CRESST REPORT 731

Mikyung Kim Wolf
Jenny Kao
Joan Herman
Lyle F. Bachman,
Alison Bailey
Patina L. Bachman
Tim Farnsworth
Sandy M. Chang

ISSUES IN ASSESSING ENGLISH LANGUAGE LEARNERS: ENGLISH LANGUAGE PROFICIENCY MEASURES AND ACCOMMODATION USES

LITERATURE REVIEW (PART 1 OF 3)

JANUARY, 2008

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Issues in Assessing English Language Learners: English Language Proficiency Measures and Accommodation Uses—Literature Review

CRESST Report 731

Mikyung Kim Wolf, Jenny Kao, Joan L. Herman, Lyle F. Bachman,
Alison L. Bailey, Patina L. Bachman, Tim Farnsworth, & Sandy M. Chang
CRESST/University of California, Los Angeles

January 2008

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the National Research and Development Centers, PR/Award Number R305A050004, as administered by the U.S. Department of Education's Institute of Education Sciences (IES).

ISSUES IN ASSESSING ENGLISH LANGUAGE LEARNERS: ENGLISH LANGUAGE PROFICIENCY MEASURES AND ACCOMMODATION USES—LITERATURE REVIEW[1]

Mikyung Kim Wolf, Jenny Kao, Joan Herman, Lyle F. Bachman,
Alison Bailey, Patina L. Bachman, Tim Farnsworth, & Sandy Chang
CRESST/University of California, Los Angeles

Abstract

The No Child Left Behind (NCLB) Act has made a great impact on states' policies in assessing English language learner (ELL) students. The legislation requires states to develop or adopt sound assessments in order to validly measure the ELL students' English language proficiency (ELP), as well as content knowledge and skills. Although states have moved rapidly to meet these requirements, they face challenges to validate their current assessment and accountability systems for ELL students, partly due to the lack of resources. Considering the significant role of assessments in guiding decisions about organizations and individuals, it is of paramount importance to establish a valid assessment system. In light of this, we reviewed the current literature and policy regarding ELL assessment in order to inform practitioners of the key issues to consider in their validation processes. Drawn from our review of literature and practice, we developed a set of guidelines and recommendations for practitioners to use as a resource to improve their ELL assessment systems. We have compiled a series of three reports. The present report is the first component of the series, containing pertinent literature related to assessing ELL students. The areas being reviewed include validity theory, the construct of ELP assessments, and the effects of accommodations in the assessment of ELL students' content knowledge.

Introduction

Public Law 107-100, the NCLB Act of 2001 (NCLB, 2002) makes clear that states, districts, schools, and teachers must hold the same high standards for ELL students as for all other students, and that educators must be accountable for assuring that all students, including ELL students, meet high expectations. By mandating that ELL students be included in annual state assessments, be subjected to annual assessments of ELP, and be included in reporting of Adequate Yearly Progress (AYP) performance targets, the federal legislation

[1] We would like to thank the following for their valuable comments and suggestions on earlier drafts of this report: Jamal Abedi, Diane August, Noelle Griffin, Margaret Malone, Robert J. Mislevy, Charlene Rivera, Lourdes Rovira, Robert Rueda, Guillermo Solano-Flores, Lynn Shafer Willner, and David Sweet. We are also grateful to Katharine Fry for her invaluable editorial assistance.

operationalized attention to the needs and progress of ELL students in both English proficiency and school subject matter.

To assure such attention, NCLB requires that states develop proficiency standards and annual measurable achievement objectives (AMAO) for student achievement in content areas and in English language development. States also must develop or identify assessments to measure ELL students' language proficiency attainment as well as reasonable, valid accommodations to measure ELL students' academic achievement (see Appendix A for the relevant parts of the NCLB legislation regarding the assessment of ELL students).

Yet the development of such measures is fraught with measurement challenges. The meaning and appropriate measures of language proficiency are open to some debate as are existing theories and literature on what constitutes fair and valid accommodations that can enable ELL students to show what they know on assessed constructs. These issues are critically important in that ELL assessment not only operates as a measure of ELL student learning—and faulty information can lead to faulty decisions—but also plays a fundamental role in leveraging reform to improve teaching and learning. As with any accountability assessment, the assessment of ELL students functions to communicate learning goals, model appropriate pedagogies, and motivate attention to students' needs and to the use of assessment to guide decision making. The establishment of sound measures to serve these functions requires a process of continuous validation.

The report that follows provides a review of the available literature in these areas to inform key issues related to assessing ELL students. The report is also intended to be used by practitioners as a resource in validating their ELL assessment systems. Before turning to that literature, the next section of the report summarizes background information about the nature, size and academic needs of the ELL population with additional rationale for the review.

Background and Rationale

Who Are ELL Students?

NCLB uses the term Limited English Proficient (LEP) and defines an ELL student as an individual who (a) is age 3 to 21 years; (b) is enrolled or preparing to enroll in elementary or secondary school; (c) was not born in the U.S. or whose native language is not English; (d) is a Native American, Alaskan Native, or a resident of outlying areas; (e) comes from an environment in which a language other than English has had a significant impact on an individual's ELP; (f) is migratory and comes from an environment where English is not the dominant language; and (g) has difficulties in speaking, reading, writing, or understanding the English language that may deny the individual the ability to meet the state's proficient

level of achievement, to successfully achieve in classrooms where English is the language of instruction, or to participate fully in society (NCLB, 2002). Based on the NCLB definition, states typically employ a home language survey, an ELP assessment, and academic achievement assessment(s) in content areas in order to identify ELL students.

It is important to note that ELL students are not a homogenous group. They come from varying cultural and linguistic backgrounds and have widely varying prior academic backgrounds and degrees of language proficiency. ELL students tend to be concentrated in the lower grades, with more than 44% enrolled in pre-K through Grade 3 (Kindler, 2002). A recent survey on the languages of ELL students showed that more than 400 languages were used, indicating the range of the diversity of this group (Kindler, 2002). The most frequently spoken language was Spanish, followed by Vietnamese at approximately 2%, then Hmong at less than 2% (Hopstock & Stephenson, 2003a; Kindler, 2002). A school-level survey conducted in 2000 by the Office for Civil Rights found that these students were 76.8% Hispanic, 12.9% Asian/Pacific Islander, 6.1% non-Hispanic White, 2.6% non-Hispanic Black, and 1.6% American Indian/Alaskan Native (Hopstock & Stephenson, 2003b). Data from the 2000 Census indicated that 59% of elementary school LEP students were U.S.-born children of immigrants, 18% were third generation, and only 24% were foreign-born (Capps et al., 2005). Among secondary-level LEP students, 44% were foreign-born. Also according to the 2000 Census, about two thirds of LEP children come from low-income families (Capps et al., 2005). The states with the highest number of ELL students are California, Texas, Florida, New York, Illinois, and Arizona (Kindler, 2002; Office of English Language Acquisition [OELA], n.d.). The heterogeneous background of ELL students is important to keep in mind as we examine policies and practices relating to their assessment.

The Size and Growth of the ELL Population

According to a recent U.S. Government Accountability Office (GAO) report (U.S. GAO, 2006), nearly 5 million ELL students[2] are enrolled in schools across the country and represent approximately 10% of all public school students. In some states, their numbers are even more substantial; for example, in California alone, approximately 1.6 million, or 25%, of K–12 students are considered ELLs (Gándara, Maxwell-Jolly, & Driscoll, 2005). Among the subgroups of the total population in K–12 public schools, ELL students are the fastest growing. Over the 10-year period between the 1994–1995 and 2004–2005 school years, the

[2] We use English language learner (ELL) to refer to students whose level of English language proficiency is not sufficient for full participation in English-only instructional environments. Although we prefer the term ELL as a positive alternative to Limited English Proficient (LEP), which connotes a deficit or "limiting" condition, LEP is used in legislation and often in research. In cases where we reference other researchers, we choose to retain their original terminology. Otherwise, we use the term ELL wherever possible.

enrollment of ELL students grew over 60%, while the total K-12 growth was just over 2% (OELA, n.d.). ELL student growth varied from state to state, with some states showing extreme growth. For example, in Colorado between 1994–1995 and 2004–2005, ELL student growth was 237.7%, as compared to 11.5% of the total enrollment growth. In Indiana during the same time period, ELL growth was 407.8%, compared to –5.1% of the total enrollment growth (OELA, 2006). Figure 1 presents the number of ELL students, density, and growth of the ELL student population in 50 states and the District of Columbia between the 1994–1995 and 2004–2005 school years.

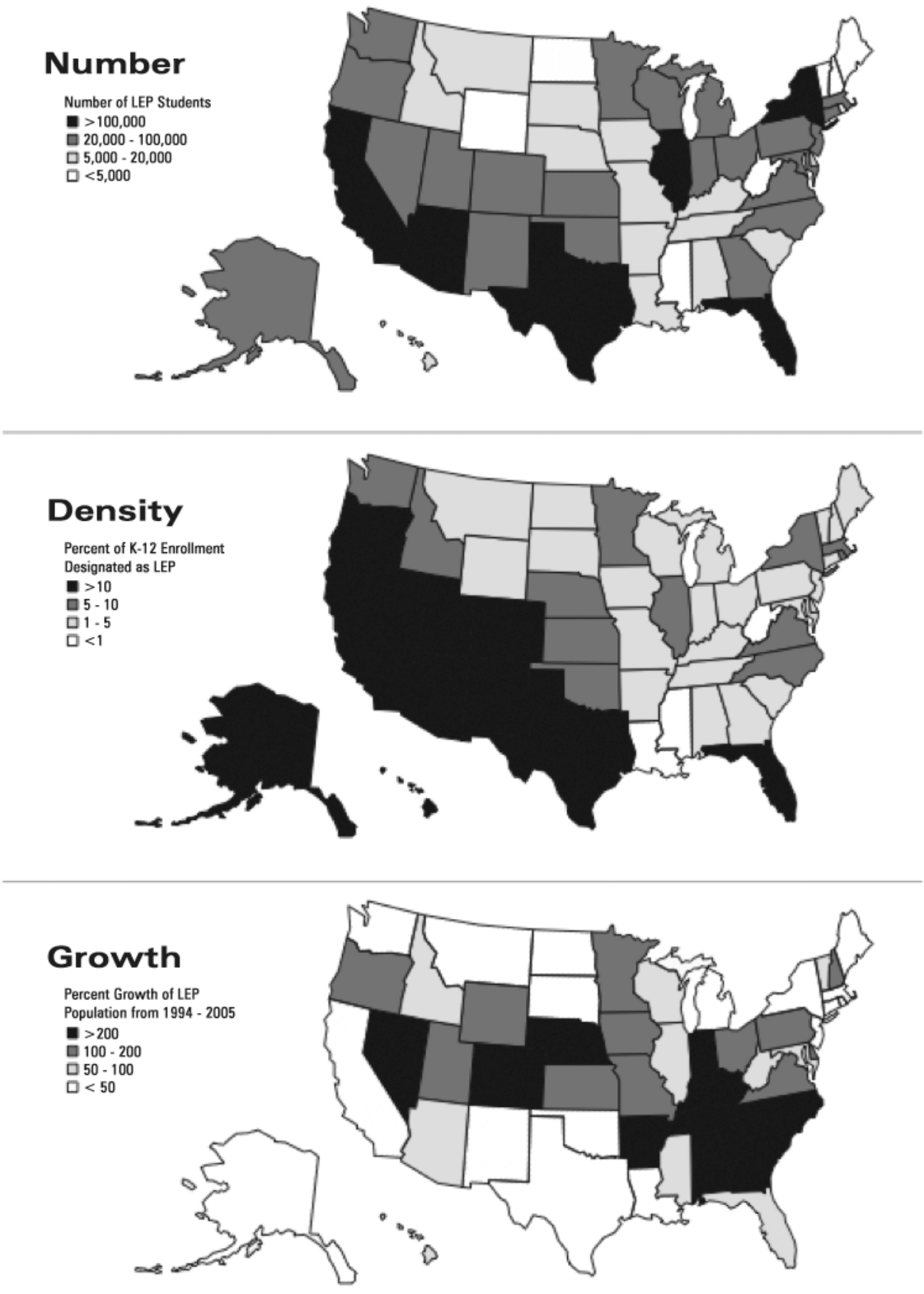

Figure 1. The number, density, and growth of LEP student population between the 1994–1995 and 2004–2005 school years. Adapted from "The Growing Numbers of LEP Students" (OELA, n.d.).

Achievement Gap Between ELL and Non-ELL Students

ELL students face a dual challenge in having to develop both academic English proficiency and content knowledge in the different subject areas. Previous research in the areas of attainment and rate of attainment of English proficiency has suggested that ELL students can take between 4 to 8 years to achieve the ELP necessary for success on academic content assessments (e.g., Collier, 1995; Cummins, 1981; Hakuta, Butler, & Witt, 2000). Furthermore, research into the poorer academic achievement of students who have recently been redesignated Fluent English Proficient (FEP) students, compared to other students in mainstream classrooms, suggests these students' difficulties may be due in part to the demands of English both in nonsheltered content classes themselves and on standardized content assessments (e.g., Bibian, 2006; Stack, 2002).

Ample data show large disparities in the achievement of ELL and non-ELL students and the need for heightened attention. The GAO reported that recent test results still demonstrate a considerable achievement gap between ELL students and the total population (U.S. GAO, 2006). Based on test scores in mathematics in school year 2003–2004 across 48 states, ELL students' math proficiency level averaged 20% lower than the overall population. Considering that mathematics may carry a lesser language demand than other content areas such as reading or science, these test results point out the urgent need to examine more closely the ways in which we assess the language proficiency and academic achievement of students who are ELLs. Similarly, results for the 2005 National Assessment of Educational Progress (NAEP) in reading showed that 73% of Grade 4 ELL students scored Below Basic as compared to 34% of non-ELL students. Similar results were seen for Grade 8, in which 71% of ELL students scored Below Basic as compared to 27% of non-ELL students (Perie, Grigg, & Donahue, 2005). For the 2005 NAEP in mathematics, 46% of Grade 4 ELL students scored Below Basic as compared to 18% of non-ELL students. In Grade 8, 71% of ELL students still scored Below Basic as compared to 30% of non-ELL students (Perie, Grigg, & Dion, 2005).

Background of the Study

In order to meet federal NCLB mandates regarding ELL students, states have moved ahead rapidly to develop an appropriate assessment system for ELL students. Using Enhanced Assessment Grants from the U.S. Department of Education (U.S. DOE, 2004), four consortia of states formed between 2003 and 2005 to collaborate on developing a common language proficiency test to adequately measure ELL students' proficiency in English. Nearly 40 states were originally represented in the four consortia, which are known as Mountain West Assessment Consortium (MWAC), Pennsylvania Enhanced Assessment

Grant (PA EAG), State Collaborative on Assessment and Student Standards (SCASS) Consortium, and World-Class Instructional Design and Assessment (WIDA) Consortium. New measures of ELP have resulted from the work of these consortia. Some states not using tests developed by the consortia have partnered with test publishers to improve their language proficiency measures.

Many states have also moved ahead with accommodation policies for assessing ELL students' academic content assessments in order to adequately measure these students' content knowledge and skills without the confounding influence of the students' limited English proficiency. However, the rush to meet NCLB assessment requirements has left states without the expertise, time, or resources to systematically document or address fundamental, underlying validity issues that are raised by the use of accommodations (U.S. GAO, 2006).

The report that follows is part of a three-pronged effort by our team of researchers at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) to help states deal with the challenges of developing sound policies and practices to support the appropriate development and use of ELL assessments. Based on a synthesis of the current research literature and a review of state practices with regard to ELL assessment, we have compiled three separate but interrelated reports. The present report we will hereafter refer to as the *Literature Review*. A second companion report, the *Practice Review*—CRESST Tech. Rep. No. 732 (Wolf, Kao, et al., 2008) summarizes the results and implications of our study of ELL assessment practices across the 50 states. The third report, *Recommendations*—CRESST Tech. Rep. No. 737 (Wolf, Herman, et al., 2008), drawn from the findings of the prior two reviews, highlights recommendations for state policy and practice as well as for future research and development.

This literature review consists of four sections. The first section reviews validity theory to understand the meaning of quality assessment and the evidence on which judgment of quality are based. The purpose of this section is to inform practitioners of general issues to consider in developing sound ELL assessments. The second section reports the latest research findings with regard to assessing ELP and development in relation to the Title III (Language Instruction for LEP and Immigrant Students) mandate of NCLB. In particular, the review focuses on the construct validity of ELL assessments by discussing the nature of the language that ELL students may need in an academic environment. The third section reviews issues in assessments of ELL students' content knowledge. Specifically, this section synthesizes the research findings on the use of accommodations that are intended to reduce

construct-irrelevant variance in measuring ELL students' knowledge and skills. The final section concludes this review with implications for both research and practice.

Issues in Validation of ELL Assessments

In order to help states monitor and validate their accountability systems under NCLB, the U.S. DOE requires that states collect and submit evidence that they meet criteria for NCLB standards and assessment requirements (U.S. DOE, 2004). The collected evidence is examined by a team of peer reviewers comprised of experts in the relevant areas using guidance developed by the DOE (hereafter referred to as the Peer Review Guidance document) to define critical elements that states must provide. The Peer Review Guidance document (U.S. DOE, 2004) makes it explicit that states must include evidence of validity, reliability, fairness, accessibility, and comparability of their state assessments, as well as valid interpretations and uses of results. The concerns raised by the Peer Review Guidance document have unique implications for ELL assessment.

In this section, we will review current validity theory, which also serves to justify the Peer Review Guidance criteria. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; hereafter referred to as *Standards*) and more recent elaborations by measurement theorists ground our theoretical framework for considering quality in state assessment systems.

Modern Validity Theory

The current definition of validity stems from Messick (1989) and the *Standards* (AERA, APA, & NCME, 1999). Messick (1989) described validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13; original emphases in italics). Similarly, the *Standards* define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). The *Standards* also conceptualize validity as accumulating evidence to provide a scientifically sound argument to justify the intended interpretation of test scores. As Kane (2001) summarized, validity issues are concerned with "the adequacy and appropriateness of the interpretations and the degree to which the interpretation is supported by the collected evidence" (p. 328).

Thus, from the perspective of modern theory, it is not the test itself that is validated; rather, validation applies to specific interpretations and uses intended by a particular test. Tests themselves are not valid or invalid; rather, validation is the accumulation of evidence

that particular interpretations or uses of a test are justified. As Bachman (2004) pointed out, validation is a two-part process that involves both "articulating an interpretive argument (also referred to as a validation argument), which provides the logical framework linking test performance to an intended interpretation and use" and "collecting relevant evidence in support of the intended interpretations and uses" (p. 258).

Likewise, Mislevy, Steinberg, and Almond (2002) proposed an evidence-centered design (ECD) approach to the design and development of assessments. In ECD, an evidentiary argument is a key concept linking the intended score-based interpretations to different kinds of evidence to support these interpretations. That is, a validity argument lays the foundation for collecting evidence that supports proposed interpretations, and the types of evidence that need to be collected to support validity depend on the intended uses of the assessment.

The validity theories reviewed above suggest that validity is an integrated, unified argument. One articulates a validity argument for evaluating the proposed interpretations and intended uses of test scores by laying out a set of claims that must be warranted or justified to support the intended use. For example, on a general level, if an ELL assessment is to be used to evaluate student progress from year to year, the items and tasks must reflect the construct of language proficiency (claim 1), must address state ELP standards (claim 2), must provide a reliable, coherent score of the construct (claim 3), must be comparable from year to year (claim 4), must be sensitive to opportunity to learn (OTL; claim 5), etc. The validation process collects evidence related to each of the claims, and an argument or judgment is made about the assessment's validity for assessing ELL progress based on the accumulated evidence. The *Standards* (AERA, APA, & NCME, 1999) and the Peer Review Guidance document (U.S. DOE, 2004) discuss several types of evidence that can be collected to support such validity arguments. One type of evidence involves the content of the test. Evidence of *content-relatedness* is concerned with the extent to which the content of the test adequately represents the construct that a test is intended to measure. Alignment studies conducted to examine the relationship between state standards and assessments represent this first type of evidence.

A second type of evidence entails an analysis of the intercorrelations among the test scores and other variables. Evidence of *convergence* (also referred to as "concurrent relatedness") considers the extent to which the test scores and other measures of the same construct are highly correlated; evidence of *discrimination* considers the extent to which the test scores are not highly correlated with measures of different constructs. For example, in providing evidence that a new measure of reading actually addressed the construct of

reading, one might look for convergence (high correlations) between the score on this measure and other concurrent measures of reading and evidence of divergence (low correlations) with measures of other constructs, for example, mathematics. Evidence of *predictive validity* considers the extent to which the test scores positively predict criterion scores that are obtained at a later time; for example, if a language proficiency test were used to place students in different instructional programs, one might look to see whether the score and the placement predicted subsequent success. Another example, common in college admissions testing, is the expectation that scores on the admissions test predict subsequent success in college.

A third type of evidence is based on the internal structure of the test and is concerned with the degree to which the test items and components conform to the way the construct is defined. For instance, one may examine whether different item formats (e.g., multiple choice and constructed response) that intend to measure the same content knowledge produce similar outcomes in students' performance. The last type of evidence is based on test takers' response processes. Here one examines the processes that students actually use in responding to test items and tasks to see whether those processes match what the test is supposed to measure. In the following section, we will discuss considerations and issues in applying validity theory to ensure the technical quality of assessments for ELL students. More detailed descriptions of types of validity evidence will be presented. In addition, other validation considerations such as reliability, fairness, and utility will be briefly discussed.

Validity Theory in Practice

Recently, Rabinowitz and Sato (2006) proposed a comprehensive set of validation criteria to evaluate the technical adequacy of an ELL assessment system. Based on the *Standards* (AERA, APA, & NCME, 1999) and an extensive review of the research, they included validity, reliability, and testing-system adequacy as major validation components. In an attempt to provide concrete and comprehensive research-based guidance, the researchers specified evidence and criteria that could be relevant for addressing each major component (see Rabinowitz & Sato, 2006, for details). For example, validity includes the specific criteria of field testing, design, content validity, construct validity, criterion validity, consequential validity, and freedom from bias. Specific evidence for content validity, for example, is based on content alignment studies and bias/differential item functioning (DIF) analysis. For freedom from bias, diverse sources of evidence include content, ethnicity, gender, disabilities, universal design, DIF, and linguistic, socioeconomic, and geographic factors.

It is notable that Rabinowitz and Sato (2006) addressed a wide variety of evidence in evaluating assessment quality. However, although their framework provides a set of rigorous

validation criteria, it is somewhat unwieldy and, as the researchers themselves noted, may benefit from simplification when translated into actual practice. The framework would benefit from a greater emphasis on construct validity as an integrative concept for validity rather than a separate criterion among many, with different types of evidence collectively contributing to support an integrated construct validity argument.

The validation process begins with consideration of the construct to be measured (e.g., ELP), the interpretations that are to be drawn from the test (e.g., what level of language proficiency students possess), and the purposes of a test (e.g., for placement, for determining progress, for making redesignation decisions). This basic specification is then the basis for asserting claims and directing the specific types of validity evidence that should be collected. The purposes of an ELP test typically can entail evidence for identifying ELL students and their ELP levels, placing an ELL student into an appropriate instructional program, monitoring student progress or attainment of English language development, and determining the redesignation of ELL students. Each of these purposes may require unique evidence, and a test of ELP that has been validated as serving one purpose cannot be assumed to serve another. On the other hand, a test of content achievement is typically intended to measure an understanding and achievement of content knowledge. And states' standards-based tests of content—for example, in reading, mathematics, and science—are intended to measure students' possession or attainment of content proficiency, typically at the levels of advanced, proficient, basic, and below basic. An assessment could reveal a student's content understanding without it being a very good measure for differentiating proficiency levels, but for ELL students an additional key validity concern in assessing content knowledge is reducing the confounding influence of language to enable students to show what they know and understand. Measurement experts call this "construct-irrelevant variance"—that is, variation in test performance that is caused by factors unrelated to the construct being measured. For example, when a math test is administered in English, test results may be a function of ELL students' ability to understand the language of the test as well as their mathematics attainment. Students' English language skills may confound their ability to show their mathematics knowledge. Given these considerations, mounting a validity argument for either ELP or content assessments requires a variety of sources of evidence. Based on the *Standards* (AERA, APA, & NCME, 1999) and the framework suggested by Rabinowitz and Sato (2006), we will discuss types of validity evidence that should be considered and how they can be associated with intended uses of the test results specific to ELL students.

Evidence based on the content of the test. The essential question here is whether the content of the assessment matches the intended construct. Sources of evidence here appeal to scientific theory about the nature of the construct (and its development) and alignment with state standards and grade-level targets, and both typically are based on documentation of the test specification and development process and on expert review. For example, to support a claim that a state's ELP test adequately measures students' English language development, one can examine the extent to which the content of the ELP test is aligned with the construct of ELP as defined in the state's ELP standards. Such evidence based on the content of the test provides some support for the proposed purpose of the test. For another example, the items in a content area test can be examined to investigate for the presence of any unnecessary linguistic complexity in measuring students' content knowledge. In practice, test developers typically conduct an expert review of the content of the items to provide validity evidence to warrant the purpose of measuring intended constructs of the test. Of important note in conducting validity studies of alignment is that there should be attention to both the content and cognitive demands of test items and tests (Herman & Baker, 2005; Herman, Webb, & Zuniga, 2003; Porter, 2002; Webb, 1999). An examination of depth-of-knowledge level of the items will provide evidence for addressing the cognitive and intellectual demands posited in standards. Content analyses and alignment studies for both the ELP and content area tests are thus one essential type of validity evidence that needs to be provided in support of the appropriate measurement of ELL students' language and content knowledge proficiency. As an example, Herman et al.'s (2003) alignment study is summarized in Appendix B.

Evidence based on the interrelations among the test scores. Examining the extent to which scores on one test are related to scores on other measures of the same or similar constructs can provide additional streams of validity evidence. For instance, in order to warrant a claim that a given ELP test measures the constructs defined in ELP and academic standards in Reading/English language-arts, one can examine the relationship between the students' scores on the ELP and Reading tests. A positive, high interrelation between the two test scores will provide a piece of validity evidence to support the claim. As another example, if one attempts to validate the purpose of an ELP test—that the test appropriately places the ELL students into different proficiency levels and thus different instructional programs—one can compare the students' ELP test scores with their performance on the same test at a later time. A positive relationship between the test scores again will provide one piece of evidence to sustain the claim about the test's purpose. With regards to a content area test, one can make a claim that accommodated and non-accommodated versions of a test measure the same construct. In order to warrant this claim, one can collect evidence by examining the test scores of students who took both versions of the test. A study conducted by Solano-Flores

and Li (2006) is a good example to illustrate how to collect this type of validity evidence (see Appendix B for a summary). The study examined where the source of measurement error lay in using different versions of accommodated tests. In addition to the test-criterion relationship, OTL and instructional sensitivity are other critical criteria to validate the uses of ELL assessments. That is, if the quality of instructional programs influences ELP progress and thus academic performance in content areas, then one needs to conduct studies of instructional sensitivity to assure that test scores differentiate programs with different levels of OTL (see Herman & Abedi's (2004) study summarized in Appendix B for this line of research).

Evidence based on the internal test structure. As mentioned earlier, analyzing the degree to which test items and components conform to the construct provides evidence for a validity argument for the proposed interpretations. Various statistical analyses (e.g., factor analysis and structural equation modeling) are available to investigate test structure. For instance, if one claims that the items of an ELP test measure both social and academic English constructs, one can investigate how test items are related to one another to measure diverse features of the constructs. Durán and Lee's (2007) study provides a good example of a method to collect the type of validity evidence based on internal test structure. They used a structure equation modeling approach to examine the construct of an ELP test (see Appendix B for a summary of the study).

Evidence based on observations of response processes. Examining how a test taker engages in a test provides another source of validity evidence. For instance, if an ELL student struggles with a general English word on a test that intends to measure mathematics skills, the student's test score may not adequately indicate the student's mathematics skills. Similarly, for an ELP test, investigating the response processes of students with different levels of ELP will provide evidence for a claim that the test measures the intended construct. The evidence based on the test taker's response process should ensure that the test taker uses the presumed construct (e.g., problem solving, application for a content-area test), rather than test-taking strategies in his or her response processes.

For the purpose of illustrating how each type of evidence can be collected in practice, a brief description of a recent ELP assessment development project is provided in Figure 2. In order to ensure technical adequacy and to serve the proposed purposes of the assessment, the WIDA consortium conducted an extensive validation study for its ELP assessment, Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs®). The major purpose of this assessment is to measure ELL students' social and academic English proficiency, based on the WIDA ELP

Standards in order to determine ELL students' language proficiency level appropriately and thus place them into different instructional programs. The WIDA Standards were formed by the consortium member states agreeing on a common set of ELP standards. The WIDA Standards then became both the official standards for each member state and the theoretical construct of ACCESS for ELLs®. The standards include five proficiency levels: Entering, Beginning, Developing, Expanding, and Bridging. Field testing was conducted in 2004 in two of the member states, Illinois and Wisconsin. A total of 6,662 students who were identified as ELL by the states participated in the field test, approximately evenly distributed among the four grade bands (i.e., K–2, 3–5, 6–8, 9–12). Sixty-one percent of the sample were native Spanish speakers with the remainder coming from diverse language groups. The validation study described in Figure 2 (see gray box below) is drawn from this 2004 field test. The study primarily focused on the content of the assessment and its relation to other measures. The study does not include observations of students' response processes.

Claim supporting use of the assessment: ACCESS appropriately measures different levels of social and academic English proficiency of ELLs.

Evidence based on the content of the test. The performance level definition and the model performance indicators of the WIDA ELP Standards were developed. Based on these two documents, a content review committee consisting of 16 expert teachers reviewed the items to assure the alignment of the content of the assessment with the construct defined in the WIDA Standards. This process is ongoing as new items are developed for ACCESS for ELLs®. After items pass the content review stage, they are examined for bias/fairness by a separate panel of bias experts from diverse ethnic groups. The WIDA reports that, on average, 33% of the items are annually replaced after this content review process. In addition, a study was conducted questioning whether the items were ordered by difficulty predicted by the WIDA Standards. Approximately 6,500 students' scores on the Listening and Reading sections were used for this study from the 2004 field test data. Using a Rasch model, average item difficulty was obtained and compared for each proficiency level. The results showed that the average item difficulty increased as the items' target proficiency level increased. The researchers concluded that the results provided evidence that the test items assess five levels of developing English proficiency as established by the WIDA Standards.

Evidence based on interrelations among the test scores. The ACCESS for

ELLs® test was specifically intended to measure an academic language construct and to closely reflect the academic language needs of school-age children, in contrast to traditional ELP assessments, which were developed to measure a general language ability construct. Therefore, the developers hypothesized that the ACCESS for ELLs® test should correlate moderately with some older ELP assessments. Scores on the field test version of the ACCESS for ELLs® were compared with scores on four older and established tests of English language proficiency using correlation methods. Data were collected from a sample of approximately 5,000 students that participated in the 2004 field test. The four tests that were used were the IDEA Proficiency Test (IPT), the Language Assessment Scale (LAS), the Language Proficiency Test Series (LPTS), and the Revised Maculaitis II (MAC II). The developers found that the ACCESS for ELLs® test had, overall, moderate and strong correlations with the other measures across grade bands and across the language domains of reading, writing, speaking, and listening (ranging from .47 to .77). Because the ACCESS for ELLs® construct was somewhat different from the constructs of the older tests, the researchers concluded that the moderate correlations constituted positive evidence of construct validity.

Evidence based on the internal test structure. The ACCESS for ELLs® test was developed as a set of four subtests measuring listening, speaking, reading and writing ability. For each subtest, the 2004 field test data were analyzed using Rasch modeling, a procedure that can be used to estimate the degree to which each item measures the latent trait. Two statistics were used to identify items that were not measures of the latent trait underlying each of the subtests: the Infit and Outfit statistics. A criterion of Infit and Outfit mean square values above 1.20 was established, and items that did not meet this criterion were examined for problems such as missing information or unclear directions. Items for which no obvious problems were found were removed from consideration in the operational test. In this manner, the items chosen for inclusion on the operational test were determined to measure the same latent trait. The researchers reported that this was considered good evidence that the test had acceptable internal structure.

Overall, WIDA conducts ongoing validation studies, including the item review and replacement process based on the WIDA Standards and Harcourt item-writing specifications. As the consortium notes, their validation process is ongoing. As the validity evidence accumulates, a longitudinal study to examine students' progress on the test could provide an importance piece of evidence to validate the purpose of the

test, measuring students' English language development progress. An examination of students' test-taking processes will also add to the current evidence that the test measures the intended construct. WIDA publishes their validation study documents, and a series of ACCESS for ELLs® technical reports are available from: http://www.wida.us/assessment/ACCESS_techReports/index.aspx

Source: Kenyon, D. (2005). *WIDA: Development and Field test of ACCESS for ELLs®*. Washington, DC: Center for Applied Linguistics.

Figure 2. Example of types of validity evidence in practice.

As noted earlier, the validity argument requires attention to both the interpretation of scores (e.g., students' proficiency level, students' attainment of proficiency and standards both in English language and academic performance) and the context or use for which that interpretation is intended (e.g., for purposes of redesignation, for measurement of progress, for instruction, for placement). A variety of evidence must be gathered to support both elements of the argument, that is, the interpretation and the context of use. The example described above illustrates how the validation process considers these two elements in practice.

Considering Other Validation Issues

Although measurement theorists consider validity as the overarching concept for examining assessment quality, the Peer Review Guidance document (U.S. DOE, 2004) draws special attention to characteristics that are essential to the validity argument for state accountability tests: reliability, fairness/accessibility, comparability of results, and procedures of test administration. That is, these characteristics also are components of validity in that a test cannot serve its intended purposes without them. Herman and Baker (2005) and both the *Standards* (AERA, APA, & NCME, 1999) and the Peer Review Guidance document (U.S. DOE, 2004) address concerns for utility and consequences as part of validity.

Reliability refers to the extent to which a test consistently measures what it is supposed to measure. For example, if a test taker takes the same test twice, the test scores should be consistent. However, measurement errors may cause inconsistency. Such errors, for example, may result from inconsistent rater scoring, an item's confusing wording, or random errors. Based on the *Standards* (AERA, APA, & NCME, 1999), the Peer Review Guidance document (U.S. DOE, 2004) includes three types of reliability criteria: (a) reliability for test scores, (b) consistency in student classification, and (c) generalizability for all relevant

sources of variance. Reliability studies need special attention in the assessment of ELL students due to the heterogeneous characteristics of the population. For the ELL population where performance of a subgroup systematically differs, the test items may measure with different levels of precision for different subgroups of test takers. Thus, a generalizability study to investigate the source of errors (whether errors come from systematic group differences) and their magnitude may be important in providing evidence for the reliability of ELL assessments.

The *Standards* (AERA, APA, & NCME, 1999) acknowledge that fairness can be interpreted in many different ways such as lack of bias, equitable treatment in the testing process, equality in outcomes of testing, and equity in OTL. When considering the assessment of ELL students, validity studies in relation to fairness can entail an investigation of the effects of accommodations and an analysis of content of the test in terms of its linguistic demands and cultural knowledge. These types of fairness studies will provide evidence to ensure construct equivalence, which is also associated with evidence for validity.

The issue of comparability of results from different forms of assessments is also of particular interest in assessing ELL students. For instance, multiple language versions of a content area test can be implemented as a type of accommodation for ELL students. Investigating the content of the items and students' performance on the multiple versions of the test is an inevitable validity issue. By the same token, if a state allows local educational agencies to select an ELP test from a list of approved commercial tests, the relationships among test scores from the different ELP tests need to be investigated for appropriate uses from test scores of different tests. Providing evidence of test comparability is an important type of validity evidence.

As far as consequences and utility are concerned, the *Standards* (AERA, APA, & NCME, 1999) note that evidence about consequences can inform validity decisions. For example, the results of an ELP test may determine redesignation status of an ELL student. The consequence is that redesignated students may take a standardized content test without accommodations. If the redesignated students' performance and learning trajectories on content tests are lower than those of non-ELL students of similar ability level (as indicated by another measure, such as reading), then consequential validity needs to be investigated. This might encourage a closer look at the ELP measure to see whether it adequately addressed all the language prerequisites needed for success in English-only curriculum and learning and/or whether there was some problem related to the cut-off scores. The Peer Review Guidance document (U.S. DOE, 2004) also maintains that states must consider both intended and unintended effects of assessments.

Given the concept of validity and validation considerations discussed above, we will illustrate a hypothetical scenario of a validation argument for an ELP test (see Figure 3). We will use California's ELP test as an example. In the scenario in Figure 3 (see gray box below), note that validity arguments differ depending on the purposes and uses. Whereas some of the evidence may be specific to an intended interpretation or use, some evidence may be overlapping to support validity arguments. It is clear that the example does not exhaustively list all possible types of evidence. Based on the available evidence, one can make a judgment about the degree of validity, that is, the plausibility of the intended interpretations and uses (Bachman, 2004; Kane, 1992; Linn & Gronlund, 2000). The important aspect to note is that validation is an ongoing process in that the validity arguments need to be expanded as intended uses are added. Appropriate types of evidence should be collected to evaluate the adequacy of different uses of the assessment.

The state of California provides an example of addressing validity issues for an ELP test. The state has developed an ELP test called California English Language Development Test (*CELDT Reporting Results 2005–2006*, retrieved October 20, 2006, from http://www.cde.ca.gov/ta/tg/el/documents/infoguide.pdf) for the purposes of (a) identifying ELL students, (b) determining their English language proficiency level, and (c) measuring the progress of ELL students' English language development. The state defines the construct of the test on the basis of the state's ELP standards, which focus on language proficiency in both social and academic settings. In addition, the ELP standards are designed to supplement the English language-arts content standards, implying that the construct of the ELP test employs the characteristics of academic English proficiency.

In relation to one of the intended uses, determining the level of students' English language proficiency, one claim to make is that test scores adequately represent the students' knowledge and skills stipulated in the ELP standards and can be used as a basis to distinguish students' academic English language proficiency. This claim should be judged using a variety of types of evidence. For example, an examination of alignment between the characteristics of test items and those of the ELP standards provides an essential type of evidence to support the argument. An investigation of the test structure is another critical type of evidence. Comparing students' performance at the different levels of proficiency with those on English language-arts tests can provide an additional type of evidence. Examining whether there are patterns and systematic variance in students' response processes is also a relevant type of evidence. Checking the internal consistency of the test is another

necessary type of evidence in making a judgment whether the claim is valid. Providing evidence that items do not function differentially for students at the same proficiency level can also be used to uphold the claim. One kind of consequence from the test use is that teachers will develop specific instruction to improve students' learning at the given proficiency level. Investigating this type of consequence provides another source of evidence. These different types of evidence will collectively be used to evaluate the validity argument for the specific use of the test results.

On the other hand, using the test results for measuring the progress of the students' ELP involves making a different type of claim and collecting different types of evidence to support that claim. One validity claim is that test results provide a basis to indicate the students' ELP progress. The alignment study described above can also provide a kind of supporting evidence. However, a more critical type of evidence can be collected from examining the alignment of the content between the current ELP test and the previous ELP test in order to appropriately measure progress. Examining the extent to which the progress determined from the test results has predictive power for students' performance on measures of content area achievement is another source of evidence. A longitudinal analysis of students' performance on ELP tests can provide another type of evidence to evaluate the claim. Collecting evidence for reliability may be more critical and complicated in that the claim is based on the comparison of the two (or more) ELP tests. Not only the internal consistency of each test but the consistency in determining the level of proficiency over the years is relevant to validate the argument. Based on these types of collective evidence, one can make an overall evaluative judgment about the validity for the intended use.

Figure 3. Example of articulating a validity argument in using an ELP test for ELL students.

In order to illustrate how to collect specific types of validation evidence, some validity studies have been selected as examples and summarized in Appendix B. As described in Appendix B, various methods, both quantitative and qualitative, are utilized. Qualitatively, content experts' review is often used to provide content-based validity evidence. Quantitatively, descriptive statistics and diverse statistical methods including regression analysis, hierarchical linear modeling, Generalizability theory (G-theory), and factor analysis are employed to provide validity evidence based on the internal structure of the test and/or the interrelation among test scores. For example, Solano-Flores and Li (2006) used a G-

theory approach to investigate whether translations into different dialects on a test can be a source of measurement error for ELL students. The researchers translated some constructed-response items from NAEP into a standard dialect of Haitian-Creole and two local dialects. Each version of the test was given to ELL students in a counterbalanced design. Results showed that the largest source of test score variance was the interaction between students, items, and dialect. In other words, the accommodation functioned differently for students depending upon their dialect of Haitian-Creole. The findings from this study provide validity evidence for the comparability of different assessment formats for ELL students.

Summary

Modern theory emphasizes validity as a unitary concept. That is, validity is an integrated, unified argument that an assessment provides sound information for specific, intended interpretations and uses, based upon multiple sources of evidence and relevant research studies. Validation considerations also include reliability, fairness, comparability of results, and consequences. The validation process begins with clear articulation of the intended interpretation and purpose of a given test and the construct it is intended to measure. Evidence is then accumulated and studies are conducted to build the argument that the assessment addresses the intended construct(s) and that each intended interpretation and use is justified. Such evidence is drawn from studies that analyze the content of a test, interrelations of the test scores with other criteria, the internal structure of the test, and students' test response processes. In addition, evidence of reliability, fairness, comparability, and consequences provides critical information for a validity argument.

The validity and technical quality of state assessments for ELL students are critical because inadequate and/or inappropriate assessment can lead to unwarranted decisions. For example, if a state assessment does not accurately reveal individual students' level of English proficiency, they may be placed in inappropriate academic environments and/or inappropriately transitioned to FEP status, which in turn may impede their subsequent progress. Faulty inferences about students' progress in attaining proficiency can lead to faulty conclusions in evaluating the strengths and weaknesses of particular programs and can lead to unwarranted sanctions for schools, districts, or the entire state. Many states only recently have adopted or developed a new ELP assessment for their ELL students (U.S. GAO, 2006). As Rabinowitz and Sato (2006) observed, these newly developed ELP assessments are being used while comprehensive technical quality and validity evidence is still being collected. Available evidence is thus very limited. The current available technical information of these ELP assessments is summarized in our *Practice Review*—CRESST Tech. Rep. No. 732 (Wolf, Kao, et al., 2008). A clear definition of a construct to be measured

is a fundamental first step in any validity argument, and all states will need to address this definition. In the next section, we review the literature on the construct of an ELP assessment, academic ELP.

Review of Literature on Assessing ELP

Under the NCLB Act, states are mandated to assess ELL students' language proficiency and to measure their progress in attaining English proficiency. The legislation emphasizes that all language modalities (listening, speaking, reading, and writing) and comprehension[3] should be measured. Consequently, states needed to act quickly to identify or develop appropriate tests of ELP to meet this federal mandate (Olson, 2002). This section reviews research on the assessment of ELL students' language proficiency and discusses validity issues in assuring the assessments' sound development and interpretation. Specifically, the section addresses five areas: (a) limitations of existing measures of ELL students' ELP; (b) the nature of the language ability to be assessed; (c) theory and research on defining academic ELP; (d) the incorporation of the construct of academic English into new ELP assessments (see our companion *Practice Review*—CRESST Tech. Rep. No. 732 [Wolf, Kao, et al., 2008] for a description of operationalizing an academic English construct into new ELP assessments in practice); and (e) validity questions in the current use of ELP assessments.

Limitations of Existing Measures of ELL Students' ELP

Research shows a variety of reasons why previous, widely used traditional language assessments are not adequate to meet the mandates of NCLB for annually measuring ELP and its progress. In reviewing five language proficiency tests, Del Vecchio and Guerrero (1995) noted problems in the construct definition and reliability of the classifications resulting from the tests. That is, there was no consensus across the tests in the definition of language proficiency, and different tests resulted in different proficiency classifications for the same students. Instead, the tests reflected a variety of definitions and different approaches to the assessment of language proficiency, some with more empirical support than others. One example is the view of discrete structuralism that permeated many tests, such that each item on the test must measure a particular discrete language skill (e.g., phonology, morphology, syntax, etc.; Davidson, Kim, Lee, Li, & Lopez, 2007). Valdés and Figueroa (1994) concurred that the majority of existing language proficiency assessments tended to

[3] NCLB does not define comprehension. States typically report comprehension scores as the combined scores of the listening and reading components of their ELP assessments.

measure discrete language skills, and argued such tests may not be appropriate to measure ELL students' language ability in an academic context.

More recently, language testing experts and language researchers, including those at the CRESST, have also highlighted the inadequacy of many assessments in tapping the development of the *academic* English language skills students need to be successful in school settings (Bailey & Butler, 2003; Bailey, Butler, Stevens, & Lord, 2007; Butler & Castellon-Wellington, 2000; Collier & Thomas, 1989; Garcia, McKoon, & August, 2006; Hakuta & Beatty, 2000; Stevens, Butler, & Castellon-Wellington, 2000). The results highlighted the need for measures of language proficiency consistent with the language demands of standardized content assessments (Butler & Castellon-Wellington, 2000).

To summarize, the previous body of literature has identified critical limitations in traditional language proficiency assessments as follows:

1. The construct of the assessment is concerned mainly with social, everyday language, and the results do not reflect whether the student is at the level of readiness or competency to perform in an academic setting (Butler, Stevens, & Castellon-Wellington, 2007).
2. There is likely a mismatch between the language skills traditionally tested and the language demands that are expected in school (Stevens et al., 2000).
3. There are both great variety and a lack of consensus in what areas of language ability are addressed and the types of tasks used in the assessments (Zehler, Hopstock, Fleishman, & Greniuk, 1994).
4. Existing assessments do not address all key language use activities (i.e., listening, reading, speaking, and writing).
5. The assessments are not systematically designed to measure progress in the attainment of English proficiency (Garcia et al., 2006).

These limitations in traditional language tests, as well as the requirements of NCLB, have spawned the development of a new generation of ELP measures. The following subsections will describe the theoretical and research background to support the next generation of ELP assessments.

The nature of language ability to be assessed. As described in the discussion of validity above, being clear on the construct to be measured and the purpose to be served is a fundamental concern. Considering that ELP assessments may be used as one criterion in making a variety of academic decisions about ELL students, the constructs addressed by ELP assessments would need to reflect language ability required to perform in an academic context. The Peer Review Guidance document (U.S. DOE, 2004) also stresses that ELP assessments should be aligned with the state's ELP standards.

As briefly mentioned above, a series of studies have been conducted at CRESST to investigate the nature of the language with which ELL students have to cope in academic settings and for both language proficiency assessments and content area assessments (Bailey & Butler, 2003; Bailey et al., 2007; Butler & Castellon-Wellington, 2000; Stevens et al., 2000). Stevens et al. (2000) examined the relationship between language proficiency and student performance on content standardized tests, comparing the type of language assessed on language tests and the language used on content tests. Comparing the language in the Language Assessment Scales (LAS; De Avila & Duncan, 1990) and a standardized social studies test for the seventh grade, the researchers found that the language proficiency test addressed more general language whereas the content test employed more academic language. In the LAS, syntax was less complex, vocabulary consisted generally of everyday words, and discourse was less demanding to process. On the other hand, the content area test had academically more demanding language including academic vocabulary, various syntactic structures, and more specific linguistic registers. Based on their findings on the mismatch of language between the two tests, Stevens et al. (2000) argued that there is a need for the development of language proficiency assessments that measure students' academic language proficiency.

A series of CRESST research studies in this line suggest caution in drawing inferences based on traditional language assessments, indicating that such assessments should be carefully used because they may indicate a limited range of academic language ability. That is, social English is not highly correlated with the more demanding language of school, and the research highlights the importance of aligning ELP assessments with the academic language requirements of content assessment, instruction, and state ELP standards (Bailey, 2007; Bailey & Butler, 2004). The alignment between the language demands implicit in ELP standards and those in content standards also raises a concern about the meaning of language proficiency in academic settings, that is, whether, on the face of it, achieving language proficiency standards prepares students for the language needed to achieve content standards (Bailey, Butler & Sato, 2005). The following subsection reviews different approaches to define academic English language and its characteristics.

Defining academic ELP. In the fields of English as a second language (ESL) and applied linguistics, teaching and learning English for academic purposes (EAP) has been developed as its own discipline. These three fields have advanced several different approaches to characterizing academic English and general English. The following review of the literature relies predominantly upon studies previously reviewed by CRESST researchers

(e.g., Bailey & Butler, 2003; Bailey, 2007) as these are most directly pertinent to the definition of academic English.

One approach is based on the cognitive and contextual demands of language use (Cummins, 1981, 1983, 2000; Mohan, 1986). Cummins (1981, 2000) distinguished academic language from interpersonal and conversational language, labeling the former as Cognitive Academic Language Proficiency (CALP) and the latter as Basic Interpersonal Communicative Skills (BICS). Cummins argued that CALP is both cognitively demanding and context-reduced, whereas BICS are cognitively undemanding and context-embedded. Context-embedded communication includes many features that support participant comprehension, such as gesture, intonation, and reference to concrete objects. One-on-one conversation and storytelling would both be examples of context-embedded communication. Context-reduced communication includes fewer aids for the participant, relying more on abstract ideas, and thus placing more cognitive demands on the recipient of the message. Examples of context-reduced communication would include reading a piece of fiction or a note from a friend, or listening to a radio broadcast.

According to Cummins (1981, 2000), there are clear differences in the acquisition and developmental patterns between BICS and CALP. The context-reduced communication of CALP often involves a high degree of lexical variety and syntactic sophistication. In contrast to CALP, meanings in BICS use are conveyed with the help of contextual support and paralinguistic cues, facilitating communication, and thus making the development of BICS easier and faster. Table 1 summarizes the key features of BICS and CALP.

Table 1

Key Features of BICS and CALP

BICS	CALP
Conversational language	Cognitively demanding and academic language
Dimensions of language proficiency related to social skills	Dimensions of language proficiency related to cognitive and academic skills
Proficiency through interpersonal interactions	Proficiency through schooling and literature
Context-embedded communication	Context-reduced communication
Development plateaus through conversational usage	Development is continual through schooling
May develop more quickly and easily than CALP	May take longer to develop than BICS

Note. BICS = Basic Interpersonal Communicative Skills, CALP = Cognitive Academic Language Proficiency.

The language function perspective represents a second approach to defining academic language, originating in Halliday's (1978) view that the purpose of language use is to

accomplish specific tasks such as informing, analyzing, and comparing. From this perspective, Chamot and O'Malley (1994) defined academic language as "the language that is used by teachers and students for the purpose of acquiring new language and skills . . . imparting new information, describing abstract ideas, and developing students' conceptual understanding" (p. 40). They also described academic language functions as the task that language users must be able to perform in the content areas. Similarly, from an analysis of features of academic English in the text and discourse of a social studies classroom, Short (1993) demonstrated that, in order to achieve effective communication in American history classes, students must be able to use the following language functions: explaining, describing, defining, justifying, giving examples, sequencing, comparing, and evaluating. Short also argued that those language functions play an important role for ELL students in acquiring content knowledge.

Another approach to defining academic English is through the analysis of linguistic elements that make up the register of schooling. Schleppegrell (2001) provided such an analysis of school-based texts and labeled academic English as "language of schooling." In terms of lexical features, Schleppegrell described the lexical choices of school-based texts as specific and technical rather than generic. With respect to grammatical features, she maintained that central grammatical features of school-based text, such as the use of lexical subjects and nominalizations, enable the language of schooling to present information in highly structured ways.

Solomon and Rhodes (1995) utilized the "registers" concept at the discourse level. According to them, academic language is associated with "stylistic register," where styles are tied to specific academic tasks and broad discourse levels, and not limited to sentence-level linguistic features. For example, "story retelling" is characterized by the specific style (i.e., precise chronological order) that teachers expect students to follow when they retell a story.

In efforts to develop a framework of academic language at the college level, Scarcella (2003) integrated discrete linguistic features (phonological, lexical, and grammatical components), the language functions perspective (sociolinguistic component), and the stylistic register perspective (discourse component). As the basis of the framework, Scarcella also adopted the "communicative competence" model proposed by Canale and Swain (1980) and Bachman (1990), which includes grammatical competence, sociolinguistic competence, discourse competence, and strategic competence.

Based on a series of studies with CRESST colleagues investigating the nature of language in academic texts, tests, and teacher discourse, Bailey (2007) defined the ability to be academically proficient as "knowing and being able to use general and content-specific

vocabulary, specialized or complex grammatical structures, and multifarious language functions and discourse structures—all for the purpose of acquiring new knowledge and skills, interacting about a topic, or imparting information to others" (p. 10). Similar to Scarcella's (2003) framework, Bailey also encompassed academic features at the lexical, grammatical, and discourse levels, as well as language functions. For instance, at the lexical level, academic English vocabulary comprises general academic vocabulary that can be used across content areas (e.g., *represent, demonstrate*, used in different disciplines) and specialized academic vocabulary within a specific content area (e.g., *diameter, radius*, predominantly used in mathematics).

Operationalizing academic English in an ELP assessment. The complex nature of academic English as described in the previous section raises several issues to consider in examining or developing the construct for an ELP assessment. Bachman (2006) pointed out that distinguishing academic English proficiency from content knowledge is a serious conundrum in assessing ELL students' language proficiency. It is partly because the construct of language ability should be viewed from an interactional framework of language use (Bachman & Palmer, 1996). In an interactional framework, language knowledge, topical knowledge, and strategic competence can all be considered in defining the construct for a language assessment, depending on its purpose. For an ELP test, it is debatable which component(s) of language ability and use should be included to define its construct. Bachman (2006) discussed how construct may be defined within the framework of "language assessment for specific purpose" as identified by Douglas (2000). For language assessment for academic purposes, the construct may be defined as the ability to use language to process information about academic content (Bachman, 2006). Furthermore, Bachman stressed that research is needed to investigate the extent to which English proficiency and content knowledge affect test performance differently.

As mentioned earlier, a number of recently developed ELP measures have attempted to incorporate varying degrees of academic English proficiency in defining the construct to be assessed (Bailey, 2007). For example, the New IDEA Proficiency Test (New IPT) assessment for pre–K/K ELL students (Ballard & Tighe, 2005) has adopted the academic English language construct framework, coupled with the analysis of various state-level ELP and content standards. The current California English Language Development Test (2004) has similarly been developed on the basis of the California ELP standards, which include characteristics of academic English. The four consortia have also endeavored to operationalize the construct of academic English proficiency into their assessments. For instance, the ACCESS for ELLs® test by the WIDA consortium specifically delineates its

construct as academic ELP. The construct includes language ability used in specific academic content areas such as language arts, mathematics, science, and social studies. It is also designed to be aligned with collaborating states' ELP standards.

However, it should be noted that definitions of academic English proficiency, as described just above, still vary. The lack of a commonly accepted framework of academic English language for K–12 students poses a challenge for states and test developers in operationalizing the construct for their ELP assessments (see our *Practice Review*—CRESST Tech. Rep. No. 732 [Wolf, Kao, et al., 2008] for examples of diverse definitions of the academic English construct delineated in states' standards and assessments).

Issues in Using Language Proficiency Assessments for Making Decisions

The previous section primarily discussed issues in the definition of the construct(s) to be measured by an ELP assessment itself. This section discusses how results from such assessments are used, a second critical issue for validation purposes. In practice, ELP assessments are used for multiple purposes including identifying ELL students, determining levels of proficiency for instructional placement, redesignating ELL status, and tracking the progress of students' developing language proficiency. ELP assessment results are part of inclusion decisions for large-scale, standardized, content-area testing (i.e., math and science testing). ELP assessment results also are the most common source of evidence used to determine specific accommodations needed for large-scale content-area testing. Yet the available validity evidence to support such use is scant. For example, if ELP results are to be used to support inclusion and accommodation decisions, there should be evidence that performance on the ELP assessment is relevant to students' ability to access (understand) content area tests in English. The comparability of language complexity and demands between ELP and large-scale academic achievement assessments can provide one source of validity evidence for such a claim. That is, if ELP assessments are aligned with the language demands of content tests, then we have evidence that performance on the ELP test tells us something important about the extent to which students are able to handle the language demands of the content area tests given in English and about whether students may need accommodations. Yet Stevens et al.'s (2000) comparison study cited earlier revealed substantial variation in the language demands of the reading section of a commonly used ELP assessment and the social science section of a large-scale standardized academic achievement assessment

Evidence on the use of ELP assessments for ELL redesignation criteria and transition decisions also is problematic. For example, if students are judged proficient based on an ELP assessment and are transitioned into Fluent English Proficient status, we might expect their

subsequent performance to be similar to that of English only or non-ELL students. However, research on the academic achievement of students who were recently redesignated FEP yielded mixed results in terms of the students' subsequent academic performance relative to non-ELL students (Abedi, Leon, & Mirocha, 2003; Bibian, 2006; Stack, 2002). Grissom's (2004) study on redesignation in California similarly raises questions about the meaning of being judged proficient based on ELP assessment results. At the time of the study, the state employed multiple criteria collectively in its redesignation process for ELL students. The criteria included (a) an assessment of ELP; (b) teacher evaluation; (c) parent opinion and consultation; and (d) academic achievement based on content-based standardized tests. Grissom found that performance on content-based standardized tests was the best predictor for the redesignation rate of the ELL students. Findings of this study suggest that the results of a language proficiency assessment alone are insufficient for redesignation decisions (see Appendix B for the summary of this study).

The present review has briefly discussed validation research related to inclusion and redesignation uses. Notably, there is a paucity of research on using ELP assessments to make other decisions. Even for the areas discussed above, moreover, additional research is needed to determine what other criteria, beyond ELP assessments, states should use to redesignate ELL students, and how language proficiency assessments can most consistently predict the readiness of ELL students for the mainstream English classroom. The limited and problematic results of available research in this area are partly due to the recent implementation of new ELP assessments as a result of NCLB. These new assessments should diminish the problems of inattention to academic language and to discrepancies in construct definition when using different commercially available tests to classify and redesignate ELL students. Future research should provide evidence of the validity of using ELP assessment results for the full range of decision-making purposes for which they are intended, including redesignation, accommodation, placement, and progress monitoring. One item for the research agenda on utility issues is to determine the extent to which the results of this type of assessment are used in a valid way as a part of the decision-making process.

Summary

Traditional language tests of the last generation have been criticized for their inadequate construct representations in an academic context. That is, test scores from these tests may not necessarily indicate an ELL student's ability to handle the language demands of materials on standardized content tests or in classroom curriculum. Considering that one of the common uses of these assessment results is to make redesignation decisions for ELL students, the limitations of such tests raise serious validity concerns.

To meet the NCLB requirements as well as to overcome these limitations, the next generation of ELP assessments has begun to respond. In particular, the newly developed ELP assessments have incorporated academic English proficiency in defining the construct to measure. Current test developers and users face challenges in accumulating evidence of validity for the specific uses for which these assessments are intended as well as in establishing their technical quality. As reported by the GAO (U.S. GAO, 2006) and Rabinowitz and Sato (2006), to date, there has been little comprehensive research on whether and how the newly developed ELP measures address previous limitations and whether and how they provide accurate information to improve decision making for ELL students. Future research should address this need as well as the need for further research and consensus on meaning and defining characteristics of academic English language to more firmly ground the development and use of these measures.

Review of Literature on Assessing Academic Achievement with the Use of Accommodations

In this section, we move from the assessment of ELL students' language skills to the consideration of validity issues in assessing their achievement in content knowledge and skills. In addressing these issues, some researchers have argued the importance of considering both language and cultural factors in the testing of ELL students (Geisinger, 2003; Solano-Flores & Trumbull, 2003; Tippeconnic & Faircloth, 2002). The *Standards* (AERA, APA, & NCME, 1999) note that for "all test takers, any test that employs language is, in part, a measure of their language skills" (p. 91). Thus, if students have not yet acquired sufficient language skills, they may not be able to adequately demonstrate their knowledge in a content-based assessment. Research has suggested that the linguistic complexity of assessments in content areas could affect the validity of assessments, particularly for ELLs (Abedi, 2002; Abedi, et al., 2003). This may partly explain why there are persistent achievement gaps between ELL students and their non-ELL counterparts.

Language proficiency interferes with ELL students' ability to show what they know in content tests given in the English language. There is a large and growing body of research investigating specific aspects of the content of academic achievement tests that may differentially affect the performance of ELL students. Studies conducted at CRESST, for example, have suggested that ELL students have more difficulty responding to test items that are linguistically complex (Abedi, Courtney, & Leon, 2003a; Abedi, et al., 2003; Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2003; Sireci, Li, & Scarpati, 2003). This differential difficulty may threaten the validity and reliability of content-based assessments of ELL students (Abedi, 2002; Abedi &

Hejri, 2004; Abedi & Lord, 2001) because ELL students could have trouble interpreting the vocabulary, or could misinterpret words literally (Durán, 1989; Garcia, 1991). In their analyses of mathematics and science subsections of 3rd- and 11th-grade standardized content assessments, Imbens-Bailey and Castellon-Wellington (1999) found that two thirds of the items included general vocabulary considered uncommon or used in an atypical manner. One third of the items included complex or unusually constructed syntactic structures. To accurately assess knowledge within content areas, students must comprehend what the items are asking and understand the response choices. The purpose of content-based standardized achievement tests is to measure students' knowledge of specific content areas, not to test non-content vocabulary.

In other words, the linguistic complexity of test items may compromise the construct validity of content assessments for ELL students because it interferes with these students' ability to respond and thus demonstrate their content knowledge (Abedi, 2006; Haladyna & Downing, 2004; Messick, 1994). Some studies have shown that reducing the unnecessary linguistic complexity of test items helps improve the performance of ELL students without compromising the validity of the assessment (Abedi & Lord, 2001; Abedi, et al., 2000; Kiplinger, Haug, & Abedi, 2000; Maihoff, 2002). More details on these studies appear in the next section of this report.

Aside from linguistic complexity, cultural variables may influence student performance on assessments. Such variables include student disinclination to ask questions, attitudes toward competition, attitudes toward individualism versus collectivism, gender roles, attitudes toward the use of time, attitudes toward the demonstration of knowledge, use of body movements and gestures, and use of eye contact (Liu, Thurlow, Erickson, Spicuzza, & Heinze, 1997). Abedi, Lord, and Hofstetter (1998) found that student background variables such as language background, length of stay in the United States, overall grades, and the number of school changes were valuable predictors of ELL student performance in math and reading.

Accommodations as a Way to Provide Valid Content Assessments for ELL Students

Because the language of a test may introduce construct-irrelevant components to the testing process, as the *Standards* (AERA, APA, & NCME, 1999) noted, "it is important to consider language background in developing, selecting, and administering tests and in interpreting test performance" (p. 91). One way in which language background has been addressed in practice is by the introduction of testing accommodations, which are meant to help ELL students overcome the language barriers they may face in understanding and responding to content assessments. In other words, accommodations help "level the playing

field" with mainstream students. Accommodations are strategies intended to reduce threats to test score validity. They thus serve the dual related purposes of both leveling the playing field and producing valid assessment outcomes.

Accommodations generally refer to changes to a test itself, or changes to the way a test is administered. Koenig and Bachman (2004) drew on the *Standards* to define an accommodation as "the general term for any action taken in response to a determination that an individual's disability or level of English language development requires a departure from established testing protocol" (p. 1). According to Rivera, Collum, Shafer Willner, and Sia (2006), accommodations should "address the unique needs of the students for whom they are provided without invalidating the test construct" (p. 1). This means that ELL students should be provided with assistance to overcome both linguistic and sociocultural barriers that make the content of a test inaccessible to them. ELL students should be able to demonstrate their knowledge of the content with minimal interference from test language. At the same time, the accommodation should not provide ELL students with unfair advantages over their peers or change the nature of the construct that is being measured.

The concept of providing testing accommodations stems from assessing students with disabilities. For example, a student who is visually impaired could perhaps benefit from receiving a test in large print. Similarly, ELL students could benefit from receiving accommodations that specifically address their language barriers. Thus, accommodation strategies that address such needs can help make tests fairer and more accessible to these students. Because accommodations were first developed for students with disabilities, many accommodations allowed for ELL students were ones originally created and designated for students with disabilities. Therefore, research has been conducted and should continue to be conducted to examine the effectiveness of accommodations that are used for ELL students. Research on accommodations should examine three criteria: effectiveness, validity, and feasibility.

Accommodations must both be effective and produce valid assessment results. An effective accommodation raises the performance of ELL students, and a valid accommodation does not alter the nature of the task (Abedi, Courtney, & Leon, 2003a; Koenig & Bachman, 2004). Past research has attempted to identify effective and valid accommodations by focusing on a select few language-related accommodations (Abedi, Courtney, & Leon, 2003a; Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005; Abedi, Hofstetter, & Lord, 2004).

Researchers have suggested an "interaction hypothesis" (or "maximum potential thesis" [Zuriff, 2000, as cited by Sireci et al., 2003]) to justify the use of test accommodations. The

hypothesis states that students who need a particular type of accommodation would benefit from it, and that other students who do not need that type of accommodation would not benefit from it (Sireci et al., 2003). As long as this is true for a specific accommodation, then scores from an accommodated assessment will be a valid indicator of the construct being assessed.

To examine a particular accommodation under the interaction hypothesis, some researchers have conducted experimental studies (detailed in the next section) by including both ELL students and non-ELL students who are tested under both accommodated and non-accommodated conditions (see Abedi, Courtney, & Leon, 2003a, 2003b; Abedi et al., 2005). If ELL students' performance under accommodated conditions is higher than under non-accommodated conditions, then that indicates the accommodation is effective. However, if non-ELL students' performance under accommodated conditions is also higher, then that indicates the accommodation may not be valid, because the data suggest that students using the accommodation are receiving an unfair advantage. By referring to an accommodation as valid, we mean the accommodation produces valid assessment outcomes, and results are comparable to and can be aggregated with non-accommodated test scores. If an accommodation gives an unfair advantage to those receiving it, or, if everyone including non-ELL students would benefit from an accommodation, then the accommodated test results could be inflated (Sireci et al., 2003).

Additionally, accommodations must also be feasible, or practical in large-scale assessments, to implement. One should also not assume a "one size fits all" approach to accommodations, as ELL students are a heterogeneous group, and there could also be a differential impact of accommodations. There is also the issue of content assessments being unavoidably intertwined with language skill, and in some cases, a content target construct could be related to a language skill construct. Accommodating ELL students can therefore be especially challenging as compared to accommodating students with disabilities, and research on accommodations must take into account all of these factors.

Research on the Effects of Accommodations

Research on accommodations used for ELL students has mainly focused on only a handful of those used in practice, leaving many used accommodations still yet unexamined (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Sireci et al., 2003). Although Rivera et al.'s (2006) review of state policy documents found 75 accommodations mentioned in 47 states during the 2000–2001 school year, not nearly as many have been studied in the literature. Furthermore, their examination found that 31 of the 75 accommodations did not support the linguistic needs of ELL students (but rather, supported students with physical or

cognitive disabilities). Sireci et al.'s (2003) review of the literature found a total of 6 different accommodation strategies that were investigated, most of which were for the dictionary or glossary accommodation. Francis et al. (2006) selected true experimental studies for their meta-analysis on the effects of accommodations and found that the empirical studies focused mainly on the following accommodations: simplified English, customized English dictionaries or glossaries, bilingual dictionary, glossary or marginal glossaries, extra time, dual-language test booklets, and native language tests.

For the few accommodations that have been studied, results have been mixed. For instance, providing translated assessments can introduce other complications (Hambleton, 2001), and some accommodations may actually provide an unfair advantage to those receiving them, as mentioned above. Furthermore, accommodations must not only be effective and valid, but they must also be feasible to implement. Abedi et al. (2005) found that dictionaries were physically cumbersome and not always useful to students, and that commercially published English dictionaries sometimes provided information on what the test was asking students to recall. Brown (1999) found no significant differences when offering students two different test versions (original and "plain language"). Consequently, research that identifies accommodations that are effective, valid, and feasible is needed.

Following are brief descriptions of several ELL accommodations more prevalent in the research literature. The studies cited here are also summarized in Appendix B with respect to study data and major findings.

Extended time. Providing students with extended or extra time is one of the most commonly used accommodations for ELL students (Rivera, Stansfield, Scialdone, & Sharkey, 2000). Extended time was one of the top five most frequently cited accommodations in the GAO's review of 42 state documents (U.S. GAO, 2006). Although it can be provided as a single accommodation, extended time is often provided in conjunction with other accommodations (Rivera et al., 2006).

Extended time was found to increase the performance of both ELL and non-ELL students when used either alone or simultaneously with another accommodation (Abedi, Lord, Hofstetter, et al., 2000; Chiu & Pearson, 1999; Hafner, 2001; Thurlow, 2001). Miller, Okum, Sinai, and Miller (1999) found inconclusive results in their study of extra time and other accommodations.

Dictionaries and glossaries. A dictionary as an accommodation must be provided along with extended time to avoid the problem of information overload and to provide the time to enable dictionary or glossary use. Some studies have found that providing a

commercially published English dictionary and extended time affects performance of all students (Maihoff, 2002; Thurlow, 2001; Thurlow & Liu, 2001). Commercially published dictionaries often vary widely in the vocabulary difficulty level of their definitions (Kopriva, 2000). Furthermore, by gaining access to definitions of content-related terms, recipients of a published English dictionary may be advantaged over those who do not have access to such a dictionary, and this may compromise the validity of the assessment (Abedi et al., 2005). Also, providing a dictionary can be logistically difficult and burdensome, which makes it less feasible to implement (Abedi et al., 2005).

In one study, a *customized* English dictionary was introduced as a more valid alternative to an entire published dictionary (Abedi, Courtney, & Leon, 2003a). The customized dictionary was a literal "cut-and-paste" of actual dictionary entries; only terms found in the test and not related to the content were included in a separate, stapled booklet. Results of this study suggested that it was an effective and valid accommodation for ELL students. Another study incorporated the customized concept into a "pop-up" glossary requiring computer administration (Abedi, Courtney, & Leon, 2003b). In this study, when students pointed a computer mouse over non-content terms, an English glossary definition would appear in a pop up window. This accommodation, which was provided with extended time, was found to be valid and effective for ELL students in Grade 8.

A test-specific English glossary can be provided as an alternative to a dictionary. The distinction between dictionary and glossary is that, generally speaking, a dictionary provides a definition of a word, but a glossary customizes the definition for specific contextual needs (Abedi, Hofstetter, Baker, & Lord, 2001). However, although ELL students' performance increased by 13% when they were tested using a glossary combined with extended time, non-ELL students' performance in the same circumstances increased by 16% (Abedi et al., 2001). Thus, while study results show evidence of the effectiveness of the accommodation for ELL students, they raise concerns about the validity and comparability of accommodated and non-accommodated results.

Bilingual dictionaries were cited as one of the top two most commonly used accommodations by the GAO's review of 42 state documents (U.S. GAO, 2006). However, the effect of bilingual dictionaries on test validity is still a concern. As with an English dictionary, students have access to content-related terms and thus perhaps gain an advantage over those who do not have such access (Abedi, Lord, Boscardin, & Miyoshi, 2000; Abedi et al., 2005). Another major limitation with a bilingual dictionary is the content equity issue. Different published bilingual dictionaries present a substantial range of content coverage (Abedi et al., 2005). Furthermore, students who speak another language at home may not be

literate or fully literate in their home language and might not find a bilingual dictionary useful. They may also not be used to using one in the classroom, which makes it unfamiliar.

More research appears to be necessary to investigate the validity of using bilingual dictionaries and glossaries. The literature that is available, however, suggests that English glossaries are preferred over Spanish glossaries, and that English language dictionaries (and glossaries) were the only accommodation found to have a statistically significant and positive average effect size based on Francis et al.'s (2006) meta-analysis of empirical accommodation studies. Six other accommodation strategies were investigated, including bilingual dictionaries and glossaries, but did not show a positive effect (Francis et al., 2006). Rivera et al. (2006) observed that more research is needed to clarify the distinction between dictionaries and glossaries in order to adequately investigate their separate effects on test validity.

Native language tests. In the GAO's survey (U.S. GAO, 2006), 16 states reported offering statewide native language assessments in language arts or mathematics in some grades for the 2004–2005 school year. However, there are many concerns over the use of native language testing, and translating a test can make the instrument easier or harder in another language as some cultural phrases and idioms can be difficult to translate (Hambleton, 2001). Furthermore, students may be proficient only in *speaking* their home language, not in reading it, and the language of instruction and the language of assessment must be aligned in order for native language testing to be useful (Abedi, Lord, Hofstetter, et al., 2000). If ELL students are instructed in English, then a native language test could negatively impact their score. Additionally, Solano-Flores, Trumbull, and Nelson-Barber (2002) contended that test translation suffers from serious theoretical, methodological, and practical limitations relating to culture and word sensitivity. Developing a test involves more than translation; the test must undergo a rigorous validation process that could be quite time-consuming and costly. Solano-Flores et al. (2002) recommended developing two language versions concurrently. These are practical issues for a state to consider when investigating the use of native language tests, especially if only a small number of ELL students speak a particular language (U.S. GAO, 2006).

Dual-language or side-by-side bilingual test versions. The concept of dual-language test versions stems from native language testing. In this format, the same information is presented in two languages on the same page, often side-by-side in separate columns. Some researchers have examined the use of dual-language tests in which test booklets contain original English items with corresponding items on facing pages translated in students' home language.

Duncan et al. (2005) created dual-language test booklets with Spanish versions of test items on the left-hand pages, and English versions of the items on the right-hand pages. Quantitative analyses indicated psychometric equivalence between the dual-language and English-only test booklets, and 85% of students responding to a questionnaire reported the dual-language test as being "useful" or "very useful." Furthermore, students given the dual-language test booklet preferred the dual-language format over a Spanish-only format, and strongly preferred the dual-language format over having an English-only test booklet with a bilingual dictionary. However, despite the preferences, no differences in test performance were detected.

Likewise, no differences in performance were detected in a similar study by Abedi, Courtney, Leon, Kao, and Azzam (2006) in which they created dual-language test booklets by placing English versions of mathematics test items in one column and Spanish versions of the test items in another column on the same page.

Sireci and Khaliq (2002) explored psychometric properties of a dual-language version of a fourth-grade mathematics test, which was given as part of a state-mandated testing program. To allow for greater confidence in drawing conclusions, multiple statistical methods were applied to evaluate the equivalence of the English and English-Spanish versions of a statewide mathematics assessment. Results suggested slight structural differences across the two versions of the test, which may be in part because of performance differences of the studied groups. The authors asserted that use of dual-language test booklets deserves further study.

Linguistic modification. Linguistic modification of test items can be defined as modifying the language of the test text to reduce linguistic complexity while maintaining the construct of the test. Other researchers refer to this as linguistic *simplification*[4] (Rivera & Stansfield, 2004) or simplified English (Francis et al., 2006). Other terms include plain English, language modification, and language simplification. Assessments that are linguistically modified may facilitate students' negotiation of language barriers. This may be accomplished by shortening sentences, removing unnecessary expository material, using familiar or frequently used words, using grammar considered more easily understood (such as present tense), and using concrete rather than abstract formats (Abedi, Lord, & Plummer, 1997).

[4] We prefer the term "linguistic modification" because simplification can have the connotation of "dumbing down" a test. We follow the work of Abedi and colleagues by contending that the linguistic structures of test items are not simplified, but rather, *modified* to reduce or eliminate factors that can interfere with comprehension and are irrelevant to the construct. Sometimes modified test items can contain more words and/or sentences than the original items in order to reduce the number of complex linguistic features.

The LEP Consortium of the Council of Chief State School Officers (CCSSO)/ State Collaborative on Assessment and Student Standards (SCASS) made seven recommendations for improving accessibility of text material (Kopriva, 2000). Table 2, as cited by Abedi, Courtney, and Leon (2003a), summarizes research findings of Abedi et al. (1997), accompanied by practical recommendations from Kopriva (2000) and Shuard and Rothery (1984).

Table 2

Linguistic Complexity Research Findings and Practical Recommendations

Research findings[a]	Practical recommendations[b]
Short words (simple morphologically) tend to be more familiar and, therefore, easier.	Use high-frequency words.
Passages with words that are familiar (simple semantically) are easier to understand.	Use familiar words. Omit or define words with double meanings or colloquialisms.
Longer sentences tend to be more complex syntactically and, therefore, more difficult to comprehend.	Retain subject-verb-object structure for statements. Begin questions with question words. Avoid clauses and phrases.
Long items tend to pose greater difficulty.	Remove unnecessary expository material.
Complex sentences tend to be more difficult than simple or compound sentences.	Keep to the present tense, use active voice, avoid the conditional mode, and avoid starting statements and questions with clauses.

Note. Adapted from Abedi, Courtney, and Leon (2003a).
[a]Based on Abedi et al. (1997). [b]Based on Kopriva (2000) and Shuard and Rothery (1984).

Previous research examining the language of math problems found that making minor changes in the wording of a problem affected student performance (Cummins, Kintsch, Reusser, & Weimer, 1988; De Corte, Verschaffel, & DeWin, 1985; Hudson, 1983; Riley, Greeno, & Heller, 1983). Larsen, Parker, and Trenholme (1978) compared student performance on math problems that differed in sentence complexity and familiarity levels of the non-math vocabulary. For example, low-achieving Grade 8 students scored significantly lower on the items with more complex language.

Abedi et al. (1997) created revised versions of test items using recommendations for reducing linguistic complexity, and found significant differences with respect to performance between student scores on complex items and less complex items. Abedi and Lord (2001) found that modifying the linguistic structures in math word problems can affect student performance. In interviews, students indicated preferences for items that were less linguistically complex, and they also scored higher on linguistically modified items. The

linguistic modification accommodation had an especially significant impact for low-performing non-ELL students and ELL students, but did not affect higher performing non-ELL students.

In studies using items from NAEP, student scores on actual NAEP items were compared with scores on parallel modified items in which the math task and math terminology were retained but the language was modified. One study (Abedi et al., 1998) of 1,394 Grade 8 students in schools with high enrollments of Spanish speakers showed that linguistic modification of the items contributed to improved performance on 49% of the items. Results indicated that students generally scored higher on shorter problem statements. Another study (Abedi et al., 2001) tested 946 Grade 8 students in math with different accommodations including linguistic modification, extra time, and glossary. Among these accommodations, only linguistic modification narrowed the score gap between ELL and non-ELL students.

Abedi and Lord's (2001) study of 1,174 Grade 8 students found small but significant score differences for students in low- and average-level math classes. Among the linguistic features that appeared to contribute to the differences were low-frequency vocabulary and passive-voice verb constructions (see Abedi et al., 1997, for a discussion of the nature of and rationale for the modifications).

In another study, Abedi, Courtney, and Leon (2003a) investigated 1,854 Grade 4 students and 1,594 Grade 8 students from 40 school sites using NAEP science items. No performance differences were seen in Grade 4 for the linguistic modification accommodation; however, differences were seen for Grade 8. The linguistically modified test version increased the performance of ELL students, but did not affect the performance of non-ELL students given the same accommodation.

Other studies have also employed language modification of test items. Rivera and Stansfield (2001, 2004) compared student performance on regular and simplified Grade 4 and Grade 6 science items. The study had a very small sample of ELL students, which did not show significant differences in their scores. However, the study did demonstrate that linguistic modification did not affect the scores of non-ELL students, indicating that linguistic modification was not a threat to score comparability.

In Francis et al.'s (2006) meta-analysis of accommodation research studies, the "simplified English" accommodation was shown to not have a significant effect on ELL students' performance. This study, however, was limited in that a number of studies were excluded from the analysis. The meta-analysis excluded studies not published in peer-

reviewed journals. Some of the studies included in the analysis had small sample sizes (e.g., Rivera & Stansfield, 2004). Another study, using Kansas state data, found items in "plain English" to actually be more difficult for students (Shaftel et al., 2003). That study, however, did not describe the procedures for creating "plain English" items. Results from research on linguistic modification can vary based on the methods of modifying the language, which is one of the criticisms this kind of research receives.

It is important to note that the linguistic modification approach also raises serious questions about the nature of the construct "academic ELP" discussed above. An additional question is the extent to which the language that is specific to the subject matter (e.g., *hypotenuse* for math, or *electron* for physics) is part of the construct being assessed, and thus the extent to which simplifying this language alters the construct.

Other accommodations. As previously mentioned, Rivera et al.'s (2006) review found 75 accommodations cited by state policies; however, only a selection of these have been investigated in empirical studies. Moreover, of the 75 accommodations, 31 were found to specifically address the needs of students with physical or cognitive disabilities and not the linguistic needs of ELL students. The GAO's review of 42 states (U.S. GAO, 2006) found the following accommodations to be most frequently cited: bilingual dictionary (32 states); reading items aloud in English (32); small group administration (29); extra time (27); individual administration (27); separate location (25); extra breaks (25); directions in student's native language (24).

Summary

In order to measure academic achievement as accurately and fairly as possible, the test accommodations sometimes used to level the playing field for ELL students are intended to reduce threats to test score validity. Various accommodations are used by states with some—extended time, dictionaries and glossaries, native language testing, dual-language test versions, linguistic modification—more commonly used than others. Despite a considerable body of research, findings are inconclusive on the effects of accommodations and the valid interpretations of scores from accommodated tests. For example, a number of studies suggest that accommodations related to linguistic modification are the most effective in reducing the gap between ELL students and non-ELL students. Although Francis et al.'s (2006) recent meta-analysis indicated that many of the linguistic modification accommodations had little or no effect on ELL students' performance; their analyses chose to exclude studies that were not published in peer-reviewed journals (e.g., Kiplinger et al., 2000; Maihoff, 2002). Other studies suggested promising potential for linguistic modification. As Koenig and Bachman (2004) contended, the existing research about accommodations is insufficient to provide

empirical support for many decisions associated with ELL students. States need to consider these issues when making decisions about accommodation use and interpreting the results from assessments with accommodation use. This area of concern continues to be one of active research and exploration.

Conclusion

Federal legislation under NCLB has brought issues of the assessment of ELL students to the forefront. NCLB mandates that states include ELL students in state and national content area assessments and report measurable yearly progress for ELL students in their English language development. However, these mandates regarding the assessment of ELL students have often not been met with systematic, research-based practice on the part of states, and many key issues regarding ELL testing and accountability systems remain unanswered.

The present document has discussed modern validity theory, which considers links between test scores and their interpretations for intended uses as well as their consequences. The validation process entails collecting various types of evidence to make an integrated validity argument. Articulating validation arguments for ELL assessments poses complexities due to the difficulties of defining the underlying language construct, the difficulties of differentiating content from language knowledge, and the varying characteristics of the ELL test taker population.

Among a variety of validity concerns, this document has focused primarily on reviewing issues related to defining the construct for ELP assessments and the effects of accommodations for ELL students in states' content area testing. These two key issues are interrelated in that states may use ELP test results for identifying ELL students' language needs for an academic setting and thus for determining an appropriate selection of accommodations. An investigation of the construct that an ELP test measures is a fundamental concern that has to be addressed. The current body of literature suggests that academic ELP is a construct that should be included in language proficiency assessments. As shown in this review, academic English is distinguished from general English in terms of lexical, syntactic, discourse, and pragmatic levels as well as its language functions. However, it is difficult to operationalize such a multifaceted construct for purposes of assessment, and firm consensus does not exist on its definition. This conundrum poses great challenges in developing an ELP test and validating the purposes of the test.

Although many states have begun to implement new ELP assessments, little research is available regarding their technical qualities or the adequacy and accuracy of ELP assessment

interpretations and uses. Given the inception of a new era of ELP assessments, evidence of validity is essential. Such evidence should be based on a strong, theory- and research-based framework defining the academic English construct. Sources of validity evidence should include analysis of the content of an ELP test, its alignment with the ELP and academic standards, the structure, reliability, and accuracy of the test, and the relationship between academic English measures, content test scores and other variables to demonstrate that specific uses of test results are justified. Results from this line of research will not only provide insights into how to operationalize the academic English construct to develop a sound ELP assessment, but may also provide evidence to support the validity of intended interpretations and uses of these assessments.

In relation to assessing academic performance in content areas, the use of accommodations is intended to assure the validity and reliability of ELL students' content test scores and interpretations. However, the research has yielded inconsistent findings and provides little evidence to assure valid procedures for applying accommodations. Considering the heterogeneity of ELL students in their language backgrounds, different levels of language proficiency, and language development rate, the challenge of establishing uniform and valid accommodation procedures are obvious. As Koenig and Bachman (2004) argued, comparing the scores of students who took accommodated and non-accommodated tests is not sufficient as a validity study, partly due to the different characteristics of the two groups. Just as for the ELP assessment, validity research on the effects of accommodations needs to provide various types of evidence based on the content of the test and the relationship between the accommodated and non-accommodated scores along with other external variables. This will provide information to justify the use of accommodations and improve our understanding of ELL students' content knowledge. In addition, although it is costly and time-consuming, investigating students' responses may provide a critical piece of information for the validity of test results.

Although more progress is needed to ensure the academic achievement of ELL students, much progress has been made in recent years, beginning with increased collaboration on the development of new ELP tests. Researchers and practitioners alike have been mobilizing to enhance the assessment and ultimately the educational achievement of ELL students. Also promising is a new taxonomy for testing ELL students, which is being developed by the University of Maryland, in collaboration with the South Carolina Department of Education. The Selection Taxonomy for English Language Learner Accommodations (STELLA) is a computerized decision-making system to help practitioners define and identify ELL students and match these students to the appropriate

accommodations (Kopriva & Carr, 2006; Zehr, 2007). The validity of this taxonomy system is, of course, a research agenda to be addressed.

It is evident that more research is needed to address issues in assessing ELP and academic performance of ELL students. Moreover, it is essential to review and analyze the current status of policies and practices that states are implementing for ELL assessments. The review of policy and practice will identify more specific and concrete issues to be considered in validation research. Scientific research-based uses of ELL assessments will improve the quality of assessments and assessment practices. In companion reports, we include both a review of current state practices and practical research-based implications and suggestions for practitioners to address the strengths and weaknesses of those practices associated with ELL students.

References

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

Abedi, J., Courtney, M., & Leon, S. (2003a). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CRESST Tech. Rep. No. 608). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Abedi, J. (2006). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-398). Mahwah, NJ: Lawrence Erlbaum Associates.

Abedi, J., Courtney, M., Leon, S., Kao, J., & Azzam, T. (2006). *English language learners and math achievement: A study of opportunity to learn and language accommodation* (CRESST Tech. Rep. No. 702). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodation for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CRESST Tech. Rep. No. 666). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (CRESST Tech. Rep. No. 536). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CRESST Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

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

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., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CRESST Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge, England: Cambridge University Press.

Bachman, L. F. (2006, September). *Issues in assessing the English proficiency and academic achievement of English language learners.* Paper presented at the MidWest Association of Language Testers Conference, Urbana-Champaign, IL.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice.* Oxford, England: Oxford University Press.

Bailey, A. L. (2007). *The language demands of school: Putting academic English to the test.* New Haven, CT: Yale University Press.

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

Bailey, A. L. & Butler, F. A. (2004). Ethical considerations in the assessment of the language and content knowledge of English language learners K-12. *Language Assessment Quarterly, 1,* 177-193.

Bailey, A. L., Butler, F. A., & Sato, E. (2005). *Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards* (CRESST Tech. Rep. No. 667). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey A. L., Butler, F. A., Stevens, R., & Lord, C. (2007). Further specifying the language demands of school. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 103-156). New Haven, CT: Yale University Press.

Ballard and Tighe Publishers. (2005). *IPT/PEM (IDEA proficiency test/Pearson Educational measurement) Title III testing solution.* Brea, CA: Author.

Bibian, G. C. (2006). *An ethnographic case study of a reconstituted urban middle school and the factors that contribute to improved ELL redesignation rates.* Unpublished doctoral dissertation, University of California, Los Angeles.

Brown, P. B. (1999). *Findings of the 1999 plain language field test* (Rep. No. T-99-013.1). Newark, DE: Delaware Education Research and Development Centers.

Butler, F. A. & Castellon-Wellington, M. (2000). *Students' concurrent performance on tests of English language proficiency and academic achievement.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., Stevens, R., & Castellon, M. (2007). ELLs and standardized assessments: The interaction between language proficiency and performance on standardized tests. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 27-49). New Haven, CT: Yale University Press.

California English Language Development Test: Form D. (2004). Monterey, CA: CTB/McGraw-Hill.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1(1),* 1-47.

Capps, R., Fix, M., Murray, J., Ost, J., Passel, J. S., & Herwantoro, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act*. Washington, DC: The Urban Institute. Retrieved July 6, 2007, from http://www.urban.org/publications/311230.html

Chamot, A. U. & O'Malley, J. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley.

Chiu, C. W. T. & Pearson, P. D. (1999, June). *Synthesizing the effects of test accommodations for special education and limited English proficiency students*. Paper presented at the National Conference on Large-Scale Assessment, Snowbird, UT. (ERIC Document Reproduction Service No. ED433362)

Collier, V. (1995). *Acquiring a second language for school: Directions in language and education, 1*(4). Washington, DC: National Clearinghouse for Bilingual Education.

Collier, V. & Thomas, W. (1989). How quickly can immigrants become proficient in school English? *The Journal of Educational Issues of Language Minority Students, 5,* 26-38.

Crockett, T., Hock, M., Kurtz, T., Magill, S., & Snider, M. A. (2007, June). *Validating the proficiency standards on an English language assessment: Results from three states*. Paper presented at the annual meeting of the Council of Chief State School Officers, Nashville, TN. Retrieved July 16, 2007, from http://www.ccsso.org/content/PDFs/57-Michael%20Hock.pdf

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3-49). Los Angeles: National Dissemination and Assessment Center.

Cummins, J. (1983). Language proficiency and academic achievement. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 108-126). Rowley, MA: Newbury House.

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon, England: Multilingual Matters.

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

Davidson, F., Kim, J. T., Lee, H., Li, J., & A. Lopez. (2007). Making choices in academic English language testing: Evidence from the evolution of test specifications. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 157-170). New Haven, CT: Yale University Press.

De Avila, E. A. & Duncan, S. E. (1990). *Language assessment scales, oral administration manual, English: Forms 2c and 2d*. Monterey, CA: CTB McGraw-Hill.

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

Del Vecchio, A. & Guerrero, M. (1995). *Handbook of English language proficiency tests*. Albuquerque: New Mexico Highlands University, Evaluation Assistance Center-West.

Douglas, D. (2000). *Assessing language for specific purposes: Theory and practice.* Cambridge, England: Cambridge University Press.

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

Durán, R. P. (1989). Assessment and instruction of at-risk Hispanic students. *Exceptional Children, 56*, 154-158.

Durán, R. P. & Lee, Y. (2007, June). *English language proficiency tests, One dimension or many? The WLPT II.* Paper presented at the annual meeting of the Council of Chief State School Officers, Nashville, TN.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments.* Portsmouth, NH: RMC Research Corporation, Center on Instruction. Retrieved November 21, 2006, from http://www.centeroninstruction.org/files/ELL3-Assessments.pdf

Gándara, P., Maxwell-Jolly, J., & Driscoll, A. (2005). *Listening to teachers of English language learners: A survey of California teachers' challenges, experiences, and professional development needs.* Santa Cruz, CA: The Center for the Future of Teaching and Learning.

Garcia, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly, 26,* 371-391.

Garcia, G., McKoon, G., & August, D. (2006). Language and literacy assessment of language-minority students. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners* (pp. 597-624). Mahwah, NJ: Lawrence Erlbaum Associates.

Geisinger, K. F. (2003). Testing students with limited English proficiency. In J. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 147-159). Greensboro, NC: CAPS Press.

Grissom, J. B. (2004). *Reclassification of English learners. Education Policy Analysis Archives, 12*(36). Retrieved October 31, 2006, from http://epaa.asu.edu/epaa/v12n36/

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

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

Hakuta, K., Butler, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Santa Barbara: University of California, Linguistic Minority Research Institute.

Haladyna, T. M. & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Halliday, M. A. K. (1978). *Language as social semiotic.* Baltimore, MD: University Park Press.

Hambleton, R. K. (2001). The next generation of the ITC test translation and application guidelines. *European Journal of Psychological Assessment, 17,* 164-172.

Herman, J. L. & Abedi, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics* (CRESST Tech. Rep. No. 633). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L. & Baker, E. L. (2005). Making benchmark testing work for accountability and improvement: Quality matters. *Educational Leadership, 63*(3), 48-55.

Herman, J. L., Webb, N., & Zuniga, S. (2003). *Alignment and college admissions: The match of expectations, assessments, and educator perspectives* (CRESST Tech. Rep. No. 593). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hopstock, P. J. & Stephenson, T. G. (2003a). *Descriptive study of services to LEP students and LEP students with disabilities. Special topic report #1: Native languages of LEP students.* Arlington, VA: Development Associates, Inc. Retrieved July 6, 2007, from http://www.ncela.gwu.edu/resabout/research/descriptivestudyfiles/native_languages1.pdf

Hopstock, P. J. & Stephenson, T. G. (2003b). *Descriptive study of services to LEP students and LEP students with disabilities. Special topic report #2: Analysis of Office for Civil Rights (OCR) Data Related to LEP students*. Arlington, VA: Development Associates, Inc. Retrieved July 6, 2007, from http://www.ncela.gwu.edu/resabout/research/descriptivestudyfiles/OCR2.pdf

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

Imbens-Bailey, A. & Castellon-Wellington, M. (1999, September). *Linguistic demands of test items used to assess ELL students.* Paper presented at the annual conference of the National Center for Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

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.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). *Measuring math—not reading—on a math assessment: A language accommodations study of English language learners and other special populations.* Presentation at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

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

Kopriva, R. & Carr, T. G. (2006, June). *STELLA: A computerized system for individually assigning test accommodations to ELLs*. Paper presented at the annual meeting of the Council of Chief State School Officers, San Francisco, CA.

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

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching (*8th ed.). Upper Saddle River, NJ: Prentice-Hall.

Liu, K., Thurlow, M., Erickson, R., Spicuzza, R., & Heinze, K. (1997). *A review of the literature on students with limited English proficiency and assessment* (Rep. No. 11). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

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 of Measurement in Education, Montreal, Canada.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*, 477-496.

Mohan, B. A. (1986). *Language and content.* Reading, MA: Addison-Wesley.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Office of English Language Acquisition (OELA), Language Enhancement, and Academic Achievement for Limited English Proficient Students. (n.d.). *The growing numbers of limited English proficient students: 1994/95-2004/05* [poster]. Washington, DC: U.S. Department of Education. Retrieved July 6, 2007, from http://www.ncela.gwu.edu/policy/states/reports/statedata/2004LEP/GrowingLEP_0405_Nov06.pdf

Office of English Language Acquisition (OELA), Language Enhancement, and Academic Achievement for Limited English Proficient Students. (2006). *California: Rate of LEP growth 1994/1995-2004/2005* [poster]. Washington, DC: U.S. Department of Education. Retrieved July 6, 2007, from http://www.ncela.gwu.edu/policy/states/reports/statedata/2004LEP/California-G-05.pdf

Olson, L. (2002, December 4). States scramble to rewrite language-proficiency exams. *Education Week, 22*(14), 10. Retrieved November 15, 2006, from http://www.edweek.com

Perie, M., Grigg, W., & Dion, G. (2005). *The nation's report card: Mathematics 2005* (NCES 2006-453). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Perie, M., Grigg, W., & Donahue, P. (2005). *The nation's report card: Reading 2005* (NCES 2006-451). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Porter, A. C. (2002). Measuring the content in instruction: Uses in research and practice. *Educational Researcher, 31*(7), *3-14.*

Rabinowitz S. & Sato E. (2006). *Technical adequacy of assessments fro alternate student populations: Technical review of high-stakes assessment for English language learners.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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., Collum, E., Shafer Willner, L., & Sia, J. K., Jr. (2006). An analysis of state assessment policies regarding the accommodation of English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective* (pp. 1-173). Mahwah, NJ: Lawrence Erlbaum Associates.

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. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment, 9*, 79-105.

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: The George Washington University, Center for Equity and Excellence in Education.

Scarcella, R. (2003). *Academic English: A conceptual framework* (Tech. Rep. No. 2003-1). Santa Barbara: University of California, Linguistic Minority Research Institute.

Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, *12,* 431-459.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. R., & Poggio, J. P. (2003). *The differential impact of accommodations in statewide assessment: Research summary.* Minneapolis: MN: University of Minnesota, National Center on Educational Outcomes. Retrieved July 10, 2007, from *http://education.umn.edu/NCEO/TopicAreas/Accommodations/Kansas.htm*

Short, D. J. (1993). Expanding middle school horizons: Integrating language, culture, and social studies. *TESOL Quarterly, 28*, 581-608.

Shuard, H. & Rothery, A. (Eds.). (1984). *Children reading mathematics*. London, England: J. Murray.

Sireci, S. G. & Khaliq, S. N. (2002). *An analysis of the psychometric properties of dual language test forms* (Center for Educational Assessment Res. Rep. No. 458). Amherst: University of Massachusetts, School of Education.

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 Res. Rep. No. 485). Amherst: University of Massachusetts, School of Education.

Solano-Flores, G., & Li, M. (2006). The use of Generalizability (G) Theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3-13.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*, 107-129.

Solomon, J. & Rhodes, N. C. (1995). *Conceptualizing academic language* (No. RR15): National Center for Research on Cultural Diversity and Second Language Learning.

Stack, J. (2002, September). *California English Language Development Test (CELDT): Language proficiency and academic achievement.* Paper presented at the Annual Conference of the National Center for the Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of ELLs* (CRESST Tech. Rep. No. 552). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Thurlow, M. L. (2001, April). *The effects of a simplified-English dictionary accommodation for LEP students who are not literate in their first language*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Thurlow, M. & Liu, K. (2001). *State and district assessments as an avenue to equity and excellence for English language learners with disabilities* (LEP Projects Report 2). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tippeconnic, J. W., III, & Faircloth, S. C. (2002). *Using culturally and linguistically appropriate assessments to ensure that American Indian and Alaska Native students receive the special education programs and services they need* (EDO-RC-02-8). Charleston, WV: ERIC Clearinghouse on Rural Education and Small Schools.

U.S. Department of Education. (2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001.* Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.

U.S. Government Accountability Office. (2006). *No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency (GAO-06-815)*. Washington, DC: Author. Retrieved July 20, 2006, from: http://www.gao.gov/new.items/d06815.pdf

Valdés, G. & Figueroa, R. (1994). *Bilingualism and testing: A special case of bias.* Norwood, NJ: Ablex.

Webb, N. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Washington, DC: Council of Chief State School Officers.

Wolf, M. K., Kao, J., Griffin, N., Herman, J. L., Bachman, P. L., Chang, S. M., et al. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses—Practice review* (CRESST Tech. Rep. No. 732). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wolf, M. K., Herman, J. L., Bachman, L. F., Bailey, A. L., & Griffin, N. (2008). *Recommendations for assessing English language learners: English language proficiency measures and accommodation uses—Recommendations* (CRESST Tech. Rep. No. 737). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Zehler, A. M., Hopstock, P. J., Fleishman, H. L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students.* Arlington, VA: Development Associates, Inc.

Zehr, M. A. (2007, February). Pilot program could help English-learners. *Education Week, 26*(22), 15-16. Retrieved August 10, 2007, from http://www.edweek.org/ew/articles/2007/02/07/22testing.h26.html

Appendix A:
NCLB Act (2002) Legislation Concerning Assessing ELL Students

Title I: Improving the Academic Achievement of the Disadvantaged

SEC.1111 (b) (2) (C) (v). Adequate yearly progress shall be defined by the State in a manner that includes separate measurable annual objectives for continuous and substantial improvement for each of the following; (II) the achievement of students with limited English proficiency.

SEC.1111 (b) (3) (C) (ix) (III) Academic assessments shall 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 administered to such students under this paragraph, including, to the extent practicable, assessment in the language and form most likely to yield accurate data on what such students know and can do in academic content areas, until such students have achieved English proficiency.

SEC.1111 (b) (7) Academic assessment of English language proficiency – Each State plan shall demonstrate that local educational agencies in the State will, beginning not later than school year 2002–2003, provide for an annual assessment of English proficiency (measuring students' oral language, reading, and writing skills in English) of all students with limited English proficiency…

Title III: Language Instruction for Limited English Proficient and Immigrant Students

SEC.3121 (d) (1) A State shall approve evaluation measures for use under subsection (c) that are designed to assess the progress of children in attaining English proficiency, including a child's level of comprehension, speaking, listening, reading, and writing skills in English.

SEC.3121 (d) (2) A State shall approve evaluation measures for use under subsection (c) that are designed to assess student attainment of challenging State student academic achievement standards on assessments.

SEC.3122 (a) (1). Achievement objectives – Each state educational agency or specially qualified agency receiving a grant under subpart 1 shall develop annual measurable achievement objectives for limited English proficient children served under this part that relate to such children's development and attainment of English proficiency while meeting challenging State academic content and student academic achievement standards.

SEC.3122 (a) (3) (A). Contents – Such annual measurable achievement objectives shall include (i) at a minimum, annual increase in the number or percentage of children making progress in learning English; (ii) annual increase in the number or percentage of children attaining English proficiency by the end of each school year, as determined by a valid and reliable assessment of English proficiency; (iii) making adequate yearly progress for limited English proficient children.

Appendix B:
Summary of Reviewed Studies Regarding Validity Evidence
(A list of acronyms is provided at the end of the table.)

Authors	Sample / Grades	Methods	Major Findings	Implications	Types of Validity Evidence
Butler & Castellon-Wellington (2000) http://www.cresst.org/products/reports/r663.pdf	California 778 - Grade 3 • 38% EO • 10% FEP • 52% LEP 184 - Grade 11 • 63% EO • 16% FEP • 21% LEP ELL status determined by test scores and the district at the time of student's arrival.	Correlations Multivariate analysis of variance (MANOVA) Assessments: An ELP test (LAS), Content test (Stanford 9)	Investigated the relationship between same-student performance on a standardized content assessment (Stanford 9) and an ELP test (LAS). Students differ on content test performance based on English language proficiency, as measured by the LAS. Differences of mean performance between EO, FEP, and LEP for every subtest of the Stanford 9 students were statistically significant.	LEP students are doing less well than the non-LEP students when tested. There is a strong relationship between the English language proficiency of ELL students and their performance on content assessments. When ELL student means are in the mid 90s on the LAS, those students' performance are similar to EO performance on content area tests. Thus, content assessments are likely valid measures of ELL students' content knowledge.	Content-related Interrelations (Criterion-related)

Authors	Sample / Grades	Methods	Major Findings	Implications	Types of Validity Evidence
Crockett, Hock, Kurtz, Magill, & Snider (2007) http://www.ccsso.org/content/PDFs/57-Michael%20Hock.pdf	New Hampshire ELL students 1843 - Grade 3–8 Rhode Island ELL students 4324 - Grade 3–8 Vermont ELL students 853 - Grade 3–8	Regression analysis Assessments: An ELP test (ACCESS for ELLs®) Content test (New England Common Assessment Program)	Examined the relationship between language proficiency measured by ACCESS for ELLs® and academic proficiency (measured by standardized content assessment). Scores of ACCESS for ELLs® predicted academic proficiency across grades and content areas in all three states.	The ACCESS for ELLs® test accomplishes its goal of measuring both social and academic language for all three states, across content areas and grade levels.	Interrelations (Criterion-related)
Durán & Lee (2007) http://www.ccsso.org/content/PDFs/48-Duran%20Lee%20yoonsun%20lee.pdf	Washington state ELL students 15,000 - Grades K–12	Confirmatory Factor Analysis Assessment: State ELP test (WLPT-II, 2006)	Examined a hypothesis that English language proficiency is on a single underlying continuum in Grades K–12. A model with a single factor with errors correlated within Reading, Writing, Listening, and Speaking subtests produced a good fit to the data.	There was no significant evidence to threaten construct validity by adding augmented items to the existing language test.	Internal structure (Construct-related)

Authors	Sample / Grades	Methods	Major Findings	Implications	Types of Validity Evidence
Grissom (2004) http://epaa.asu.edu/epaa/v12n36/	California EO and ELL students 693,821 - Grades 2–5	Regression analysis Assessments: CA standardized achievement tests (STAR) (1998–2002)	Examined the redesignation rates over time and factors to predict redesignation (e.g., gender, language, content test score). Content test score (Reading) was the best predictor of redesignation.	Academic achievement test scores can be used as one of the valid criteria for making a redesignation decision.	Interrelations (Criterion-related)
Herman &Abedi (2004) http://www.cresst.org/products/reports/r633.pdf	50 - Grade 8, Algebra class (pre-pilot tested and observed) 602 - Grade 8, Algebra class (surveyed) 271 - Grade 8, Algebra class (observed)	Multiple regression analyses HLM analyses Assessments: An ELP test Standardized achievement tests OTL measures • Surveys • Classroom observation	Examined whether opportunity to learn (OTL) variables predicted students' performance on an algebra achievement test. OTL made a more significant contribution in predicting student math performance for ELL students than EO students. HLM results indicated about 25% of variance of OTL was explained by students' language proficiency.	OTL in the classroom affects ELL students' performance on achievement tests. ELL students may receive different OTL than non-ELL students, which can compromise the interpretation of test scores.	Interrelations (Criterion-related)

Authors	Sample / Grades	Methods	Major Findings	Implications	Types of Validity Evidence
Herman, Webb, & Zuniga (2003) http://www.cresst.org/products/reports/tr593.pdf	Content experts (10 high school mathematics teachers and 10 faculty members from universities)	Expert review Kappa coefficients of agreement Correlation Assessment: Golden State Exam (GSE) High School Mathematics test	Examined the test items in terms of alignment between the content of items and standards. Depth of knowledge correlated positively with item complexity; item complexity did not correlate significantly with item difficulty. Overall there was good alignment between the Golden State Exam and the Statement on Competencies.	Alignment considerations should precede test development to ensure that intended purpose is measured.	Content
Jepsen & de Alth (2005) http://www.ppic.org/content/pubs/report/R_405CJR.pdf	California ELL students in K–12, 2002 and 2003	Hierarchical Linear Modeling (HLM) Assessments: State ELP test (CELDT, 2002, 2003) Standardized content tests	Examined the redesignation rate and its determinants in relation to school and student characteristics. At a school level, academic achievement test scores had good predictability of redesignation rates. Because school districts in CA have great latitude in weighing the state board of education's guidelines for reclassification, CELDT scores alone did not determine reclassification.	State policies aimed at improving CELDT performance, which include providing necessary resources to schools, will likely improve reclassification rates. Monitoring and modifying the reclassification process would be more effective and easier if districts applied state guidelines.	Interrelations (Criterion-related) Consequence

Authors	Sample / Grades	Methods	Major Findings	Implications	Types of Validity Evidence
Kopriva, Wiley, & Emick, (2007) http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/2b/65/9f.pdf	2,502 - Grades 3 & 5 ELL Levels per grade: Grade 3 • 52 Beginning • 198 Intermediate • 75 Advanced • 245 Exited • 711 Non-ELL Grade 5 • 46 Beginning • 148 Intermediate • 55 Advanced • 256 Exited • 719 Non-ELL Teacher ratings of students' abilities	Regression analysis Assessments: District benchmark mathematics test developed to mirror the state's large-scale assessment	Accommodations include modified mathematics test items that provide more access for ELL students (linguistic simplification, more accessible problem contexts, clearer formatting, use of pictures, etc.) and bilingual and picture-word glossaries. Examined the influence of accommodations and evaluated the effect of accommodations for the validity of score inferences across ELL students. ELL students were compared with native English speakers who were poor readers. Validity data from multiple-choice item type results for ELL students were generally poor compared to the control group, but comparable for constructed-response item types. For students classified as knowing some math on a criterion measure, analysis indicated there were significantly higher misclassification rates of test score data for lower English proficient ELL students compared to the control group.	Questions about validity of inferences drawn from large-scale assessments for students with lower English language proficiency should be raised. The effectiveness of measuring content knowledge with some item types for students of varying levels of English language proficiency should be questioned. Item type may interact with the grade level of test takers.	Interrelations (Criterion-related)

Authors	Sample / Grades	Methods	Major Findings	Implications	Types of Validity Evidence
Solano-Flores & Li (2006)	Florida, New York 170 ELL students identified by districts Grades 4 & 5 Haitian-Creole speakers	Generalizability (G)-theory Assessments: A set of 12 NAEP open-ended items in mathematics Four versions of the test; English, Haitian-Creole Standard, Haitian-Creole Dialect A and B	Examined where the source of measurement is found among rater, item, and code (dialect) of the test. The largest measurement error was found in the interaction of student, item, and code (dialect).	Not only a student's first language but also dialect variation should be considered in assessing ELL students' content knowledge.	Interrelations (Criterion-related) Reliability
Stevens, Butler, & Castellon-Wellington (2000) http://www.cresst.org/products/Reports/TR552.pdf	121 - Grade 7 • 19 EO • 102 ELL	Qualitative content review Correlations Item-response pattern analysis Assessments: An ELP test (LAS Reading component), Content test (Iowa Tests of Basic Skills [ITBS] Social Studies Test for Seventh Grade)	Analyzed the language of content and ELP tests, and examined the relationship between student performance on these two types of tests. The social studies test had more complex vocabulary (academic language) and sentence structure than the ELP test. An ELL student with a low language score likely did poorly on the content assessment, but an ELL student with high language proficiency did not necessarily do well on the content assessment.	An ELP test assessing social language ability may not be adequate to predict ELL students' readiness in a mainstream classroom. Opportunity to learn (OTL) and content knowledge may be more important factors on how well a student does on a content assessment than test scores on an ELP assessment.	Content-related Interrelation (Criterion-related)

List of Acronyms

ACCESS for ELLs® = Assessing Comprehension and Communication in English State-to-State for English Language Learners (a consortium test)

CELDT = California English Language Development Test

ELL = English Language Learner

ELP = English language proficiency

EO = English-only

FEP = Fully English Proficient

GSE = Golden State Exam

HLM = hierarchical linear modeling

LAS = Language Assessment Scales (a commercial test)

LEP = Limited English Proficient

NAEP = National Assessment of Educational Progress

OTL = opportunity to learn

STAR = Standardized Testing and Reporting (used in California)

WLPT-II = Washington Language Proficiency Test

Appendix C:
Summary of Reviewed Accommodation Studies
(A list of acronyms is provided at the end of the table.)

Authors/Link	Sample Size/ Grade Level	Major Findings
		Extended Time *(allowed in 38 states)*
Abedi, Lord, Hofstetter, & Baker (2000) See also: Abedi, Hofstetter, Baker, & Lord (2001) http://www.cresst.org/products/Reports/newTR536.pdf	946 - Grade 8	Investigated NAEP math items. Four accommodation types were investigated: • modified English • glossaries • extra time • glossary plus extra time Glossary plus extra time increased students' performance the most. Only extra time also increased student performance, for both ELL students and non-ELL students.
Chiu & Pearson (1999) (ERIC Document Reproduction Services No. ED433362)	N/A	This was a meta-analysis of 30 studies. Timing effects suggest that all students benefit from extended time, but an advantage for targeted population.
Hafner (2001) (ERIC Document Reproduction Services No. ED455299)	292 - Grade 4 159 - Grade 7	Extra time accommodation showed higher mean scores for both LEP and non-LEP. Mean scores were even higher for non-LEP.
Miller, Okum, Sinai, & Miller (1999)	601 - Grade 8	Study of high school proficiency exam in New Jersey. Accommodations investigated: • extra time • bilingual dictionary • dictionary plus extra time • Results were inconclusive.

Authors/Link	Sample Size/ Grade Level	Major Findings
Dictionaries *(dictionaries or glossaries or word lists are allowed in 43 states)*		
Abedi, Courtney, & Leon (2003a) http://www.cresst.org/products/Reports/R608.pdf	1,854 - Grade 4 1,594 - Grade 8	Investigated science items from NAEP and TIMSS. Three accommodation conditions: • customized English dictionary • bilingual dictionary • linguistic modification All conditions included extra time. Customized English dictionary did not significantly increase the scores in either grade for either group of students.
Abedi, Courtney, Mirocha, Leon, & Goldberg (2005) http://www.cresst.org/products/reports/r666.pdf	421 - Grade 4 190 - Grade 8	Investigated NAEP science items. Three conditions investigated: • published English dictionary • bilingual dictionary • linguistic modification Accommodations increased the mean scores of ELL students and did not affect the scores of non-ELL students. English dictionary was more effective for Grade 4. Linguistic modification was more effective for Grade 8. Published dictionary seemed to provide too much information.
Abedi, Lord, Boscardin, & Miyoshi (2000) http://www.cresst.org/products/reports/TR537.pdf	422 - Grade 8	Investigated NAEP science items. Two conditions: • English glossary with Spanish translations • customized English dictionary ELL students scored highest on customized English dictionary. Accommodations did not affect non-ELL students.

Authors/Link	Sample Size/ Grade Level	Major Findings
Glossary *(dictionaries or glossaries or word lists are allowed in 43 states)*		
Abedi, Courtney, & Leon (2003b) http://www.cresst.org/products/Reports/TR586.pdf	607 - Grade 4 542 - Grade 8	Investigated math items from NAEP and TIMSS. For Grade 4, four accommodations types were studied: • computer testing with pop-up glossary • extra time • customized English dictionary • small-group testing For Grade 8, two accommodations were studied: • computer testing with pop-up glossary • customized English dictionary Pop-up glossary was effective and valid for Grade 8, but not for Grade 4.
Abedi, Hofstetter, Baker, & Lord (2001) http://www.cresst.org/products/Reports/newTR536.pdf	946 - Grade 8	Investigated NAEP math items. Four conditions were investigated: • modified English • glossaries • extra time • glossary plus extra time Glossary plus extra time increased students' performance the most. Performance of both ELL and non-ELL students increased.

Authors/Link	Sample Size/ Grade Level	Major Findings
Dual-Language/Side-by-Side Bilingual Test Versions *(allowed in 9 states)*		
Abedi, Courtney, Leon, Kao, & Azzam (2006) http://www.cresst.org/products/reports/R702.pdf	2,321 - Grade 8	Investigated math items from NAEP and TIMSS. Two accommodation types: • dual-language (English and Spanish) test • linguistic modification Accommodations did not affect student performance.
Duncan, del Rio Parent, Chen, Ferrara, Johnson, Oppler, & Shieh (2005) http://www.leaonline.com/doi/abs/10.1207/s15324818ame1802_1	402 - Grade 8 tested (68 in focus groups)	NAEP math items were presented in a dual, side-by-side English and Spanish format. No differences in performance were detected. However, Spanish-speaking students reported finding the dual-language test as useful or very useful.
Linguistic Modification *(this accommodation is not mentioned in existing state policies)*		
Abedi, Courtney, & Leon (2003a) http://www.cresst.org/products/Reports/R608.pdf	1,854 - Grade 4 1,594 - Grade 8	Investigated science items from NAEP and TIMSS. Three accommodation conditions: • customized English dictionary • bilingual dictionary • linguistically modification All conditions included extra time. In Grade 8, linguistically modified version was the only accommodation to significantly impact performance. (No differences were seen in Grade 4.)
Abedi, Hofstetter, Baker, & Lord (2001) http://www.cresst.org/products/Reports/newTR536.pdf	946 - Grade 8	Investigated NAEP math items. Four conditions were investigated: • modified English • glossaries • extra time • glossary plus extra time No significant results were found with the modified English test version.

Authors/Link	Sample Size/ Grade Level	Major Findings
Linguistic Modification *(this accommodation is not mentioned in existing state policies)*		
Abedi & Lord (2001) http://www.leaonline.com/doi/abs/10.1207/S15324818AME1403_2	1,174 - Grade 8	NAEP math items linguistically modified. Increases in scores were seen for low-performing students (which includes ELL students).
Abedi, Lord, & Hofstetter (1998) http://www.cresst.org/products/Reports/TECH478.pdf	1,394 – Grade 8	Three test booklets were administered: original English, linguistically modified English, and original Spanish. LEP students performed best on linguistically modified English.
Abedi, Lord, & Plummer (1997) http://www.cresst.org/products/Reports/TECH429.pdf	1,031 – Grade 8	NAEP math items were linguistically modified. Students in low and average math classes scored higher in the linguistically modified version, but results were not significant.
Shaftel, Belton-Kocher, Glasnapp, & Poggio (2003) http://education.umn.edu/NCEO/TopicAreas/Accommodations/Kansas.htm	617 ELL students - Grades 4, 7, & 10	Original test forms and plain English test forms. Items in plain English were more difficult for students than in original form.
Rivera & Stansfield (2001; 2004) 2001: http://www.doe.k12.de.us/AAB/aera%20linguistin%20simplification.pdf 2004: http://www.leaonline.com/doi/pdf/10.1207/s15326977ea0903&4_1	11,306 non-LEP and 109 LEP in Grades 4 and 6	Linguistic simplification did not affect the performance of non-LEP in science. The sample size for LEP was too small for meaningful results.

List of Acronyms

ELL = English Language Learner, LEP = Limited English Proficient, NAEP = National Assessment of Educational Progress, TIMSS = Third International Mathematics and Science Study