A meta-analytic model that assumes that there is no heterogeneity in individual effect sizes around a single mean effect size, beyond normal sampling error.

Random-effects model. A meta-analytic model that accounts for the fact that individual effect sizes may differ significantly from each other, beyond normal sampling error

effects differed across samples within accommodation), while Tables 1 and 2 provide results from both analyses in the interest of comprehensive reporting. We also provide results from additional analyses that investigated whether the effects of individual accommodations differed by grade level, content area, or study design (i.e., repeated-measures designs compared with between-groups designs).

Accommodations with evidence of effectiveness

Simplified English. Simplified English had a statistically significant, small effect on ELLs' performance (mean effect size = 0.14; s.e. = 0.05 in the random effects model) based on 24 samples drawn from 12 studies. The mean effect size was statistically significant across a variety of sensitivity analyses, which involved excluding individual studies.[2] When comparing this effect size to the estimates in Table 3, the effect can be considered equal to a 9 percent to 19 percent reduction in the observed achievement difference between ELLs and non-ELLs, depending on the specific estimate used.

The test of heterogeneity indicated that the effect sizes differed across samples (Q = 49.27; p = .001). Moderator analysis indicated that the effect size differed significantly between grades (Q between grades=17.91; p=.001 in the random effects model), with significantly positive effects found for the five samples in grade 7 (mean effect size=0.35; s.e.=0.07 in the random effects model) and for the eleven samples in grade 8 (mean effect size =.13; s.e.=0.04 in the random effects model), but non-significant and/or negative effects in grades 4, 5, and 6. However, we hesitate to attribute this difference simply to grade level because of the small number of samples at each grade level. In particular, the relatively large grade 7 effects contributed by Aguirre-Munoz (2000) and Sato et al. (2010) may have driven this moderating effect and may be due to specific features of these two studies other than student grade level; without these two studies, the moderating effect of grade no longer appears significant, while the average mean effect of this accommodation remains significant. Therefore, we feel confident about the average positive effect of simplified English and the effect variation across samples, but we

[2] In particular, we were concerned that the large effects contributed by Aguirre-Munoz (2000) may have driven this result. This is the only dissertation in the meta-analysis and differs in other, potentially important ways from the other included studies (e.g., it focused on a performance assessment in history unlike the large majority of other studies which focused on more traditional large-scale assessments in math or science). Nonetheless, a sensitivity analysis excluding this study yielded a statistically significant mean effect size (.10; s.e.=0.05) that was only slightly smaller than the mean effect size including this study.

lack confidence that this effect size variance can be clearly attributed to grade. Effects of this accommodation differed also by content area, with much larger and significant effects for the three student samples who took history tests (mean effect size=0.45; s.e.=0.11; all drawn from Aguirre-Munoz, 2000), smaller but significant effects for the 11 samples who took math tests (mean effect size=0.12; s.e.=0.05), and non-significant effects for the 10 samples who took science tests (mean effect size=0.06; s.e.=0.12). Again, we hesitate to attribute this difference to content area because it may be due to idiosyncratic characteristics of the Aguirre-Munoz (2000) study.

Effect sizes did not differ by study design, i.e., whether the study involved a between-groups or repeated-measures comparison (Q between designs=1.71; p=.279 in the random effects model); the mean effect sizes were roughly similar for the two samples from repeated-measures studies (0.07) and the 22 samples drawn from between-groups studies (0.16).

English dictionaries and glossaries. Providing English dictionaries and glossaries had a small but statistically significant effect on students' performance (mean effect size = 0.16; s.e. = 0.05 in the random effects model) based on 18 samples drawn from nine studies. When compared to the estimates in Table 3, this effect equals a reduction of between 11 percent and 21 percent of the observed achievement difference between ELLs and non-ELLs, depending on the specific estimate used. The test of heterogeneity of effect sizes indicated that the effect of this accommodation differed across samples (Q=29.30; p=.021), although none of the hypothesized moderators was found to predict this between-sample variation. Moderator analyses indicated that the effect of this accommodation did not differ between grades (Q between grades=3.21; p=.201 in the random effects model) or between content areas (Q between content areas=0.71; p=.700 in the random effects model).

The format of the glossary or dictionary did not appear to alter the effect size. More specifically, the mean effect size for the four samples in which computerized glossaries were provided (mean effect size=0.27; s.e.=0.09 in the random effects model) appeared to be notably greater than for the 14 samples in which paper and pencil dictionaries or glossaries were provided (mean effect size=0.13; s.e.=.06 in the random effects model); however, moderator analyses indicated that this difference was not statistically significant (Q between computerized and paper and pencil=2.06; p=.151 in the random effects model).

Effect sizes also did not differ by study design, i.e., whether the study involved a between-groups or repeated-measures comparison (Q between designs=0.44; p=.506 in the random effects model); the effect size from the one sample from a repeated-measures study (0.08) appeared similar to the mean effect size from the seventeen samples drawn from between-groups studies (0.17).

Finally, further moderator analyses did not support the need for extra time in order for glossaries and dictionaries to be effective. Specifically, the effect of English dictionaries and glossaries was not significantly different when bundled with extra time (Q between extra time and no extra time=0.001; p=.975 in the mixed effects model). That is, the average effect sizes for the five samples that provided extra time along with English dictionaries/glossaries was equal to that for the thirteen samples that did not provide extra time with this accommodation.

Extra time. Our results indicated that extra time, when provided alone, yielded a statistically significant effect on ELLs' performance (mean effect size=0.23; s.e.=0.10 in the fixed effects model). This mean effect size was based on three samples drawn from three different studies. When compared to the estimates in Table 3, this effect size can be considered equal to a 15 percent to 31 percent reduction in the observed achievement difference between ELLs and non-ELLs, depending on the specific estimate used. The test of heterogeneity indicated that the effect sizes did not differ across samples (Q=0.21; p=.900), albeit with a very small collection of samples. Given this result and the small number of samples, moderator analyses were not conducted.

We also examined the effect of bundling extra time with other accommodations by examining whether the average effect aggregated across accommodations differed by bundling with extra time. This analysis indicated that effects did not differ as a function of bundling with extra time (Q between extra time and no extra time=0.78; p=.377 in the random effects model), while mean effect sizes aggregated across accommodations had similar magnitudes when bundled with extra time (mean effect size=0.08; s.e.=0.06 in the random effects model) and when not bundled with extra time (mean effect size=0.15; s.e.=0.04 in the random effects model). We add that it is somewhat contrary to expectations that effect sizes would diminish when bundled with extra time. Overall, these results indicate that bundling extra time with other

accommodations does not appear to increase the effectiveness of those accommodations.

Accommodations with limited evidence of effectiveness

Native language accommodations. We found no significant mean effect for any of the native language accommodations, but also found that there continue to be few studies of any specific native language accommodation as well as great heterogeneity among the effects found. In particular, providing bilingual dictionaries and glossaries did not have a significant effect on ELLs' performance (mean effect size = -0.03; s.e. = 0.12 in the random effects model). This result is based on six samples from four studies, including one sample from one study not included in the Francis et al. (2006) meta-analysis. The test of heterogeneity indicated that the effect sizes differed across samples (Q = 13.97; p = .016).

Similarly, translated versions of tests (all studies involved translation into Spanish) did not have a statistically significant mean effect on ELLs' performance (mean effect size = 0.46; s.e. = 0.34 in the random effects model); this mean effect is based on five samples from two studies, including three samples from one study not included in the Francis et al. (2006) meta-analysis. The test of heterogeneity indicated that the effect sizes differed significantly across samples (Q = 55.47; p < .001).[3]

In addition, dual language booklets did not have a statistically significant effect on ELLs' performance (mean effect size=-0.01; s.e.=0.07 in the fixed effects model), based on five samples from three studies. The test of heterogeneity indicated that the effect of this accommodation did not differ across samples (Q=2.94; p=.568).

Similarly, dual language booklets read aloud did not have a significant effect on ELLs' performance, based on one sample (effect size=0.27; s.e.=0.20 in the fixed effects model). Given the small number of samples for any particular native language accommodation, moderator analyses were not conducted.

[3] One well-designed quasi-experimental study (Robinson, 2010) was not included in the current meta-analysis due to design and data reporting differences, but its results are nonetheless worth noting. Robinson (2010) used a regression discontinuity design to evaluate the effects of test translation for a math test in kindergarten and first grade, drawing on nationally-representative data from the Early Childhood Longitudinal Study—Kindergarten, 1998 cohort. Robinson found large effects (Cohen's *d*s > 0.85) for test translation for young Spanish-speaking ELLs and found little evidence that effects were moderated by the language of instruction. These large effects provide additional evidence for the heterogeneity of effects of native language accommodations; specifically, we suspect that such large effects are particular to the age group and specific early math skills investigated by Robinson (2010), in that math assessment in kindergarten and first grade will likely draw heavily on early numeracy skills (and related language skills) learned at home.

Read aloud. Reading tests aloud was not found to have a statistically significant effect on ELLs' performance (mean effect size=.09; s.e.=0.16 in the fixed effects model). This mean effect is based on two samples in two studies. Given the small number of samples, moderator analyses were not conducted.

Small group administration. Based on the effect reported in a single study, administering tests in small groups did not yield a statistically significant effect on ELLs' performance (mean effect size = -0.54; s.e. = 0.32 in the fixed effects model).

Effects of accommodations for non-ELLs

Consistent with Francis et al. (2006), none of the accommodations was found to significantly improve the performance of non-ELLs, providing supporting evidence for the conclusion that these accommodations do not alter the construct being measured for ELL students on these assessments. Forty-eight samples drawn from sixteen studies provided data to test the effects of accommodations on non-ELLs. Overall, the effect of accommodations on non-ELLs was very small and not statistically significant (mean effect size=0.01; s.e.=0.02 in the random effects model). The test for heterogeneity indicated that effect sizes differed across samples (Q=84.18; p=.001). When examined individually, none of the accommodations studied had significant positive effects for non-ELLs. Spanish translation had a significant negative effect for non-ELLs, based on one sample (effect size=-0.88; s.e.=0.25 in the fixed effects model), which is understandable given that most non-ELLs can be expected to lack proficiency in Spanish. Overall, this evidence suggests little reason to believe that the use of accommodations with ELLs threatens the validity of the assessments based on their providing benefits to non-ELLs, though of course the issue of validity is complex and involves many other facets (Pennock-Roman & Rivera, 2011).

CENTER ON
INSTRUCTION