Assessment Accommodations for English Language Learners: Implications for Policy-Based Empirical Research

Jamal Abedi, Carolyn Huie Hofstetter and Carol Lord

REVIEW OF EDUCATIONAL RESEARCH 2004 74: 1
DOI: 10.3102/00346543074001001

Review of Educational Research
Spring 2004, Vol. 74, No. 1, pp. 1–28

Assessment Accommodations for English Language Learners: Implications for Policy-Based Empirical Research

Jamal Abedi
University of California, Los Angeles
Carolyn Huie Hofstetter
University of California, Berkeley
Carol Lord
California State University, Long Beach

Increased attention to large-scale assessments, the growing number of English language learners in schools, and recent inclusionary policies have collectively made assessment accommodations a hotly debated issue, especially regarding the validity of test results for English language learners. Decisions about which accommodations to use, for whom, and under what conditions, are based on limited empirical evidence for their effectiveness and validity. Given the potential consequences of test results, it is important that policymakers and educators understand the empirical base underlying their use. This article reviews test accommodation strategies for English learners, derived from "scientifically based research." The results caution against a one-size-fits-all approach. The more promising approaches include modified English and customized dictionaries, which can be used for all students, not just English language learners.

KEYWORDS: accommodations, assessment, bilingual, English language learner (ELL), limited English proficient (LEP).

Historically, English language learners[1] in the United States were excluded from participation in large-scale student assessment programs; there were concerns about the confounding influences of language proficiency and academic achievement. In the last 40 years, however, a series of antidiscrimination laws, court cases, and, more recently, standards-based legislation, most notably the No Child Left Behind Act of 2001, have prompted marked changes in the education and assessment of students. States are responsible for developing challenging academic content and achievement standards, as well as statewide assessment systems for monitoring schools and districts to ensure that they are making adequate yearly progress toward educating all students to these high standards. The assessments, in addition to being technically sound and aligned with the state standards, must be valid and reliable as determined by "scientifically based research" and must meet various inclusion requirements, such as adaptations or accommodations for students with limited English proficiency (LEP) and students with disabilities.

The emphasis on inclusion introduces new, unintended consequences. We are not likely to obtain accurate and relevant information regarding a student's content knowledge of science, for example, by administering a science test in a language that the student does not understand. Accordingly, English learners are eligible for accommodations—changes in the test process, in the test itself, or in the test response format. The goal of accommodations is to provide a fair opportunity for English language learners to demonstrate what they know and can do, to level the playing field, so to speak, without giving them an advantage over students who do not receive the accommodation.

As of 1998–1999, thirty-seven states reported using test accommodations (Rivera, Stansfield, Scialdone, & Sharkey, 2000). The widespread use of accommodations, however, raises a number of issues and questions. Does using accommodations yield more valid inferences about an English learner's knowledge? Which students should be eligible and what criteria should be used to decide their eligibility? What type of accommodation should be used? Are some accommodations more effective than others—and if so, are they more effective in general or only for particular students? Do accommodations give students who receive them an unfair advantage? Is it meaningful to compare English learners' accommodated scores with English-proficient students' non-accommodated scores? What implications do test accommodations have for test administration and testing policy more generally? When we look for answers to these questions in studies of content area assessments, we are confronted with a striking lack of empirical research.

To address these issues, we focus our discussion accordingly:

1. What is the policy context for the use of test accommodations?
2. Who are English language learners?
3. What is the relationship between language proficiency and test performance?
4. What are accommodations, and who uses them?
5. How are accommodations used?
6. What does empirical research on accommodations tell us?
7. What are the key issues in deciding among accommodation options?
8. What are the implications for education policy and practice?

Much has been written about assessment for English language learners (ELL; see Sireci, Li, & Scarpati, 2003). For this review, we selected only studies that are related specifically to assessment accommodations for ELL students and that are based on an experimental design approach. Research using translations of tests into other languages, and the problems inherent in that approach, have been reviewed elsewhere and are discussed briefly here. Our research focus arises from the need for a thorough review of the validity of research on accommodations. Even if highly effective, an accommodation that alters the construct being measured may not produce desirable results because the accommodated and non-accommodated results may not be combined. The best way to examine the validity of an accommodation is to offer it to both ELL and non-ELL students in a randomized approach. This requirement excludes many studies based on existing data from national and state assessments. Typically, in those assessments accommodations are not provided to non-ELL students, and ELL students are not randomly assigned to various forms of accommodation. For example, Abedi, Leon, and Mirocha (2003) found that only LEP students at the lowest level of English proficiency were provided with accommodations.

What Is the Policy Context for the Use of Test Accommodations?

Decisions regarding the education, inclusion, and assessment of all students, regardless of gender, race, national origin, or language background, are founded on considerable historical, legal, and judicial precedent (for a detailed description see American Institutes for Research, 1999). Policies have emerged from various antidiscrimination laws (e.g., Title VI of the Civil Rights Act of 1964; the Equal Educational Opportunities Act of 1975), federal court cases (e.g., *Lau v. Nichols,* 1974; *Casteneda v. Pickard,* 1981), and standards-based legislation (e.g., Goals 2000; Titles I and VII of the Improving America's Schools Act of 1994; Titles I and III of the No Child Left Behind Act of 2001). The exclusion of certain groups from large-scale assessments may have been well-intentioned with regard to fairness and validity, but it has resulted in a lack of representation in broader educational policy and accountability debates and thus has diminished educational opportunities for English learners (August & Hakuta, 1997; Hakuta & Beatty, 2000).

To further student achievement, representation, and accountability in American schools, legislative reforms mandate that *all* children participate in large-scale statewide assessments. The notion of appropriate assessment accommodations was present in the Improving America's Schools Act and continues today. Title 1, Section 1111(b)(3)(C)(9)(III) of the No Child Left Behind Act further states that such assessments must provide for "the inclusion of limited English proficient students, who shall be assessed in a valid and reliable manner and provided reasonable accommodations on assessments, . . . including, to the extent practicable, assessments in the language and form most likely to yield accurate data on what such students know and can do in academic content areas." The only students exempt from state assessments are those who have not attended schools under the local educational agency for a full academic year.

Not surprisingly, the recent focus on testing has yielded considerable controversy. Proponents highlight the importance of attaining and maintaining standards; they regard tests as a primary mechanism for rewarding high-performing schools and helping and/or sanctioning low-performing schools. Critics declare that standardized assessments measure socioeconomic status, innate ability, and non–instructionally related material and thus yield little valid information about student achievement. Furthermore, the time commitment for test preparation robs classroom teachers of valuable instructional time and tends to water down instruction by indirectly encouraging teachers to "teach to the test" (Kohn, 2000; McNeil, 2000; Popham, 2001). This controversial, multilayered dialogue occurs in the public spotlight through daily newspapers (Lewin, 2002), popular national magazines (McGinn, Rhodes, Foote, & Gesalman, 1999), and radio reports (Sanchez, 2001, 2002a, 2002b).

Inclusion of English learners and the use of test accommodations with this population exacerbate these tensions. Validity concerns emerge from a potential "mainstream bias" (August & Hakuta, 1997; Garcia & Pearson, 1994); the use of a small, unrepresentative sample of English learners for test norming; and test content and procedures that reflect the dominant culture (Kopriva, 2000; Rivera & Vincent, 1997; Wong Fillmore, 1982). The question remains: How valid are inferences about students' knowledge when they are based on a test administered in a language that the student may not understand? Recognizing the importance of this issue and the growing presence of English learners in schools, a recent edition of *Standards for Educational and Psychological Testing* (American Educational

Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) devoted an entire section to testing language-minority students, cautioning that test results should be weighed carefully because of the potential for confounding influences on students' performance.

If questions about the validity of existing assessment tools for ELL students and the powerful impact of assessment on instruction result in reduced testing of ELL students, the quality of instruction for those students may be affected. ELL students may be omitted from the accountability picture; they may not be considered in state or federal policymaking; and their academic progress, skills, and needs may not be appropriately assessed. In view of the many issues concerning the inclusion of ELL students in national and state large-scale assessments, it is clear that excluding those students from assessments does more harm than good. Efforts must be made to modify assessment tools to make them more relevant to ELL students while not altering the construct being measured.

Who Are English Language Learners?

English language learners represent a rapidly growing, culturally and linguistically diverse student population in the United States. In 2000–2001, LEP students comprised nearly 4.6 million public school students.[2] The majority were Spanish speakers (79.0%), followed by Vietnamese (2.0%), Hmong (1.6%), Cantonese (1.0%), and Korean (1.0%). Since the 1990–1991 school year, the LEP population has grown by approximately 105%, while the overall school population has increased by only 12% (Kindler, 2002).

English learners matriculate in schools throughout the nation, but most frequently in large urban school districts in the Sun Belt states, in industrial states in the Northeast, and around the Great Lakes. This trend, however, is changing as immigrants move to more affordable suburban and rural areas and to areas where language-minority families are relative newcomers, such as the Midwest. More than half (56.1%) reside in four states alone: California (32.9%), Texas (12.4%), Florida (5.6%) and New York (5.2%) (Kindler, 2002). English learners represent one in four K–12 students in California schools (California Department of Education, 2000).

This population includes recent immigrants as well as children born in the United States. In the 2000–2001 school year, more than 44% of all LEP students were enrolled in Pre-K through Grade 3; about 35% were enrolled in Grades 4–8; and only 19% were enrolled at the high school level (Kindler, 2002). Many LEP students attend schools where most of their peers live in poverty. There are numerous differences among English learners; for example, Spanish-speaking families tend to have lower parental educational attainment and family incomes than Asian- or Pacific-language families (August & Hakuta, 1997).

Defining and identifying English learners remains problematic for educational practitioners and researchers. There have been problems with definitions or guidelines for identifying which students are "English language learners" (Anstrom, 1996; Rivera, Vincent, Hafner, & LaCelle-Peterson, 1997). Although the federal government provided definitions of LEP for purposes of funding allocations, specific operational guidelines are not available, allowing varying interpretations across school districts and states. Thus a student designated as LEP in one school district may not receive that designation in a neighboring district. The lack of consensus has prompted a call for states to adopt a common definition and to provide specific iden-

tification guidelines (August, Hakuta, & Pompa, 1994). To date, however, no such guidelines have been developed.

Many criteria are used across the nation for identification of ELL students. Among the most commonly used criteria are Home Language Survey results and scores from English proficiency tests. There are reasons, however, to believe that the Home Language Survey results may not be valid because of parents' concern over equity in education for their children, parents' citizenship issues, and communication problems (Abedi, 2004b). Similarly, there are concerns about the validity of current English proficiency tests, such as the Language Assessment Scales and other commonly used assessments (Zehler, Hopstock, Fleischman, & Greniuk, 1994). Criterion-related validity coefficients, or the correlation between English proficiency tests and other existing valid measure of English proficiency, are not strong, explaining less than 5% of the common variance (Abedi). Finally, in terms of content and construct validity, there is little evidence that the contents of existing English proficiency tests align sufficiently with commonly accepted English language proficiency standards, such as standards by Teachers of English to Speakers of Other Languages (Bailey & Butler, 2003).

What Is the Relationship Between Language Proficiency and Test Performance?

Research has documented amply the impact of students' language background on test performance. Students who lack proficiency in the language of the test consistently perform at lower levels, and changes in the language of the test can result in changes in student scores (Aiken, 1971, 1972; Cocking & Chipman, 1988; De Corte, Verschaffel, & De Win, 1985; Jerman & Rees, 1972; Kintsch & Greeno, 1985; Larsen, Parker, & Trenholme, 1978; Lepik, 1990; Mestre, 1988; Munro, 1979; Noonan, 1990; Orr, 1987; Rothman & Cohen, 1989; Spanos, Rhodes, Dale, & Crandall, 1988).

Researchers examining the content and technical quality of English proficiency tests have expressed concerns regarding the technical aspects of such tests, including the norming of the tests when they are used as normed referenced tests. For example, in reviewing technical characteristics of several commonly used English proficiency tests, Zehler et al. (1994) had concerns about the limited populations on which the test norms were based.

Similarly, reviewers of standardized achievement tests have expressed concerns about their use for ELL students because the norming groups for the tests are not representative of the ELL population. Linn (1995) cites the inclusion of all students as one of the three most notable features of current reform efforts. The issue of inclusion of students in assessment has also been among the major issues for the National Assessment of Educational Progress (NAEP; see, for example, Mazzeo, Carlson, Voelkl, & Lutkus, 2000). Navarrette and Gustke (1996) expressed concerns that assessments were "not including students from linguistically diverse backgrounds in the norming group, not considering the match or mismatch between a student's cultural and school experiences, and not ensuring for English proficiency," resulting in "justified accusations of bias and unfairness in testing" (p. 2).

Recent experimental design studies conducted by the National Center for Research on Evaluation, Standards, and Student Testing further demonstrated that (a) test scores of English learners are substantially lower than those of native Eng-

lish speakers in all subject areas, and (b) the linguistic complexity of test items may threaten the validity and reliability of tests of content area achievement, particularly for English learners (Abedi, 2002; Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000; Abedi, Leon, et al., 2003). These studies indicated that, as the language demands of individual test items decrease, the performance gap between English learners and English-proficient students decreases. Specifically, among students in Grades 10 and 11 at four selected sites in the United States, the English-proficient students performed higher on NAEP items than did their less English-proficient peers in reading and writing; the gap narrowed for less linguistically challenging items, such as science and math problems. The performance gap virtually disappeared in math computation, where test item language demands were minimal (Abedi, Leon, et al.).

These studies suggest that reducing the impact of language factors on content-based assessments can improve the validity and reliability of such assessments for English learners, resulting in fairer assessments. Furthermore, the findings suggest that various language-related accommodation strategies should be developed to minimize the impact of language and thus reduce the performance gap between English learners and other students.

What Are Accommodations, and Who Uses Them?

Definitions of test accommodations vary but tend to focus on their function as "support provided students for a given testing event either through modification of the test itself or through modification of the testing procedure to help students access the content in English and better demonstrate what they know" (Butler & Stevens, 1997, p. 5). Sometimes referred to as modifications or adaptations, such accommodations may be used to obtain a clearer picture of what students know and can do, especially with regard to content-based assessments (e.g., mathematics and science tests), where performance may be confounded with the students' English or native language proficiency or other background variables. The goal is to remove sources of difficulty that are irrelevant to the intent of the measurement, according to the Standards for Educational Accountability Systems (Baker, Linn, Herman, & Koretz, 2002), and thereby to level the playing field for English learners, without giving them an advantage over students who are not receiving accommodated assessments (Baker et al., 2002; Thurlow, Liu, Erickson, Spicuzza, & El Sawaf, 1996). Ideally, an assessment accommodation should yield an interaction effect, where the accommodation improves the performance of English language learners but not the performance of native English speakers. Such effects can be demonstrated through experimental design studies (Shepard, Taylor, & Betebenner, 1998).

Thirty-seven of 40 states (93%) with assessment accommodation policies for English learners in 1998–1999 allowed use of accommodations. Although there is little consistency in how they are administered, the most common accommodations have been extra testing time (allowed by 65% of the states on all or some test components), small group administration (59%), individual administration (53%), testing in a separate location or carrel or with more breaks (47%), and the use of a bilingual dictionary or word list (43%). Language-based changes are used less frequently; they have included translation of the test into the student's native language (22%), a bilingual version of the test (8%), and a modified (or "sheltered") English version (4%). Although the goal of accommodation for English learners is to

address their linguistic needs, language-based accommodations are permitted less frequently than other types. There have, however, been calls for closer examination of language-based accommodations, as they directly address the primary need and argument for accommodations (Rivera et al., 2000).

How Are Accommodations Used?

Accommodation use raises a number of thorny issues that are not easily resolved (Shepard et al., 1998). Should only certain students receive accommodations? Which students are eligible and who should make that decision? What type of accommodation should a student receive? Are accommodated test results comparable with non-accommodated test results? How do we ensure fairness so that accommodations enable English learners to demonstrate what they know without giving them an unintended advantage over other students? These questions are addressed below.

Deciding Who Is Eligible to Receive an Accommodation

Individual states have different eligibility criteria for determining which students receive accommodations and which do not. The most commonly reported criteria are formal assessments of English language proficiency (23%), time spent in the United States or in English-speaking schools (18%), informal assessments of English proficiency (14%), language program placement (11%), and performance on other tests (11%). Less common criteria include time in the state's schools (9%), performance on schoolwork (6%), teacher observations or recommendations (6%), parents' or guardian's opinions (3%), students' native language proficiency (3%), and academic background in the home language (3%). Some states use a single criterion, some use combinations of criteria (Rivera et al., 2000).

Further, the criteria themselves may be problematic. Butler and Stevens (1997) question the validity of formal language assessments, and numerous researchers criticize the use of standardized tests for English learners. Language of instruction is an important factor in determining the language for testing, but this factor is frequently ignored. Like many criteria for eligibility, language of instruction is not easily identifiable. Professional judgments are important in decisions about test accommodations, and they may vary with decision makers' levels of knowledge and previous experience in testing English learners. The usual decision makers are school or district officials (25%), the student's parents or guardians (20%), a local committee (18%), and the student's classroom teacher (14%) (Rivera et al., 2000).

Reporting Test Scores

Another issue is the potentially negative consequences of reporting accommodated scores. If a student receives an accommodation, should the test score be flagged as "accommodated" or "nonstandard" and the actual accommodation(s) listed as part of the student's academic record? Such information can stigmatize the student, not only within the classroom but also in the way school records are accessed and used to identify eligible students for programs. The potential repercussions of reporting test accommodations and effects on students' futures are unclear at this point.

Administering accommodations also introduces issues related to fairness and equity. It can be argued that accommodations should be provided for *all* students,

not just for select groups, so that no single group receives special advantages. For example, a study of student testing in Kentucky found that accommodations produced implausibly high mean scores for some groups of students with disabilities; excessive use of accommodations was cited (Koretz, 1997).[3] There have been reports of parents, hoping to improve their children's test performance, who tried to obtain special designations for their children so that they would be eligible for the accommodations provided for students with disabilities. Perceptions of favored treatment may contribute to a growing public backlash against the use of accommodations, in addition to the backlash against the expansion of large-scale testing programs (Rothstein, 2001). On the other hand, stakeholders may hold ideological or political views that some accommodations (e.g., extra time and modified English) are equitable and valid only for specified populations, not for all students. Legal and ethical issues have been raised with respect to testing and accommodations for students with disabilities; similar issues may be raised concerning English learners.

In summary, for any group of students, accommodations must be administered, used, and interpreted cautiously. This concern is reflected in *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999, p. 102):

> Judgment comes into play in determining whether a particular individual needs accommodation and the nature of such accommodation. . . . The overarching concern is the validity of the inference made from the scores on the modified test: fairness to all parties is best served by a decision about test modification that results in the most accurate measure possible of the construct of interest.

What Does Empirical Research on Accommodations Tell Us?

Since the standards-based reforms of the 1990s, the adoption of assessment accommodations in state- and district-level testing programs has far outpaced the research on their validity and their impact on testing programs. Many accommodation strategies have been proposed, and many used, with little knowledge of their actual effect. To date, only a handful of research studies exist. This section summarizes "scientifically based research" on test accommodations, meaning "research that involves the application of rigorous, systematic and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs" (No Child Left Behind Act of 2001, p. 1964). The selection of studies for review here was based on criteria for "scientifically based research" as outlined by the act:

i. Employs systematic, empirical methods that draw on observation or experiment;
ii. Involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn;
iii. Relies on measurements or observational methods that provide reliable and valid data across evaluators and observers, across multiple measurements and observations, and across studies by the same or different investigators;
iv. Is evaluated using experimental or quasi-experimental designs in which individuals, entities, programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interests,

with a preference for random assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition controls;

v. Ensures that experimental studies are presented in sufficient detail and clarity to allow for replication, or, at a minimum, offer the opportunity to build systematically on their findings;

vi. Has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review (pp. 1964–1965).

Our focus here is on research on accommodations for English learners. Although we do not review here the many studies of accommodations for students with disabilities, there has been extensive research in that area (see, for example, Mazzeo et al., 2000; Thurlow, McGrew, Tindal, Thompson, Ysseldyke, & Elliott, 2000; Thompson, Blount, & Thurlow, 2002; Tindal, Anderson, Helwig, Miller, & Glasgow, 2000; Tindal & Fuchs, 2000). Here we are interested in the performance of K–12 students in content area assessments (such as mathematics, science, and social studies) and, in particular, in accommodations that directly address the students' anticipated difficulty with the language of the written text. We first discuss assessments in the student's native language and then consider the use of modified English, extra time, dictionary, glossary, oral administration, and other accommodation approaches.

Tests in Students' Native Language

Upon initial consideration, it may seem reasonable simply to translate all content-knowledge tests into the student's native language. This approach has been tried, with numerous problems. In a background paper prepared for the National Assessment Governing Board regarding the accommodation of LEP students in the Voluntary National Test, research on test translation was reviewed (American Institutes for Research, 1999). The review identified validity problems with native language testing, largely because of the difficulty of maintaining construct equivalence when tests are translated.

Whether a test is translated directly from one language into another, or tests in two languages are developed in parallel, there is a high risk that the two versions will differ in content and construct (Kopriva, 2000). Even with efforts to devise ways to equate tests (Sireci, 1997) and the development of international guidelines for test translation and adaptation (Hambleton, 1994), translated assessments are technically difficult, time consuming, expensive to develop, and not widely used. A word or phrase in one language, for instance, may not have an equivalent in another language. Even if an equivalent word in the other language exists, an equivalent translation of the test item may be difficult to achieve; if the equivalent word is used less often in the other language, student performance may suffer (Hambleton & Patsula, 1998). In addition, some languages, such as Spanish, have multiple dialects, thus limiting the appropriateness of the translated version for some student populations (Olson & Goldstein, 1997).

Native-language assessments are useful only when students can demonstrate their content knowledge more effectively in their native language, typically because they have received content-area instruction in that language. Otherwise, translated items may confuse students who have learned content and concepts in English

(Butler & Stevens, 1997). Students who have learned subject-area-specific academic vocabulary in classes conducted in English may not be familiar with the corresponding academic vocabulary items in their native language. For those students, then, a test in the native language will contain unfamiliar vocabulary and will not provide an effective accommodation.

In a small-scale study (Liu, Anderson, Swierzbin, & Thurlow, 1999), nine Minnesota students in Grade 7 ESL/bilingual classes were offered a reading test with bilingual (English and Spanish) test items; six students reported preferring the bilingual version of the test for at least one of the two passages completed.

In an experimental study of Grade 8 mathematics performance with accommodations (Abedi, Lord, & Hofstetter, 1998), three different math test booklets with the same test items were prepared from the 1996 NAEP *Grade 8 Bilingual Mathematics* booklet: (a) original English (standard); (b) Spanish translation; and (c) modified English. (For more information on the NAEP booklet, please see Allen, Carlson, & Zelenak, 1999.) Test booklets were randomly assigned to students in intact mathematics classrooms. Hispanic English learners who received English language or sheltered English classroom instruction scored higher on the mathematics tests in English than their peers who were administered the same test in Spanish. In contrast, students who received mathematics instruction in Spanish performed significantly higher on the Spanish-language math items than their peers with the same items in English, either modified or original.

Other studies highlight the technical difficulties in developing and using translated assessments. In the NAEP LEP Special Study (Anderson, Jenkins, & Miller, 1996), participating students were instructed to respond to the math questions in their preferred language (Spanish or English) and were also given extra time. Item analyses conducted after the administration of the test suggested that the translated versions of the same items might not have been parallel to the original English versions in measurement properties. A large percentage of the Spanish items were found to have poor item statistics, dissimilar to those for the English versions. Item-level analyses suggested that English learners responded to about two-thirds of the items in the Spanish blocks in different ways. In addition, analysis of the bilingual test booklets suggested that up to 10% of English learners answered some items in English and some items in Spanish, rather than all of the items in one language, as instructed (Anderson et al., 1996).

Translation of test instructions, rather than test items, has also been used as a form of accommodation, but without conclusive results. A study of the performance of Grade 11 LEP students on the New Jersey High School Proficiency Assessment compared scores under three accommodations: translation of instructions, extra time, and a bilingual dictionary (Miller, Okum, Sinai, & Miller, 1999). Mean scores on the writing component were highest with translated instructions plus extra time, but were lowest with translated instructions without extra time on the writing and reading components.

Such research findings dispel the myth that testing English learners in their native language will in all cases yield more valid inferences about students' academic achievement. They instead suggest that assessments using languages other than English should be administered only to students who receive content-area instruction in that language and are familiar with the content terminology in that language, or students who, until recently, have been educated in that language.

Linguistic Modification of Test Items

A student who has knowledge of content in mathematics, science, or history is not likely to demonstrate that knowledge effectively if she cannot interpret the vocabulary and linguistic structures of the test. Accordingly, one approach to accommodation involves rewording of test items to minimize construct-irrelevant linguistic complexity. A number of studies have examined the language of mathematics problems and have established that making minor changes in the wording of a problem can affect student performance (Cummins, Kintsch, Reusser, & Weimer, 1988; DeCorte et al., 1985; Hudson, 1983; Riley, Greeno, & Heller, 1983). Differences in the syntactic complexity of the language of mathematics problems has been found to affect the performance of some students (Larsen et al., 1978; Wheeler & McNutt, 1983).

Recent studies have compared student scores on NAEP test items with parallel modified items in which the mathematics task and mathematics terminology are retained but the language has been modified. One study (Abedi & Lord, 2001) examined the effects of language modification of test items with 1,031 Grade 8 students in southern California. NAEP mathematics items were modified to reduce the complexity of sentence structures and to replace potentially unfamiliar vocabulary with words likely to be more familiar to the students. Mathematical terms were not changed. Original English versions of the items and modified English versions were randomly assigned to the students. The results showed small but significant differences in the scores of students in low- and average-level mathematics classes. Among the linguistic features that appeared to contribute to the differences were low-frequency vocabulary and passive voice verb constructions (see Abedi, Lord, & Plummer, 1997, for a discussion of the nature of, and rationale for, the modifications). English learners and low-performing students benefited the most from language modification of test items.

In another study, Abedi et al. (1998) examined the impact of language modification on the mathematics performance of English learners and English-proficient students. Items from the 1996 NAEP *Grade 8 Bilingual Mathematics* booklet were used to construct three different test booklets (original English, modified English, and original Spanish). The test booklets were distributed randomly to a convenience sample of 1,394 Grade 8 students in schools with high enrollments of Spanish speakers. Results showed that modification of the language of items contributed to improved performance on 49% of the items; the students generally scored higher on shorter problem statements.

A third study (Abedi, Lord, Hofstetter, et al., 2000) examined the impact of four forms of accommodation on a sample of 946 Grade 8 students, in comparison with no accommodation. The accommodations were (a) modified English, (b) extra time only, (c) glossary only, and (d) extra time plus glossary. These four accommodation forms, along with a standard test condition, were randomly assigned to the sampled students. The findings suggest that some accommodations increased the performance of both English learners and non–English learners. Among the various accommodations, only modified English narrowed the score gap between English learners and other students.

Other studies have also employed language modification; the results vary and may have depended in part on the nature and extent of the modifications made. In one study (Brown, 1999), 305 students in Grade 5 and 161 students in Grade 8 were

tested in mathematics and science with two test versions: one with items in original format, and another with modified English ("plain language"). No significant differences were found for any group of students (LEP, special education, or regular) in either subject or grade level. Brown compared the performance of LEP, special education, and regular students in science and math using both multiple-choice and open-ended test items. The results of this study were different for math and science. In science, there was no significant difference between the performance of students on the original test and the plain language version. However, on open-ended science questions, students at the higher end of the score distribution benefited from the plain language version. In math the study did not find a systematic pattern. On some of the test items students performed better with the plain language version, but on other items they did better with the original version. Brown found little difference in student performance on the original and plain language items but noted that the lack of significant difference could have been be due to small sample size.

Another study (Lotherington-Woloszyn, 1993) assessed reading comprehension of original texts and "simplified" versions for 36 intermediate-proficiency learners of English as a second language at the university-entrance level. Participants read, recalled, and commented on original texts and two versions simplified by different editors. The results suggested that none of the text versions was significantly better comprehended. However, participants could identify the simplified versions and recognize the original texts as being the most difficult to comprehend. Lotherington-Woloszyn concluded that although simplifying texts did not affect comprehension, it did affect participants' attitudes toward the text.

Rivera and Stansfield (2001) compared English learner performance on regular and simplified science items for Grades 4 and 6. All Delaware students in those grades participated in the study. Six forms of science assessment were used, four containing the original English and two consisting of the simplified versions of the same test items. A large number of regular students participated in the study (9,000 students in each grade level), but the number of LEP students was small. Because the LEP students were distributed across the six forms, their small number became even smaller (6 to 23 students per form). Although the small sample did not show significant differences in scores, the study demonstrated that linguistic simplification did not affect the scores of English-proficient students, indicating that linguistic simplification was not a threat to score comparability.

Extra Time

The provision of extra time is the most commonly permitted accommodation. It allows students more time to complete test items than is normally allotted. It is logistically feasible and does not require changes in the test itself. Some teachers feel that extra time is essential for English learners to decode the language of the test until they become fully proficient in academic English. Research to date on extra time as a valid accommodation strategy is not conclusive. Abedi, Lord, Hofstetter, et al. (2000) found that English learners benefited from extra time; however, English-proficient students benefited from extra time as well. Raising student scores overall may be a welcome result, even though the accommodation did not narrow the performance gap in this study.

Although extra time is a popular and easily implemented accommodation strategy (Hafner, 2001; Kopriva, 2000), it may not always be effective. Miller et al.

(1999) studied Grade 11 English learners using three accommodations: extra time, translation of instructions, and bilingual dictionary. The results were inconclusive, as students had the highest mathematics scores under standard testing conditions and the lowest under extra time.

Published Dictionaries

Students with limited English vocabularies may benefit from the accessibility of word definitions, and such definitions may be provided in the form of commercial dictionaries. A study of 133 Hmong students with limited English proficiency and 69 English-proficient students in three urban middle schools in metropolitan Minneapolis involved tests with four reading passages: two with a commercially published, monolingual, simplified English dictionary, and two without the dictionary. There was not a significant difference in reading comprehension scores for either group. However, among students who reported using the dictionary, those with self-reported intermediate English reading proficiency benefited; self-reported poor readers did not (Albus, Bielinski, Thurlow, & Liu, 2001).

The dictionary strategy provides an accommodation in a familiar format; students are probably already accustomed to using dictionaries. However, commercially available dictionaries differ widely in the difficulty level of the vocabulary in the definitions (Kopriva, 2000). Abedi, Courtney, Mirocha, Leon, and Goldberg (2001) used both published English and bilingual dictionaries as a form of accommodation for 611 students in Grades 4 and 8 in several locations in various regions of the United States. Both ELL and non-ELL students were tested in science. The authors found that the use of published dictionaries was not effective as an accommodation and was administratively difficult to implement. In addition, they questioned the validity of using commercial dictionaries, arguing that such dictionaries may provide information that the test is measuring. In addition, English dictionaries usually provide substantially broader levels of content coverage than the bilingual dictionaries, and some bilingual dictionaries provide more content coverage than others. As a result of their validity and feasibility concerns, Abedi et al. recommended against using published dictionaries.

A bilingual dictionary was one of the accommodations provided by Miller et al. (1999) for Grade 11 students (discussed earlier). They found that the students who received that accommodation scored lowest on the mathematics test.

Glossary and Customized Dictionary

An accommodation format might include brief glosses (definitions or simple paraphrases of potentially unfamiliar or difficult words) in the test booklet—for example, in the page margin beside the test item. One study providing this form of accommodation (Abedi, Lord, Hofstetter, et al., 2000) found that both English learners and English-proficient students performed significantly higher on mathematics test items when extra time was provided along with the glossary. For English learners, the gain was small but significant (a mean score of 13.69 out of a maximum score of 35 with glossary and extra time, as compared with a mean score of 12.07 without them). Provision of extra time only, or the glossary only, had less impact; in fact, for English learners, provision of only the glossary resulted in slightly lower scores, possibly a consequence of information overload because the students had limited time to access the additional information.

Another study of 422 Grade 8 students in science classes (Abedi, Lord, Kim, & Miyoshi, 2000) compared performance on NAEP science items in three test formats: one booklet in original format (no accommodation); one booklet with English glosses and Spanish translations of selected words in the margins; and one booklet with a customized English dictionary at the end of the test booklet. The customized dictionary included only words that appeared in the test items. The three test booklets were randomly assigned to sampled students in intact classrooms. English learners scored highest on the customized dictionary accommodation. Although the accommodations helped English learners to score higher, for the English-proficient students there was no significant difference among scores in the three test formats. This outcome suggests that the accommodation strategies did not affect the construct.

Another study of 607 students in Grade 4 and 542 students in Grade 8 (Abedi, Courtney, & Leon, 2003) compared performance on NAEP science items with two types of accommodation randomly assigned: a computer-based glossary (with a computer-based text) and a customized dictionary (in hard copy). The computer-based test, which included extra time, came with a customized "pop-up" glossary that appeared as students pointed at non-content words with a computer mouse. The results indicated that the computer-based testing accommodation was effective for ELL students without posing any threat to the validity of the assessment.

Oral Administration

The category of oral administration includes several types of accommodation. For example, the test directions and/or the test items themselves may be administered orally (rereading and paraphrasing of directions may be permitted if test security is not an issue). In addition, oral administration may be in the students' native language or in English. With these accommodations there is a risk that the test administrator will provide unintentional cues—for example, through voice, rate of reading, or body language—thus giving students more information than would normally be apparent from the printed words. Preliminary studies (Kopriva & Lowrey, 1994) suggest that students prefer oral administration of the assessment in their native language when they are new to the United States, are not literate in their home language, and have little oral or reading proficiency in English. In addition, the studies suggest that oral administration of the test in English is preferred when students have been instructed in English in the United States for a long period of time and have attained a level of conversational oral proficiency in English but are not yet literate enough in English to read the test.

Other Accommodations

Several other strategies have been proposed. Additional breaks in the test period have been suggested to counteract fatigue. Students may be tested in small groups or individual carrels to minimize distractions. Less common accommodation strategies include test forms for which the student responses do not require writing, written responses can be in the student's native language, or oral responses can be in English or the student's native language.

What Are the Key Issues in Deciding Among Accommodation Options?

In deciding whether and when to use accommodations for English language learners, four major considerations emerge:

- *Effectiveness:* What strategies minimize language proficiency effects and enable English learners to demonstrate their content area knowledge?
- *Validity:* Does provision of the accommodation help English learners by reducing the language barrier, thus narrowing the performance gap between English learners and English-proficient students? Or does it affect the content (e.g., affecting scores regardless of the students' language proficiency), possibly altering the test construct being measured?
- *Differential impact:* Do students' background variables affect the accommodated results? Are some accommodations more effective with certain groups of students than with others?
- *Feasibility:* Is the accommodation strategy practical and affordable, even for large-scale assessments?

Effectiveness

To be effective, the accommodation should improve the performance of English learners by helping them to overcome the language barrier. In other words, it should level the playing field. As discussed above, empirical studies have found that using parallel items in modified English has enabled English learners to score higher and has reduced the performance gap between English learners and other students (Abedi & Lord, 2001; Abedi, Lord, Hofstetter, et al., 2000).

Validity

A valid accommodation should help the target group but not affect scores of other students. "If assessment accommodations are working as intended, the result should show an interaction effect. The accommodation should improve performance of English-language learners but should leave the performance of native-English speakers unchanged" (Shepard et al., 1998, p. 11). As noted above, for example, Abedi, Lord, Hofstetter, et al. (2000) found that, although provision of a glossary plus extra time increased the performance of English learners by 13%, it increased the performance of English proficient students even more, by 16%. If an accommodation strategy intended for English learners raises everyone's scores, the strategy may not be addressing the intended target, the difference in language proficiency. In such a case, it is possible that the accommodation is affecting the construct being measured. The result brings into question the validity of the test results. As noted above, many states currently use accommodations without evidence of their validity. If an accommodation improves the performance of all students, then the accommodated assessment may not be valid; if it is provided for selected students only, the resulting scores should not be combined with non-accommodated results.

Differential Impact

An accommodation strategy may be effective with some students but not with others. Empirical research using students' background characteristics can seek to identify which accommodation approaches are most appropriate. In other words, there may be interaction effects between the various types of accommodations and students' background characteristics. In a study using Grade 8 NAEP mathematics test items, Abedi, Lord, Hofstetter, et al. (2000) selected several background variables for grouping students: type of mathematics class, language of instruction in

mathematics class, country of origin, and length of time in the United States. Two multiple regression models were created that used a criterion scaling approach (see Pedhazur, 1997). The mathematics test score was the criterion variable in the two models. The results suggest that students' background variables—especially those related to language, such as length of time in the United States, overall grades since Grade 6, and number of times the student has changed schools—could be useful in deciding which accommodation to use. Thus caution is warranted against blanket statements about the general effectiveness or lack of effectiveness of a particular form of accommodation for all English language learners.

Feasibility

An accommodation is not likely to be implemented if it is not logistically manageable or affordable. For national and state assessments there are substantial costs associated with providing dictionaries or glossaries (whether to all or to selected students) and with administering assessments to students one-on-one or in small groups. Translation of assessment instruments into the many languages and dialects found in the English-learner population may pose feasibility problems because of technical issues concerning translation (as noted earlier), as well as budget and resource limitations. Some forms of accommodation require initial expenditures for test design and production but do not require additional costs for administration; accommodations of this type include the use of parallel items in modified English or provision of a modified glossary incorporated as part of the test booklet. The costs of accommodations should be tracked and evaluated, and cost–benefit analyses should be considered.

Interaction of Effectiveness, Validity, Differential Impact, and Feasibility

The four issues discussed above must be considered in combination and as an interactive set when an accommodation strategy is recommended. An ideal accommodation must be effective, valid, and appropriate to the background of recipients, while at the same time feasible. For example, an accommodation that is effective in raising English language learner performance may not be usable if it also increases the performance of non–English learners, because it may change the construct being measured. An effective and valid accommodation for a particular group of English learners may not be effective and valid for other groups of English learners. Furthermore, accommodations that are effective, valid, and appropriate to the background of students may not be feasible to implement because of budgetary or logistical considerations. For ease of administration, the simplest arrangement is to provide the same test procedure and format for all students. Accordingly, the preferred accommodation is one that (a) can feasibly be provided for all students; (b) addresses the language issue for English learners and raises their performance; and (c) does not affect the scores of other students who are already proficient in English. The intent of an accommodation is to "remove irrelevant sources of difficulty, to get a fairer picture or more accurate picture of what the test-taker actually knows" (Shepard et al., 1998, p. 3). Decisions regarding the selection of particular accommodation strategies should therefore be based on their effectiveness and validity for the group being assessed.

The limited number of empirical studies to date show that these three criteria have been met by (a) providing items with controlled vocabulary and sentence struc-

tures (i.e., the "modified English" studies); and (b) providing customized glossaries (i.e., including in the test booklet brief definitions of the words in the test).

What Are the Implications for Education Policy and Practice?

Evidence from a review of the empirical research strongly suggests caution in adopting a "one size fits all" approach to test accommodations for English learners. As discussed earlier under the heading "Differential Impact," the effectiveness and validity of accommodations depend on the backgrounds of students for whom the accommodations are used. Some forms of accommodation are more appropriate for some students than for others. Students classified as LEP are not a homogeneous group. LEP students differ in many respects, including level of English proficiency (see, for example, Abedi, 2004b). The level of English proficiency of some students currently classified as LEP may be even higher than that of some low-performing native English speakers. Accommodations that are appropriate for LEP students at the higher end of the English proficiency distribution may not be relevant for LEP students at the lower end of the distribution. For example, LEP students with English proficiency sufficient to understand test questions may need additional time to process the information and may therefore make good use of an extra time accommodation. In contrast, students at the lower end of the English proficiency distribution may not understand the English vocabulary and structures in the test items and therefore not benefit from extra time. For them, a glossary of non-content terms or a customized dictionary may be more appropriate.

On the basis of previous research, we believe that the following can be said with confidence at this time:

1. Translating test items from English into other languages does not appear to be an effective accommodation strategy when students have studied the subject in a classroom where English is used. The language of assessment should match students' primary language of instruction.
2. Some accommodations are more effective with certain student groups than with others, depending on background factors such as English reading proficiency and length of time in the United States.
3. The performance gap between English learners and other students has been narrowed by modifying the language of the test items to reduce the use of low-frequency vocabulary and complex language structures that are incidental to the content knowledge being assessed. This accommodation strategy is effective; it is also valid, because it does not appear to affect the performance of English-proficient students.
4. Customized dictionaries can be an effective and valid alternative to commercial dictionaries; they have been found to help English learners while not affecting the scores of English-proficient students.

There is no reason why *all* students should not have content-area assessments that use clear language and provide sufficient time for them to show what they know. In addition, customized dictionaries or glossaries can be provided for all students, regardless of their level of English language proficiency.

One of the most promising test accommodations—modifying the language but not the content of the test item, which can reduce the gap between English learners and others by affecting the performance of only the former—is rarely used. Native

language translations are typically available only for Spanish-speaking students, because Spanish is the most common native language among English learners in the United States. Because it is not cost-effective to develop translated assessments in the many languages spoken by other English learners, this introduces questions of fairness and equity for non-Hispanic students.

One question that we have encountered is whether, in reducing the language barrier through accommodations, we are "dumbing down" our tests and lowering our expectations—for example, by using modified language or using oral presentation or response. We want all of our students to develop proficiency in English, and we need to provide them with the language experience and the tools to understand and use the academic language of our classrooms, in school books, materials, and tests. However, for assessments in content areas such as mathematics, science, and social studies, it is reasonable to minimize the use of language structures that are unnecessarily complicated and are not relevant for the knowledge or skills being measured.

From this survey of English learner accommodations and the limited research to date, the following recommendations emerge:

1. The design of future large-scale content area assessments should take English learners into account from the outset rather than as an afterthought. Test developers should use straightforward, uncomplicated language when *developing* test forms, building equitable assessments initially, rather than trying to avoid language biases by making adjustments later in the testing process. The use of clear language free of unnecessary complexity and the provision of a customized dictionary can be part of good assessment practice, not a separate adaptation. The evidence from empirical research suggests that providing these for all students does not threaten test validity.
2. The specific language demands of academic materials and assessments should be identified and provided to teachers so that they can ensure that students have the language resources to demonstrate their content-area knowledge and skills. Ensuring the opportunity to learn is essential for meaningful educational testing. Differential performance on original and modified English items can help researchers identify linguistic features that may affect the performance of some student groups.
3. Cost–benefit analyses should be conducted to compare the relative advantages and feasibility of accommodation alternatives. Some accommodation strategies may be more expensive—in various senses of the word—than others; at present, the research data on costs of accommodations are limited.
4. More research should examine the effectiveness and validity of particular accommodations for various student groups. Student background variables are strong indicators of preparedness for participation in large-scale assessments. Accordingly, in planning who will be tested and how, states should collect and consider background information, including the language spoken in the home, English proficiency level, length of time in the United States, and years of schooling in English and the native language.
5. The research base should be expanded to test and confirm the limited results reported here. The body of empirical research on accommodations for K–12 English learners in content areas is meager. Thus educational practitioners are forced to rely on anecdotal evidence and perceived notions of "best practice,"

rather than on empirical evidence of validity, to guide their decisions on the use of accommodations. New and innovative assessment techniques should be developed and empirically tested to provide approaches with proven effectiveness and validity for all of our students, including English learners.

Increased attention to assessment accommodations for ELL students is warranted as a result of the recent standards reform movement, exemplified in legislation such as the Improving America's Schools Act of 1994 and the No Child Left Behind Act of 2001. These reforms have put instruction and assessment of less advantaged subgroups of students, including ELL students, at the top of the national agenda. The legislation was created, in part, in response to the rapid increase in the number of English learners in our schools, and it has mandated inclusion of those students in national and state assessments through the use of reliable and valid measures (Abedi, 2004b; Erpenbach, Forte-Fast, & Potts, 2003; No Child Left Behind Act of 2001; Mazzeo et al., 2000). Because of equity issues in the assessment of ELL students (Abedi, 2004a), accommodations have been provided to facilitate assessments of those students and make them more equitable across LEP and regular student categories. Decision makers are encouraged to consider the full range of issues reviewed here, including validity and effectiveness, in selecting and implementing accommodations for English language learners.

Notes

The research for this article was supported in part by the Office of Educational Research and Improvement, Contract #R305B960002-01 as administered by the U.S. Department of Education. The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the Office of Educational Research and Improvement or the U.S. Department of Education. The authors acknowledge valuable contributions of colleagues Frances Butler and Joan Herman and are grateful to Eva Baker for support of this work. The three principal authors are listed alphabetically. Correspondence concerning this article should be addressed to Jamal Abedi (see author contact information after reference list).

[1] This article uses the terms "English language learner" (ELL), "English learner," and "student with limited English proficiency" (LEP student) interchangeably. All refer to students who may be in need of English language instruction. Although "LEP" is a commonly recognized term for this student population, some regard it as having a negative connotation and prefer the more positive terms "English language learner" and "English learner" (LaCelle-Peterson & Rivera, 1994; August & Hakuta, 1997; Butler & Stevens, 1997).

[2] Estimates of the English learner population vary with definitions of LEP and the estimation techniques used.

[3] Allowance of multiple accommodation strategies is more commonplace in assessments of students with disabilities. English language learners typically are given only a single accommodation.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*(3), 231–257.

Abedi, J. (2004a). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4–14.

Abedi, J. (2004b). *The validity of the classification system for students with limited English proficiency: A criterion-related approach.* Manuscript submitted for publication.

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

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2001). *Language accommodation for large-scale assessments in science.* 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 Tech. Rep. No. 603). Los Angeles: University of California, 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., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.

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

Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance: NAEP TRP Task 3D: Language background study* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Aiken, L. R. (1971). Verbal factors and mathematics learning: A review of research. *Journal for Research in Mathematics Education, 2,* 304–313.

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Educational Research, 42*(3), 359–385.

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

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 Technical Report* (NCES Publication No. 1999452). Washington, DC: National Center for Education Statistics.

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

American Institutes for Research. (1999, February). *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, 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.

Anstrom, K. (1996). Defining the limited-English-proficient student population. *Directions in Language and Education: National Clearinghouse of Bilingual Education, 1*(9), 1–9.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda.* Washington, DC: National Academy Press.

August, D., Hakuta, K., & Pompa, D. (1994). *For all students: Limited English proficient students and Goals 2000* (Occasional Papers in Bilingual Education Focus No. 10). Washington, DC: National Clearinghouse for Bilingual Education.

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K–12 education: A design document* (CSE Tech. Rep. No. 611). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* (Policy Brief 5, Winter). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing. (To appear in S. Fuhrman & R. Elmore, Eds., *Redesigning accountability systems,* Teachers College Press, New York.)

Brown, P. B. (1999). *Findings of the 1999 Plain Language Field Test* (Publication T-99-013.1). Newark, DE: Delaware Education Research and Development Center, Inclusive Comprehensive Assessment Systems Project.

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

California Department of Education. (2000). *California demographics data.* Retrieved February 24, 2003, from *http://www.cde.ca.gov/demographics/*

Castenada v. Pickard, 648 F.2d 989 (5th Cir. 1981).

Civil Rights Act of 1964, Pub. L. No. 88-352 (1964).

Cocking, R. R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement of language minority children. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics,* pp. 17–46. Hillsdale, NJ: Erlbaum.

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

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

Equal Educational Opportunities Act of 1975, 20 USC Sec. 1703 (1975).

Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB* (An Accountability Systems and Reporting State Collaborative on Assessment and Student Standards Paper). Washington, DC: Council of Chief State School Officers.

Garcia, G. E., & Pearson, P. D. (1994). Assessment and diversity. *Review of Research in Education, 20,* 337–391.

Goals 2000: Educate America Act, Pub. L. No. 103-227 (1994).

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.

Hakuta, K., & Beatty, A. (Eds.). (2000). *Testing English-language learners in U.S. schools.* Washington, DC: National Academy Press.

Hambleton, R. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*(3), 229–244.

Hambleton, R., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45,* 153–171.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development, 54,* 84–90.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. (1994).

Jerman, M., & Rees, R. (1972). Predicting the relative difficulty of verbal arithmetic problems. *Educational Studies in Mathematics, 4,* 306–323.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services, 2000–2001 summary report.* Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92*(1), 109–129.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools.* Portsmouth, NH: Heinemann.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners.* Washington, DC: Council of Chief State School Officers.

Kopriva, R. J., & Lowrey, K. (1994, April). *Investigation of language-sensitive modifications in a pilot study of CLAS, the California Learning Assessment System* (Tech. Rep.). Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Tech. Rep. No. 431). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55–75.

Larsen, S. C., Parker, R. M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. *Educational Studies in Mathematics, 21,* 83–90.

Lau v. Nichols, 414 U.S. 563, 94 S.Ct. 786 (1974).

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics, 21,* 83–90.

Lewin, T. (2002, March 18). In testing, one size may not fit all. *New York Times,* p. A16.

Linn, R. L. (1995). *Assessment-based reform: Challenges to educational measurement.* Princeton, NJ: Educational Testing Service.

Liu, K., Anderson, M., Swierzbin, B., & Thurlow, M. (1999). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 1* (Minnesota Rep. No. 20). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Lotherington-Woloszyn, H. (1993). Do simplified texts simplify language comprehension for ESL learners? In Tickoo, M. L. (Ed.), *Simplification: Theory and application* (Anthology Series 31). Singapore: SEAMEO Regional Language Centre.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES Publication No. 2000-473). Washington, DC: National Center for Education Statistics.

McGinn, D., Rhodes, S., Foote, D., & Gesalman, A. (1999, September 6). The big score: High-stakes tests are rapidly becoming a rite of passage in districts around the country. But do they really improve learning? *Newsweek,* 46–51.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing.* New York: Routledge.

Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 200–220). Hillsdale, NJ: Erlbaum.

Miller, E. R., Okum, I., 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.

Munro, J. (1979). Language abilities and math performance. *Reading Teacher, 32*(8), 900–915.

Navarrette, C., & Gustke, C. (1996). *A guide to performance assessment for linguistically diverse students.* Albuquerque: New Mexico Highlands University.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. (2002).

Noonan, J. (1990). Readability problems presented by mathematics text. *Early Child Development and Care, 54,* 57–81.

Olson, J. F., & Goldstein, A. A. (1997). *The inclusion of students with disabilities and limited English proficiency students in large-scale assessments: A summary of recent progress.* (NCES Publication No. 97-482). Washington, DC: National Center for Education Statistics.

Orr, E. W. (1987). *Twice as less: Black English and the performance of Black students in mathematics and science.* New York: W. W. Norton.

Pedhazur, E. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Harcourt Brace College Publishers.

Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). New York: Academic Press.

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

Rivera, C., Stansfield, C. W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998–1999.* Arlington, VA: George Washington University, Center for Equity and Excellence in Education.

Rivera, C., & Vincent, C. (1997). High school graduation testing: Policies and practices in the assessment of English language learners. *Educational Assessment, 4*(4), 335–355.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). *Statewide assessment program policies and practices for the inclusion of limited English proficient students.* Washington, DC: Clearinghouse on Assessment and Evaluation. (ERIC Rep. No. EDO-TM-97-02)

Rothman, R. W., & Cohen, J. (1989). The language of math needs to be taught. *Academic Therapy, 25*(2), 133–142.

Rothstein, R. (2001, May 30). Lessons: A rebellion is growing against required tests. *New York Times,* p. B9.

Sanchez, C. (Speaker). (2001, March 29). Testing backlash. *All Things Considered* [Radio broadcast]. Washington, DC: National Public Radio.

Sanchez, C. (Speaker). (2002a, February 14). Lost students. *Morning Edition.* [Radio broadcast]. Washington, DC: National Public Radio.

Sanchez, C. (Speaker). (2002b, March 21). Standardized tests: What are we trying to measure? *Talk of the Nation.* [Radio broadcast]. Washington, DC: National Public Radio.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's Grade 4 Mathematics Performance Assessment* (CSE Tech.

Rep. No. 486). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*(1), 12–19.
Sireci, S. G., Li, S. & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature* (Center for Educational Assessment Research Rep. No. 485). Amherst: University of Massachusetts.
Spanos, G., Rhodes, N. C., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 221–240). Hillsdale, NJ: Erlbaum.
Thompson, S., Blount, A., Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Tech. Rep. 34). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
Thurlow, M., Liu, K., Erickson, R., Spicuzza, R., & El Sawaf, H. (1996, August). *Accommodations for students with limited English proficiency: Analyses of guidelines from states with graduation exams* (Minnesota Rep. 6). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
Thurlow, M. L., McGrew, S., Tindal, G., Thompson, S. J., Ysseldyke, J. E., & Elliott, J. L. (2000). *Assessment accommodations research: Considerations for design and analysis* (NCEO Tech. Rep. 26). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
Tindal, G., Anderson, L., Helwig, R., Miller, S., & Glasgow, A. (2000). *Accommodating students with learning disabilities on math tests using language simplification.* Eugene, OR: University of Oregon, Behavioral Research and Teaching.
Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: University of Kentucky, Mid-South Regional Resource Center.
Wheeler, L. J., & McNutt, G. (1983). The effect of syntax on low-achieving students' abilities to solve mathematical word problems. *Journal of Special Education, 17*(3), 309–315.
Wong Fillmore, L. (1982). Language minority students and school participation: What kind of English is needed? *Journal of Education, 164,* 143–156.
Zehler, A. M., Hopstock, P. J., Fleischman, H. L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students.* Arlington, VA: Development Associates, Special Issues Analysis Center.

Authors

JAMAL ABEDI is the Director of Technical Projects at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) and a faculty member at the University of California, Los Angeles, Graduate School of Education and Information Studies, 300 Charles E. Young Drive North, Los Angeles, CA 90095-1522; e-mail *jabedi@cse.ucla.edu.* His recent studies have focused on the impact of linguistic factors and accommodations for English language learners.
CAROLYN HUIE HOFSTETTER is an Assistant Professor at the University of California, Berkeley, Graduate School of Education, 4431 Tolman Hall, Berkeley, CA 94720-1670; e-mail *chofstet@uclink.berkeley.edu.* Her research focuses on assessment methodologies and evaluation theory and practice.
CAROL LORD is an Associate Professor at California State University, Long Beach, with a joint appointment in the Department of Teacher Education and the Department of Linguistics, where she teaches courses on linguistic structure and development and language testing. She can be reached at California State University, Long Beach, 1250 Bellflower Blvd., Long Beach, CA 90840; e-mail *clord@csulb.edu.* Her research interests include characterization of linguistic complexity and sources of reading comprehension difficulty.

APPENDIX

Summary of empirical evidence related to testing accommodations for English language learners

Study	Characteristics	Outcomes
	Testing in students' native language or in English with translation of test items or directions	
Abedi, Lord, & Hofstetter (1998)	Experimental study (n = 1,394) in southern California. Three booklets containing 35 NAEP mathematics items for Grade 8 were randomly administered to English learners and non–English learners in intact classrooms: (a) original English, (b) modified English, and (c) Spanish translation.	Hispanic English learners in mathematics classes with English or sheltered English instruction scored higher on NAEP mathematics test in English (standard or linguistically modified) than did peers who were administered the test in Spanish. In contrast, English learners receiving mathematics instruction in Spanish performed significantly higher on Spanish-language mathematics items than did students with same items in English.
Anderson, Jenkins, & Miller (1996)	Analysis of data from the NAEP LEP Special Study. Students were instructed to respond to mathematics questions in their preferred language (Spanish or English) and were given extra time.	The results suggest that translated versions of items may not have been parallel to the original English versions in measurement properties. A large percentage of Spanish items were found to have poor item statistics, dissimilar to those for the English versions. English learners responded to two-thirds of the items in Spanish blocks in different ways. In addition, as many as 10% of English learners answered some items in English and some in Spanish, rather than all items in one language only, as instructed.
Miller, Okum, Sinai, & Miller (1999)	Study of LEP accommodation for 601 Grade 8 students in 17 districts (10 counties) in New Jersey. The accommodations included (a) extra time, (b) English/native-language dictionary, and (c) both extra time and dictionary.	The results were inconclusive. The mean scores on the writing component were highest with translated instructions plus extra time, but lowest with translated instructions without extra time on the writing and reading components.

(*continued*)

APPENDIX *(continued)*

Study	Characteristics	Outcomes
Miller, Okum, Sinai, & Miller (1999)	Study of Grade 11 LEP students' performance on the New Jersey High School Proficiency Assessment under three accommodations: (a) translation of instructions, (b) extra time, and (c) bilingual dictionary.	All LEP students, with or without accommodations, obtained scores below the minimum score of 100 for proficiency in each content area (reading, math, and writing). Overall, under standard conditions with translated instructions, students received the lowest mean scores for reading and writing. However, students under the standard testing conditions received the highest mean scores for the math component of the Early Warning Test.
	Testing in modified English	
Abedi & Lord (2001)	Experimental study comparing performance on original NAEP math items and parallel linguistically modified items by Grade 8 students ($n = 1{,}031$) in southern California. Original English versions of the items and modified English versions were randomly assigned to students.	English learners and low-performing students benefited most from language modification of test items. There were small but significant differences in the scores of students in low- and average-level mathematics classes. Linguistic features that appeared to contribute to the differences were low-frequency vocabulary and passive-voice verb constructions.
Abedi, Lord, & Hofstetter (1998)	Experimental study ($n = 1{,}394$) in southern California. Three booklets containing 35 NAEP mathematics items for Grade 8 were randomly administered to English learners and non–English learners in intact classrooms: (a) original English, (b) modified English, and (c) Spanish translation.	The results indicated that linguistic modification of items contributed to improved performance by English learners on 49% of the items. Students generally scored higher on shorter problem statements.
Abedi, Lord, Hofstetter, & Baker (2000)	Experimental study using Grade 8 NAEP items with several accommodations, including extra time and glossary, assigned randomly to Grade 8 students ($n = 946$).	The results indicated that some forms of accommodation increased performance by both ELL and non-ELL students. Among the accommodation options, only modified English narrowed the score gap between English learners and non–English learners.

Brown (1999)	Study to examine the effect of plain language on assessment of three groups of students—Students with Disabilities, LEP, and regular—using multiple-choice and open-ended science and math items.	Higher-performing students benefited from the linguistically modified version of the open-ended questions. No significant impact of language modification was found for math items.
Lotherington-Woloszyn (1993)	Assessment of reading comprehension of original texts and two simplified versions by 36 ESL learners at the university entrance level.	No test version produced significantly better comprehension than any other. However, participants were able to identify which texts were simplified and could rank the original texts as hardest to comprehend.
Rivera & Stansfield (2001)	Comparison of student performance on regular and simplified science items for Grades 4 and 6.	The results suggested that linguistic simplification did not affect the scores of English-proficient students, indicating that linguistic simplification is not a threat to score comparability. However, the small sample did not show significant differences in scores for English language learners.
	Allowing extra time	
Abedi, Courtney, Mirocha, Leon, & Goldberg (2001)	Study of the effects of using published dictionaries (both English and bilingual) as a form of accommodation for 611 students in Grades 4 and 8 in several locations in various regions of the United States. Both ELL and non-ELL students were tested in science.	This study found that providing published dictionaries was not an effective accommodation and was administratively difficult. The authors questioned the validity of using published dictionaries: English dictionaries often revealed content, whereas bilingual dictionaries often simply gave a direct translation. The researchers recommended against using published dictionaries.
Abedi, Lord, Hofstetter, & Baker (2000)	Experimental study using Grade 8 NAEP items with several accommodations, including extra time and glossary, assigned randomly to Grade 8 students ($n = 946$).	The results showed that English learners benefited from extra time, as did students already proficient in English.

(*continued*)

APPENDIX *(continued)*

Study	Characteristics	Outcomes
Miller, Okum, Sinai, & Miller (1999)	Study of Grade 11 LEP students with three accommodations: (a) extra time, (b) translation of instructions, and (c) bilingual dictionary.	Mean scores in mathematics were highest for students under standard testing conditions and lowest under extra time. Students who had extra time achieved the highest writing scores.
	Providing published dictionaries	
Miller, Okum, Sinai, & Miller (1999)	Study of Grade 11 students who received a bilingual dictionary for use during testing.	Students who received the bilingual dictionary scored lowest on the mathematics test.
Thurlow, McGrew, Tindal, Thompson, Ysseldyke, & Elliott (2000)	Study of a reading test for which commercially published English dictionaries were provided to urban middle school students in Minnesota.	The results indicated that students with self-reported intermediate English reading proficiency benefited from using the published English dictionary, whereas self-reported poor readers did not.
	Providing glossary and/or customized dictionary	
Abedi, Lord, Hofstetter, & Baker (2000)	Experimental study using Grade 8 NAEP items with several accommodations, including extra time and glossary, assigned randomly to Grade 8 students ($n = 946$).	Both English language learners and English-proficient students scored significantly higher when extra time and a glossary were provided. Extra time only, or glossary only, had less impact. For English learners, providing only a glossary resulted in slightly lower scores, probably a consequence of information overload.
Abedi, Lord, Kim, & Miyoshi (2000)	Study of students ($n = 422$) in Grade 8 science classes using NAEP science items in three test formats: (a) original format (no accommodation), (b) English glosses and Spanish translations for selected words in the margins, and (c) customized English dictionary at the end of the test booklet. The three types of test items were randomly assigned to the sampled students in intact classrooms.	English learners scored highest with the customized dictionary accommodation. Although all accommodations helped English learners, there was no significant difference for English-proficient students between test formats, suggesting that accommodation strategies did not affect construct.