Classification System for English Language Learners: Issues and Recommendations

Jamal Abedi, *University of California, Davis*

High-stakes decisions for the instruction and assessment of English language learner (ELL) students are made based on the premise that ELL classification is a valid dichotomy that distinguishes between those who are proficient in the use of the English language and those who are not. However, recent research findings draw a vague picture of the term "ELL" and call for a more valid classification system for ELL students. Thus, the purpose of this paper is twofold: (1) to reveal issues concerning the validity of the current ELL classification system based on the results of several empirical studies, and (2) to initiate a discussion on ways to improve the validity of the ELL classification system by proposing a system that uses existing multiple criteria in a stepwise manner. While the suggested system has its own limitations and controversies, we hope this discussion stimulates thoughts and brings much needed attention to this very important national issue.

Keywords: English language learner, validity, ELL classification, assessment, No Child Left Behind

Recent federal legislation, such as the Improving America's Schools Act of 1994 and the No Child Left Behind Act (NCLB) of 2001, address the need to advance the quality of teaching and learning for *every* child, including those who are English language learners (ELL).[1] On the other hand, research on fair, valid, and effective assessment has brought into question existing ELL classification policies and practices. Improper classification may render assessment results unfair, invalid, and ineffective, which may lead to inappropriate and inadequate instruction for ELL students. Validity problems in ELL classification and assessment may also affect accountability, such as in reporting Adequate Yearly Progress (NCLB, 2002) for ELL students. Misleading results of inaccurate classification and invalid assessment may lead to disproportionately placing ELL students in special education classrooms where it may negatively affect their academic career and may take them a longer time to graduate (Stefanakis, 1998).

Improved validity and consistency in classification of ELL students is of utmost importance as these students continue to be a fast-growing population. According to a recent report by the U.S. Government Accountability Office, about 5 million ELL students were enrolled in schools, representing approximately 10% of all public school students (GAO, 2006). Between 1990 and 1997, the number of U.S. residents born outside the United States increased by 30%, from 19.8 million to 25.8 million (Hakuta & Beatty, 2000). Approximately 1.6 million in the state of California alone are considered English learners (Gándara, Maxwell-Jolly, & Driscoll, 2005). This rapid growth demands that we consistently and accurately determine which students require English language services (Abedi & Gandara, 2006). To make this determination, we first need to accurately identify ELL students.

A discussion of the ELL classification system must consider the validity of information collection methods and how information can be used more effectively. The purpose of this paper is twofold: first, to bring issues concerning validity of current ELL classification to the attention of assessment experts, researchers, educational practitioners and policymakers; and second, to initiate a discussion on how existing information can effectively be used to improve the validity of the ELL classification system.

Revealing the Issues

For the purposes of this paper, students who are not considered ELL are referred to as "non-ELL." The non-ELL group usually consists of native English speakers, students from non-English-speaking homes who are fluent in English at the time of school entry (initially fluent English proficient, IFEP), and students who progress out of the ELL category (redesignated fluent English proficient, RFEP, National Clearinghouse for English Language Acquisition, 2002).

One would expect a uniform approach in assigning an ELL classification code (1 = ELL, 0 = non-ELL) to students across the nation. If states or at least school districts within a state have adopted a uniform definition, then one would expect a student who is classified as ELL at one school to be similarly classified at another school. However, results from several studies have suggested otherwise (see, for example, Abedi, Lord, Hofstetter, & Baker, 2000). A review of data from 12 schools revealed different systems for determining a student's level of English

Jamal Abedi is a Professor, University of California, Davis, School of Education, One Shields Avenue, Davis, CA 95616; jabedi@ucdavis.edu.

proficiency (Abedi, Lord, & Plummer, 1997). Additionally, Rivera, Stansfield, Scialdone, and Sharkey (2000) found in their review of state policies that the ELL definition provided by about half of the participating states differed widely in content. Linquanti (2001) found that the criteria used for initially classifying language-minority students as ELL in California—largely based on English language proficiency (ELP)—are different from the multiple criteria (linguistic and academic) used to reclassify them as RFEP, and that these latter criteria vary across districts within the same state.

Cisneros and Leone (1995) believed that determining the exact number of ELL students in elementary and secondary schools was not easy, since ELL definitions vary so widely from state to state. They indicated:

> Due to the broad definition of "limited English proficient" in the Bilingual Education Act (BEA) and lack of clearly outlined procedures for identifying ELL students, future reauthorization of federal legislation will need to define such terms and clearly outline procedures for identification of ELL students . . . . (p. 362)

In search of a model for classification, one might ask what the national criteria are for including ELL students in large-scale assessments. The National Assessment of Educational Progress (NAEP) does not directly define "ELL," but includes those ELL students who participate in the regular state assessments (NAEP, 2007). Thus, NAEP's identification of ELL students is based on states' classification policy.

The definition of an ELL [LEP] student, as outlined in the No Child Left Behind Act, Title IX #25 (NCLB, 2002) is: (a) age 3 through 21; (b) enrolled or preparing to enroll in an elementary or secondary school; (c) not born in the United States or whose native language is not English; (d) is a Native American, Alaskan Native, or a native resident of the outlying areas; (e) comes from an environment where a language other than English has had a significant impact on an individual's level of ELP; (f) is migratory and comes from an environment where English is not the dominant language; or (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 and the ability to successfully achieve in classrooms where English is the language of instruction, or to participate fully in society (NCLB, 2002, Title IX).

The above definition is primarily based on two sources of information: (1) students' language background information and (2) their level of English proficiency. Information on the language background of students (for example, country of birth, native language, and type and amount of a language other than English spoken at home) comes typically from a parent-completed Home Language Survey (HLS) which is described below. Information on the students' level of English proficiency in speaking, reading, writing, listening, and comprehension comes from existing tests of English proficiency. However, research shows major concerns with the reliability and validity of these sources of information.

Home Language Surveys

The HLS determines which students should undergo English language assessment and possibly receive instruction designed for speakers of other languages. The HLS (which may differ across states in term of format and type of questions) is usually used just to identify linguistic minority (i.e., potential ELL) status. The main purpose of the survey is to identify what languages are spoken at home. Some school districts require that the HLS be administered to families of every entering student and that those results be included in every student's permanent file. According to the *Survey of the States' ELL Students*, over 80% of schools made use of some form of HLS (Kindler, 2002). Unfortunately, the validity of HLS data could become questionable. Parents may give inconsistent information for a variety of reasons, including concerns over equity of opportunity for their children, citizenship issues, and poor comprehension of the survey form or interview (Abedi et al., 1997; Littlejohn, 1998).

Littlejohn (1998) questioned the validity of HLS information. He used as an example a student who has always spoken English but who had a relative in the home for a period of time who spoke Spanish would, under the Office for Civil Rights' (OCR), be classified as "PHLOTE" (primary home language other than English) (p. 8). On the other hand, some parents claimed that English was the language spoken at home because their children practiced English as a second language at home. Littlejohn indicated that this was enough of a problem that the Denver Office of Education added the cautionary note of "Do not list languages learned or used only academically" (p. 10).

As suggested above, in addition to the information from an HLS, student assessment outcomes from both content-based and ELP tests are also used as criteria for ELL initial classification (especially in later grades) and reclassification. Below is a short review of these criteria that are used for multiple purposes including for classification of ELL students.

English Language Proficiency Tests

In addition to the HLS, ELP tests are commonly used for identifying ELL students. The *Survey of the States' LEP Students* revealed that 94% of those surveyed used some type of ELP test for ELL classification and placement (Kindler, 2002). While this more objective criterion is a valuable addition to the ELL classification process, there are several major concerns with the conceptual framework and psychometric characteristics of many ELP tests.

ELP tests can be grouped in two different categories: tests prior to the implementation of NCLB (pre-NCLB) and tests that were newly developed based on the NCLB Title III requirements (post-NCLB). To ensure accurate classification of ELL students, it is necessary to examine the validity of ELP assessment outcomes (for both pre- and post-NCLB tests) as criteria for ELL classification. In doing so, we first need to understand the theories behind English language acquisition and then examine validity issues concerning these assessments. We can then demonstrate how improvements in ELP assessments affect the validity of ELL classification.

Theories of Second Language Acquisition

Understanding the process of language acquisition for ELL students is essential to developing a more valid assessment and classification system for these students. However, different views and theories on second language acquisition complicate the issue (Conteh-Morgan, 2002; Francis & Rivera, 2007; Reutzel & Cooter, 2007). The behavioral theorists believe that language development is influenced by environmental stimuli, such as imitation, rewards,

and practice. On the other hand, the innatist theorists believe that learning is a natural process through a human built-in device for learning language. For example, according to Chomsky (1968) language is modeled by internal factors and then shaped through experience. Similarly, Krashen (1988) suggests that humans are born with the ability to learn language (see also Lightbown & Spada, 2000).

Critics of the innatist theory argue against the claim that internal factors fully explain language acquisition process. They believe that environmental factors such as exposure to rich learning environments and interaction with others influence language acquisition (Reutzel & Cooter, 2007). These cognitive theorists believe that the process of language acquisition may in turn affect cognitive and social skill development (Reutzel & Cooter, 2007). Finally, the social interaction theorists believe that language acquisition is impacted by many different factors including cognitive, linguistic, social, and physical ones (Reutzel & Cooter, 2007).

As reflected in many of these well-known theories, the process of language acquisition for all children, including ELLs, is influenced greatly by environmental factors including opportunities to learn and practice at home, at school, and in society. Research literature clearly links environmental factors, such as these rich learning opportunities, with proficiency in English. For example, the number of years students live in the U.S. interacting with native speakers of English (Hakuta, Butler, & Witt, 2000), the number of English-only classes, and student level of proficiency in native language affects students' proficiency in English, helping them to be reclassified as proficient in English. Such information, if collected properly, could improve the validity of the ELL classification system substantially. More importantly, to improve the validity of the classification system for ELL students, teachers, and school officials, including bilingual and ESL/ELD coordinators, must be familiar with students' language needs and backgrounds in order to use the criteria for the ELL classification system properly. Therefore, incorporating such important information into the ELL classification system will help improve the validity of this system. We now discuss ELP assessments that were developed before and after implementation of NCLB.

Pre-NCLB English Language Proficiency Assessments

The pre-NCLB assessments were developed by different organizations at different times based on different needs and requirements. There are major limitations with many of these assessments. First and foremost are the discrepancies in the theoretical bases of these tests. The tests are based on one or more of at least three different schools of thought: (1) the discrete point approach, (2) the integrative or holistic approach, and (3) the pragmatic language testing approach (Del Vecchio & Guerrero, 1995). Consequently, the tests provide very different outcome measures. For example, Valdes and Figueroa (1994) indicated that:

> As might be expected, instruments developed to assess the language proficiency of "bilingual" students borrowed directly from traditions of second and foreign language testing. Rather than integrative and pragmatic, these language assessments instruments tended to resemble discrete-point, paper-and-pencil tests administered orally. (p. 64)

Second, a distinction exists between basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP) (Bailey & Butler, 2003; see also Cummins, 2000). In the context of assessments, language proficiency tests could vary in the extent they gauge CALP. Bailey and Butler (2003) defined academic language as "language that stands in contrast to the everyday informal speech that students use outside the classroom environment" (p. 9). In other words a student could score high in BICS but low in CALP. Therefore, it is necessary to determine that language proficiency tests adequately measure the type of language proficiency needed to be successful in mainstream English classrooms.

Zehler, Hopstock, Fleischman, and Greniuk (1994) compared English proficiency test content and structure (productive skills, receptive skills, and reading skills), the test administration procedures, the theoretical bases of the tests, and issues related to the validity and reliability of the tests. They found major differences in all of the areas in which the tests were compared. They also found that tests differed in their approaches to defining language proficiency, the types of tasks and specific item content, the grade level ranges, and the specific time limits (see also, Rossell, 2000).

Del Vecchio and Guerrero (1995) also presented a comprehensive review of some of the commonly used ELP tests prior to NCLB: (1) Basic Inventory of Natural Language (BINL), (2) Bilingual Syntax Measure (BSM), (3) Idea Proficiency Test (IPT), (4) Language Assessment Scales (LAS), and (5) Woodcock–Munoz Language Survey (WMLS). They found major differences between these tests with respect to their purpose, age and language group, administration, cost, items, scoring, test design, theoretical foundation, reliability, and validity of the tests. Such wide ranging disparities in these ELP assessments are a significant cause for concern with regard to the accuracy and consistency of the measures used to classify ELL students.

Post-NCLB English Language Proficiency (ELP) Assessments

NCLB required schools receiving Title I funding to annually assess ELL students' level of ELP using reliable and valid measures. For example, NCLB requires that ELP assessments include four modalities (reading, writing, speaking, and listening), incorporate the concept of academic language, and align the content of ELP assessments with the states' ELP standards.

Four consortia of states carried out the challenging task of developing post-NCLB assessments based on the NCLB Title III requirements (see Abedi, 2007). Test items were aligned with the states' ELP content standards and standard setting was conducted to set language proficiency levels in several categories typically distinguishing beginning, intermediate, proficient, and advanced in all four modalities. ELP tests were often developed for four or more grade clusters (typically K-2, 3–5, 6–8, and 9–12) and included common sets of items across adjacent grade clusters. The newly developed assessments underwent extensive pilot and field testing on large and representative samples of students. The content and psychometric properties of the individual items as well as the total tests were carefully examined and improvements were made where needed.

While these efforts have helped establish a strong foundation for the newly developed ELP assessments, we believe there are still some unresolved issues concerning such assessments

that may impact classification of ELL students. These issues include:

ELP standards
NCLB requires the newly developed ELP assessments to be aligned with state ELP content standards. This poses several concerns. The term *ELP* has not been clearly defined in the literature. Furthermore, many states did not have a set of defined ELP content standards prior to the implementation of NCLB, and it was technically challenging for the consortia to develop a set of standards that are truly common across the participating states in a given consortium of ELP assessment (see, for example, Fast, Ferrara, & Conrad, 2004).

Standard setting
The newly developed ELP assessment consortia conducted standard setting to create language proficiency levels for ELL classification purposes. In addition to the sources of inconsistencies due to the use of different standard-setting approaches and subjectivity involved in the standard-setting process (Giraud, Impara, & Plake, 2005; Jaeger, 1989), other factors may introduce bias into the standard-setting process. For example, inconsistencies in the proficiency levels at the different modalities may affect decisions on classifications of ELLs (see, for example, Bunch, 2006). To illustrate this point, assuming a student was rated as *proficient* in reading and writing but as *below proficient* in listening and speaking, how will this student be rated on the overall proficiency scale? Or, should the level of proficiency of this student be judged based on the total test of all four modalities? If so, then should there be evidence of unidimensionality of the test?

Dimensionality
In addition to the scores from each of these four modalities/subscales (reading, writing, listening, and speaking), composite scores of all subscales as well as other composites are commonly used by states. If the four modalities are highly correlated and if they measure a single construct, the decision to combine the different subscale scores would be less complicated than when the subscales are not measuring the same construct. Therefore, the issue of dimensionality needs to be addressed prior to deciding whether to use subscale or total scores.

The baseline for the NCLB Title III assessment
Since the newly developed ELP assessments were not available at the start of NCLB implementation, states had no other choice but to use whatever existing ELP assessment they found to be relevant for their state. Now that many states have access to new ELP assessments that meet NCLB requirements, they are faced with the quandary of linking ELP assessment results from "off-the-shelf" tests as the *baseline* with the results from their new ELP assessments. The problem is not limited to the tests having different domains of ELP content. Many of the existing ELP tests at the start of NCLB implementation were based on different theoretical emphases prevalent at the time of test development. They were also not aligned with states' ELP content standards and were not based on the concept of academic language. Therefore, even a high correlation between ELP assessments used as the baseline and the new ELP assessment would not be enough to assume a strong link between the two.

Academic English
Clearly, the focus of NCLB Title III ELP assessment is on "academic English." Therefore, many of the newly developed measures of ELP are based on academic English to facilitate learning content knowledge. However, concerns remain as to whether ELP assessment should be focused on the language of the content areas (such as mathematics and science) or the language that facilitates content learning. Fast et al. (2004) indicate that ELP assessments "are not tests of academic content, in other words, no external or prior content-related knowledge is required to respond to test questions" (p. 2). Given these concerns, test item writers for the ELP assessments are not quite certain what constitutes "*academic English"* and how it should be captured within the ELP assessments. (For a more detailed discussion of the pre- and post-NCLB ELP assessments see Abedi, 2007.)

Standardized Academic Achievement Tests

Results from standardized achievement tests which are required for all students are also used in conjunction with scores from language proficiency tests to identify but mainly to reclassify ELL students. The *Survey of the States' LEP Students* revealed that achievement tests were used by 76% of the states surveyed to help identify or reclassify ELL students (Kindler, 2002). Critics believe these tests are not designed for this purpose; rather, they are designed only to assess monolingual English students' content knowledge (Rossell, 2000).

Stefanakis (1998) indicated that a major concern in the assessment of ELL students is the lack of standardized achievement tests specifically designed to assess the content knowledge of these students. Mahoney and MacSwan (2005) argued that the use of academic tests for identifying and reclassifying ELL students is inappropriate. A review of standardized achievement tests by Zehler et al. (1994) found major differences between these tests across different areas, including their content, format, and psychometric characteristics.

Not only do the differences between tests produce inconsistent classification and assessment results for ELL students, but the unnecessary linguistic complexity of many achievement test items that are developed for native speakers of English casts doubt on the validity and reliability of these assessments when used for ELL students. For example, based on the results of many studies on the assessment of ELL students, Abedi (2006a) indicated that unnecessary linguistic complexity of test items may be an additional source of measurement error in using standardized achievement tests for ELL students (see also Figueroa, 1989, 1990; Valdes & Figueroa, 1994). This may seriously undermine the validity of inferences addressed by the assessment because it is a source of construct-irrelevant variance (see also Haladyna & Downing, 2004; Messick, 1994). Additionally, Solano-Flores and Trumbull (2003) found that language factors interact with test items. That is, items that are linguistically complex contribute largely to the measurement error variance observed for ELL students, leading students to misinterpret and misunderstand test questions.

In addition to the content and psychometric concerns with using standardized academic achievement tests as an index for ELL classification, there is disagreement on the level of student performance below which students are considered as ELL. For example, Grissom (2004) reported a cutoff score on the standardized norm-referenced test (NRT) at the 36th percentile point or

above in order for students in California to be RFEP. Gándara (2000) indicated that, "LEP in California is a child who does not understand sufficient English to pass a test of oral proficiency and *who does not score above the 35th to 40th percentile on an English standardized test*" (p. 3, emphasis added). Linquanti (2001) documented reclassification cutscores used in seven California districts that ranged from the 33rd percentile to the 40th percentile, with some utilizing these for reading, language, and/or math NRT sections. A report by the United States General Accounting Office (2001) observed a disagreement about appropriate standards for measuring English proficiency. The report interpreted the age/grade appropriate level as scoring above the 50th percentile on standardized achievement tests but also acknowledged that "some states consider students English proficient when they score at the 40th percentile or even at the 32nd percentile" (p. 14).

As can be seen from the short summary of research presented above, there is no specific indication of which tests or which cutoff score would indicate an acceptable level of English proficiency. Classifying language proficiency by arbitrarily setting a cutoff point on standardized academic achievement test scores (for example in reading/language arts) is also not a good practice since there are large numbers of native English speakers who score below these cutoff points. Should these students also be considered ELL? If the answer to this question is "Yes," then the concept and operational definition of ELL classification becomes even more controversial. On the other hand, if the answer is "No," then one must ask if low-scoring, native English speakers can truly be considered language proficient, classified as "non-ELL," and be deprived of the additional language skill development they deserve. Further, the language intervention strategies would need to be significantly different for these students than for those whose native language is not English.

Empirical Evidence on the Validity of ELL Classification

To examine the validity of current ELL classification, we present empirical data on some of the most commonly used criteria in the classification/reclassification of ELL students. Students' background characteristics, such as students' socioeconomic status (SES) and ethnicity are not directly used as criteria for ELL classifications, but because they have been found to correlate with students' academic performance, these variables may also need to be acknowledged as predictors of ELL classification outcome. Finally, we present some data to help us provide research-based recommendations on how to improve the validity of ELL classification.

Our presentations are based on data from several randomized field studies conducted at the National Center for Research on Evaluation, Standards, and Students Testing (CRESST, see for example, Abedi, 2006a, 2006b; Abedi et al., 2000; Staley, 2005) and analyses of existing data from seven locations nationwide. Table 1 presents summary information, including testing year, grade levels, student population, and type of test for the sites that provided comprehensive databases for analyses. Because of confidentiality agreements with the data providers, state and test names will not be mentioned in this paper but may be revealed by the written permission of the providers.

The sites were in the United States and varied in location and population. The student background variables included gender, ethnicity, free/reduced-price lunch participation, parent education, student ELL status, and Students with Disabilities (SD) status. Item-level standardized achievement test data were also obtained. However, the sites differed in the standardized tests used, the type of language proficiency index used, and the type of background variables provided. Comparisons of the results across the data sites provided cross-validation information. To obtain information on the consistency of results over time, we also included a few locations with very recent data. Of the seven sites, four provided assessment data from 1997 to 1998 (pre-NCLB) and the three others provided data for 2005–2006 (post-NCLB).

Empirical Data Presentation

Home Language Survey

Earlier in this paper, we cited studies that question the validity of the Home Language Survey, a commonly used

Table 1. Site Summary

Site/Grade	Data Year	Number of Students	Number of ELL Students	Percent of ELL Students	Tests Used
Site 1	1998–1999				NRT
Grade 3		36,065	7,270	20.20%	
Grade 6		28,313	3,341	11.80%	
Grade 8		25,406	2,306	9.00%	
Site 2	1997–1998				NRT
Grade 2		414,169	125,109	30.20%	
Grade 7		349,581	73,993	21.20%	
Grade 9		309,930	57,991	18.70%	
Site 3	1997–1998				NRT
Grade 10		12,919	431	3.30%	
Grade 11		9,803	339	3.50%	
Site 4	1997–1998				NRT
Grade 3		13,810	1,065	7.70%	
Grade 6		12,998	813	6.30%	
Grade 8		12,400	807	6.50%	
Site 5	2005–2006				CRT
Grade 5		33,242	5,008	15.10%	
Grade 8		33,106	3,870	11.70%	
Site 6	2005–2006				CRT
Grade 4		102,574	4,219	4.10%	
Grade 8		107,695	3,456	3.20%	
Site 7	2005–2006				CRT
Grade 4		55,724	7,090	12.70%	
Grade 8		52,900	3,026	5.70%	

Note: Data on ELL students from Site 1 is for students receiving bilingual services.

criterion for ELL classification decisions. The findings of our empirical data share similar concerns. In one study, Abedi et al. (1997) formulated the Language Background Questionnaire (LBQ) based on the HLS concept and administered it to 1,031 eighth-grade students. The LBQ included questions about languages other than English spoken at home, the number of years the student had lived in the United States, and the number of English-only classes taken. Students' responses to the LBQ were compared with school rosters reporting the students' official primary languages as identified by the parents on the district's Home Language Survey and their ELL classification where appropriate. Significant discrepancies were revealed, making the accuracy of this single source of language background data highly questionable. In many schools the record of students speaking a language other than English at home, regardless of ELL classification, was significantly lower than what the students reported in the LBQ (see also Abedi et al., 2000).

Language Proficiency Test Scores in Determining ELL Classification

Since ELL classification should distinguish between students who are proficient in English and those who are not, one would assume a high level of association between these two variables. However, research findings do not support this assumption. The lack of a strong relationship between English proficiency test scores and ELL classification may be partly due to content and psychometric shortcomings of the tests but are mainly due to validity issues in defining the ELL/non-ELL dichotomy.

To illustrate the relationship between ELL classification codes and ELP scores, we selected two data sites: one provided data from ELP tests that were developed prior to NCLB (Site 2) and one provided data from the ELP test developed based on the NCLB Title III requirements (Site 7). A comparison of the relationship indices between these two sites may reveal some information on the possible improvement of ELL classification based on the new ELP assessments.

One of the most commonly used ELP tests developed prior to the implementation of NCLB (Loop, 2002) was used for this purpose. A district within Site 2 provided an excellent opportunity to illustrate the power of this test in defining the ELL/non-ELL dichotomy. The test was administered to both ELL and non-ELL students. Performance differences between ELL and non-ELL students were estimated in terms of effect sizes (Cohen, 1988; Kirk, 1995). Additionally, the percent of variance of ELL classification categories explained by the test was also computed and was labeled as ω^2 (see Kirk, 1995, pp. 177–180). Table 2 presents ω^2 as well as effect sizes when comparing the performance of ELL and non-ELL students in grades two through twelve for a district in Site 2. The proportions of variance of ELL/non-ELL explained by the test scores (ω^2) ranged between .03 (3% of the variance for Grade 12) to .09 (9% of the variance for Grade 10), which were not large enough to suggest a strong association between English proficiency test scores and ELL classification. The effect sizes ranged between .179 (for students in Grade 12) to .319 (for students in Grade 10) with an average effect size of .239. Based on Kirk (1995) this average effect size is considered small. Thus, the results do not support the notion that the ELP test score explains much of the variance of the ELL/non-ELL dichotomy.

To compare the association between the pre- and post-NCLB assessments and the ELL dichotomy, we also used data from one of the more recent test administrations. Site 7 administered one of the post-NCLB English language proficiency assessments to both ELL and non-ELL students. At this site the non-ELL students consisted of ELL students who were reclassified as fluent English proficient (FEP). Table 2 presents the results of analyses for Site 7 as well. Unlike Site 2 for which data were available for Grades 2 through 12, for Site 7 we had access to data only for Grades 4 and 8.

As data in Table 2 show, for Grade 4 the proportions of variance of ELL/non-ELL dichotomy explained by the newly developed ELP scores (ω^2) is .142 with an effect size of .407 ($n = 7{,}957$). Comparing with the similar statistics obtained for Site 2 ($\omega^2 = .035$ with an effect size of .190), the level of the relationship between the new ELP and ELL classification code is much stronger with the new ELP assessments. A similar association was found for students in Grade 8. The proportion of variance in ELL dichotomy explained by the new ELP measure was .104 with an effect size of .341, compared with a ω^2 of .064 and an effect size of .260 for Site 2. The average effect sizes of the association between ELP assessments and ELL dichotomy across the two grades for the new test is .374, which according to Kirk (1995) is medium to high as compared with an average effect size of .239 for pre-NCLB which is considered small.

Standardized Academic Achievement Test Scores in Determining ELL Classification

Academic achievement tests were used by 76% of the states surveyed in defining the ELL/non-ELL dichotomy (Kindler, 2002). As noted above, research has identified major sources of construct-irrelevant variance with these tests when administered to ELL students. To illustrate the power of achievement

Table 2. Omega-Square, Effect Size, and Number of Students for Sites 2 and 7

Site 2				Site 7			
Grade	Omega Square	Effect Size	Number of Students	Grade	Omega Square	Effect Size	Number of Students
2	.050	.229	587				
3	.038	.199	721				
4	.035	.190	621	4	.142	.407	7,957
5	.040	.203	1,002				
6	.050	.230	803				
7	.068	.270	938				
8	.064	.260	796	8	.104	.341	4,364
9	.070	.275	1,102				
10	.092	.319	945				
11	.074	.283	782				
12	.031	.179	836				

Table 3. Site 1 and 6 Omega-Square, Effect Sizes, and Number of Students for NRT and CRT Test Scores and ELL Classification

	Site 1			Site 6	
	Reading	Math		Reading	Math
Grade 3			Grade 4		
Omega	.026	.002	Omega	.017	.034
Effect size	.162	.045	Effect size	.132	.188
Number of students	36,006	35,981	Number of students	100,992	101,652
Grade 6					
Omega	.066	.024			
Effect size	.265	.156			
Number of students	28,272	28,273			
Grade 8			Grade 8		
Omega	.067	.028	Omega	.003	.013
Effect size	.266	.170	Effect size	.058	.115
Number of students	25,362	25,336	Number of students	106,700	107,016

test scores in determining the ELL/non-ELL dichotomy, we compared the performance of ELL and non-ELL students on reading/language arts, science and mathematics subscale scores of achievement tests using data from sites 1, 2, 5, and 6. Norm-referenced tests (NRTs) in reading, science, and mathematics were used in sites 1 and 2, and state constructed criterion-referenced tests (CRTs) in reading and mathematics were used in Site 5 and 6.

Once again, performance differences between ELL and non-ELL students on the scores of achievement tests were estimated in terms of percent of variance explained and effect sizes. Table 3 presents the effect sizes and the proportion of the variance of the ELL classification code explained based on test scores for students in Grades 3, 6, and 8 in Site 1 and students in Grades 4 and 8 in Site 6. The two sites (Site 1 with the pre-NCLB data and Site 6 with the post-NCLB data) will provide the opportunity to make comparisons over time and to determine the possible impact of NCLB on the classification of ELL students.

As data in Table 3 show, ω^2 and effect sizes were very small for both sites, suggesting that there is not a strong association between standardized achievement test scores and ELL/non-ELL dichotomy. The ω^2 for the NRT test by ELL classification ranged between .002 (mathematics) for Grade 3 to .067 (reading for Grade 8), indicating that based on the average of the three grade levels (3, 6, and 8), only about 3.5% of the variance of ELL classification is explained by the NRT scores. Effect sizes for NRT across all grade levels ranged from .045 (mathematics for Grade 3) to .266 (reading for Grade 8). Once again, these effect sizes are small (Kirk, 1995).

Results from Site 6 are consistent with those reported for data in Site 1. The ω^2 values ranged from .003 (Grade 8 reading) to .034 (Grade 4 mathematics), which on the average explain about 1.7% of the common variance between test scores and ELL classification codes. The effect sizes ranged from .058 (Grade 8 reading) to .188 (Grade 4 mathematics) with an average of .123 which is quite small (Kirk, 1995).

Table 4 presents ω^2 and effect sizes for NRT and CRT tests in explaining the ELL/non-ELL dichotomy for students in Grades 3, 7, and 9 in reading, science, and mathematics in Sites 2 and for Grades 5 and 8 in Site 5. Site 2 used the NRT test scores directly as a criterion for ELL classification. Therefore, one would expect a higher level of association between the achievement test scores and ELL classification in this site. The index of strength of association (ω^2) for the NRT test ranged from .051 (Grade 9 mathematics) to .203 (Grade 7 reading) with an average of .115, indicating that NRT test scores explained about 12% of the variance of ELL classification. The effect sizes for the NRT test scores ranged from .231 for mathematics in Grade 9 to .504 in

Table 4. Site 2 and Site 5, Omega-square, Effect Sizes, and Number of Students in Different Subscales and ELL Classification

	Site 2			Site 5		
	Reading	Science	Math	Reading	Science	Math
Grade 3				Grade 5		
Omega-Square	.172	.089	.076	.108	.091	.068
Effect size	.456	.313	.286	.348	.316	.270
ELL	104,333	23,555	109,327	5,008	5,008	5,008
Non-ELL	272,653	54,300	277,042	28,118	28,118	28,118
Total	376,986	77,855	386,369	33,126	33,126	33,126
Grade 7						
Omega-Square	.203	.132	.101			
Effect size	.504	.390	.335			
ELL	69,074	24,761	71,227			
Non-ELL	267,235	77,834	268,867			
Total	336,309	102,595	340,094			
Grade 9				Grade 8		
Omega-Square	.150	.087	.051	.068	.061	.057
Effect size	.420	.309	.231	.270	.255	.246
ELL	32,515	33,032	33,311	3,870	3,870	3,870
Non-ELL	192,598	192,639	193,906	33,106	33,106	33,106
Total	225,113	225,671	227,217	36,976	36,976	36,976

reading for Grade 7. These effect sizes fall within the mid-category based on Kirk (1995).

The strength of association (ω^2) between the CRT test scores and the ELL classification code was also obtained for data from Site 5. The results of analyses summarized in Table 4 show a slightly lower level of association between the CRT test scores and ELL classification code. The index of strength of association (ω^2) ranged between .057 for Grade 8 mathematics to .108 for reading in Grade 5 with an average of .076, indicating that only about 7.6% of the variance in the ELL classification code is explained by the scores of standardized achievement tests. Similarly, the effect sizes ranged between .246 (Grade 8 mathematics) to .348 (Grade 5 reading) with an average of .284 which is considered a medium effect size.

The data presented above clearly suggest that standardized achievement test scores, including reading and language arts subscales, were not strongly associated with the levels of ELL classification. More importantly, the post-NCLB assessments did not show much improvement over the pre-NCLB assessments in providing more valid criteria for ELL classification. These data suggest that standardized achievement tests may not be a valid criterion for assessing ELL students for classification purposes as a single criterion or even when combined with other criteria.

Discrepancies in Patterns of ELL Classification in Different Districts

Validity issues and inconsistencies in ELL classification criteria have resulted in large discrepancies between states and districts in ELL classification/reclassification practices. These discrepancies will continue to persist until serious consideration is given to the validity of ELL classification. As indicated earlier in this report, standardized achievement test scores are often used as a criterion for classification of ELL students and their reclassification as "Redesignated Fluent English Proficient" (RFEP) or a similar code. To be reclassified as RFEP at Site 2 of this study, ELL students had to score above the 36th percentile on the NRT reading comprehension test, with some discretion allowed. To study the implementation of this "36th percentile policy," we compared agreement between current classification and performance on the standardized reading test.

Table 5 provides data on the agreement between the NRT reading levels and ELL classifications of students in Site 2 for some ELL students with valid scores in the reading content area. Percentages in the table represent agreement between the actual classification and the reading percentiles. As Table 5 shows, 74.5% of students scoring below the 36th percentile were designated as ELL. Of the students scoring above the 36th percentile, 73.8% were reclassified RFEP and 26.2% were still classified ELL (contingency coefficient = .390, $p = .000$). We understand that standardized achievement test scores are not often used as a single criterion but these results question the usefulness of these scores even when they are used in conjunction with other criteria.

Figure 1 shows the variation in percentages of ELL students scoring below the 36th percentile in reading who remain classified as ELL in districts with at least 200 third-grade ELL students. Each "line" in the figure represents one district. Low-scoring ELL students in grades three through five scoring below the 36th percentile tended to remain classified as ELL, while low-scoring students in higher grades were more likely to be reclassified as RFEP. As grade level increased, however, the variation in agreement among districts also increased. Parrish, Perez, Merickel, and Linquanti (2006) found that some districts used alternative reclassification criteria that lowered reclassification standards for ELL students at later grades. They cited concerns among administrators that long-term ELLs could face aggregated track placement and reduced access to courses needed for postsecondary education.

To measure the overall agreement between the current classification and the next NRT reading performance we computed *kappa* coefficient which measures exact agreement beyond chance. Figure 2 presents the *kappa* for the Site 2 districts. As Figure 2 shows, there was a wide variation in overall agreement among these districts at Site 2. Two districts were very close to the kappa = 0 line that indicates no agreement beyond chance between ELL classification and NRT scores. In comparison, three districts had kappa > .50 in middle school.

These discrepancies once again point to the validity concerns in ELL classification/reclassification practices. The districts used in the above analyses were all using the same standardized achievement test. If the 36th percentile score is a good indication of students' relative proficiency in reading comprehension (the policy set by the state), then the irony is that over 25% of the students scoring above the 36th percentile are still classified as ELL. Obviously, one can expect an even larger discrepancy across different states using different standardized achievement tests based on different content standards, with different percentile score cutscore.

Variables/Factors Unrelated to Level of English Proficiency That Might Have Long-Term Effects on Reclassification

Results of the research presented above suggest that students' level of English proficiency is not the only determinant of ELL classification. Other factors may also influence decisions on ELL classification. Grissom (2004) and Abedi (2004) both found that variables such as gender, socioeconomic status (SES, measured by free/reduced price lunch), ethnicity, and parent education are powerful predictors of ELL classification/reclassification. For example, Grissom (2004) indicated that in California multiple criteria are used for reclassifying from ELL to RFEP. These criteria according to Grissom include: (1) assessment of ELP, (2) teacher evaluation, (3) parent opinion and consultation,

Table 5. Site 2, Agreement Between NRT Reading Levels and ELL Classification

	Current Classification	
NRT Reading Level	**ELL**	**RFEP**
Below 36th percentile (*n*)	56,095	6,040
Below 36th percentile (Pct)	74.5%	26.2%
Above 36th percentile (*n*)	19,172	16,983
Above 36th percentile (Pct)	25.5%	73.8%

Note: Contingency coefficient = .390.

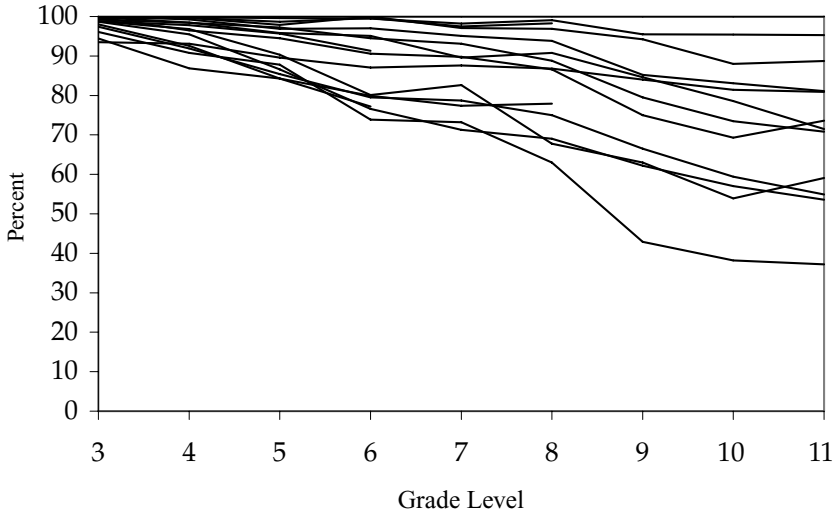


FIGURE 1. Site 2 percent of ELL students scoring below the reading 36th percentile by school district (minimum $N = 200$).

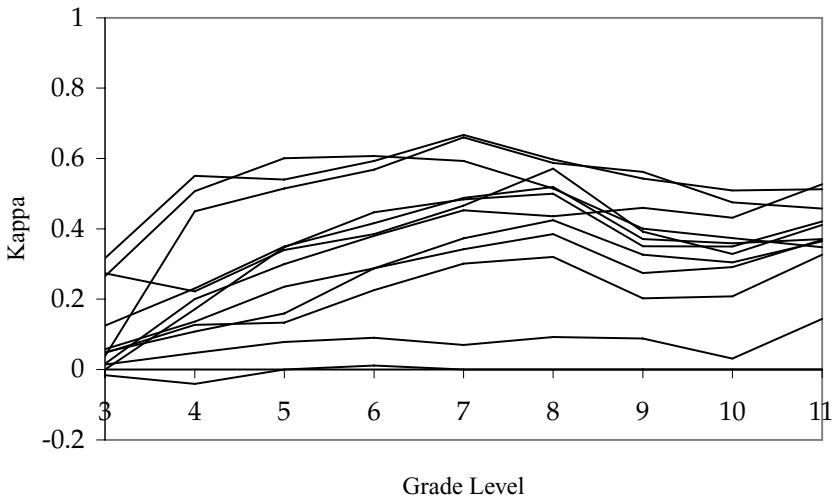


FIGURE 2. Site 2 *kappa* coefficients for agreement between ELL classification and NRT performance (minimum $N = 200$).

and (4) performance in basic skills. Moreover, Parrish et al. (2006), in their examination of nine high- versus low-reclassifying districts in California, noted that variations in the districts' reclassification criteria and cutscore, procedures and systems to monitor students' readiness to reclassify, and the importance of reclassification in local accountability explained much of the observed variation in reclassification rates.

A large district within Site 2 provided a unique opportunity to study the effects of some of these variables on ELL classification/reclassification decisions in a longitudinal setting. We created a cohort of 1993–1994 Grade 7 students ($n = 23{,}856$) and followed them for a period of 6 years (12 semesters, fall 1993 to spring 1999), conducting an event history analysis (also referred to as a *survival analysis* approach, Miller & Zhou, 1997).

Results of the event history analysis indicated that, in addition to the students' level of language proficiency, their background variables (such as ethnicity) appear to correlate with the ELL classification. Table 6 summarizes the results of the event history analyses by students' background variables and test scores. Results are reported by gender, ethnicity, free/reduced lunch participation, Title I status, and reading test scores. Table 6 shows the number of ELL students at the start of the cohort and the proportion of redesignated ELL students for the first period (first six semesters) and the second period (second six semesters). The last column of Table 6 shows the median semesters students remained as ELL.

The data show that the percentages of students that were RFEP from the first to the sixth semester vary considerably across categories of some variables. Between males and females, there was not much variation. In the time period between the first and sixth semesters, 27% of female students were redesignated compared to 23% of male students. In the second period, 62% of female students as compared with 53% of male students were redesignated. This difference was not statistically significant. Consequently, the median time in ELL status (number of semesters) was very similar for males and females (9.08 for females and 9.98 for males). Likewise, percentages of RFEP across the free/reduced lunch program were very similar (9.52 for participants versus 9.55 for nonparticipants).

In contrast to gender and SES, there was a much larger variation in percentages and median time spent in the ELL category across racial/ethnic categories. Percentages of RFEP for the first period across ethnic categories ranged between 21% for Hispanics to 55% for Asians and Caucasians. The percentages of RFEP in the second time period were substantially higher for all ethnic categories as expected. However, once again Hispanics had the smallest percent of RFEP (57% Hispanics versus 77% Asians and 68% Caucasians).

It took almost ten semesters for Hispanic students to be reclassified from ELL to RFEP, while it took half as much time for Asian and Caucasian students to be reclassified. Looking at the levels of reading test scores, it took much less time for students with higher reading scores to be reclassified than students with lower reading test scores as one might expect. However, many of these variables may be confounded with other variables. For example, a majority of students who are classified as ELL are Hispanic.

Initiating Changes

The results of analyses summarized above indicate that the current system of ELL classification produces inconsistent outcomes. This may be due to psychometric characteristics of the assessments used for the classification, the resulting accuracy of ELL classification, the weight of other factors influencing

Table 6. Site 2, 1993–1994 Cohort ELL Event History Analysis by Student Background Variables

Subgroup	Number ELL at Start	Percent RFEP Within 0–6 years	Percent RFEP Within 7–12 years	Median Time in ELL Status
Gender				
Female	11,763	27%	62%	9.08
Male	12,093	23%	53%	9.98
Ethnicity				
Asian	1,246	55%	77%	5.45
Hispanic	21,167	21%	57%	9.91
Caucasian	1,019	55%	68%	5.49
Free Lunch				
Free/reduced	19,099	24%	59%	9.52
Nonparticipant	4,757	29%	50%	9.55
Title I				
Title I	14,166	18%	56%	10.13
Non-Title I	9,690	34%	60%	8.45
93–94 NRT Reading Test				
Not tested	10,465	18%	38%	>12.00
Percentile 1–15	3,835	08%	42%	>12.00
Percentile 16–36	5,704	23%	75%	8.83
Percentile 37+	3,852	59%	94%	5.07
96–97 NRT Reading Test				
Not Tested	10,980	17%	38%	>12.00
Percentile 1–15	6,141	15%	52%	10.80
Percentile 16–36	3,747	32%	84%	7.89
Percentile 37+	2,988	58%	94%	5.13

Note: Students with missing data were deleted; therefore, totals are slightly different across different groups.

ELL classification, or to a combination of the above. More importantly, a valid classification system should be based on the theory of language acquisition and should clearly identify the level of academic language proficiency that students should reach in order to be classified as fluent in English and to be able to fully participate in English-only instruction and assessments.

It must be indicated at this point that the term "ELL classification system" that is used frequently in this paper includes initial identification as language minority (via home language survey), initial identification as English learner (typically via ELP assessments) and reclassification to fluent English proficient (typically via both ELP and achievement test scores). While we distinguish between these different purposes, the ELL classification system as discussed in this paper encompasses all these purposes.

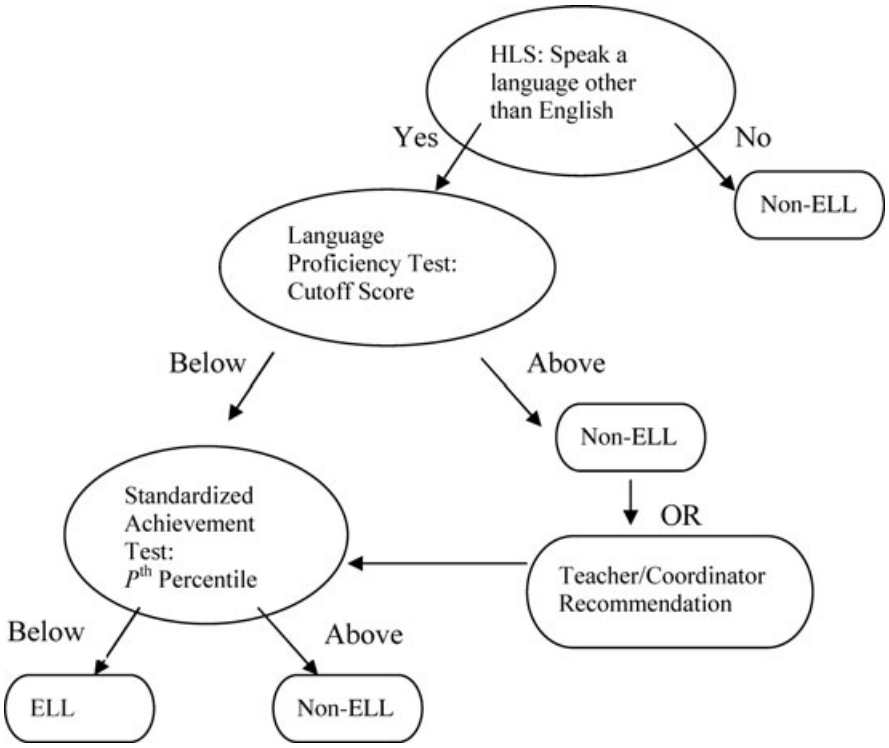
To initiate dialog on creating a model for improving validity of the ELL classification system, we support the use of multiple criteria (Tippeconnic & Faircloth, 2002) with a minimum level of redundancy in a value-added sequence of phases. We build this model on the concept of *academic English* since beyond such a practical concept there is not a commonly acceptable theoretical foundation for ELL classification. We believe the concept of academic English is a sensible base for the model for two reasons: (1) ELL student proficiency in academic English is a prerequisite of their success when both instruction and assessment are offered in English only, and (2) as noted in this paper, measures of students' academic performance are also used as criteria for ELL classification/reclassification.

Figure 3 depicts this stepwise concept. The model uses information currently available and recommends increasing the validity of the classification system by augmenting the knowledge about the student language background with multiple criteria. Obviously, as discussed earlier, the existing data that are used in this model (HLS, ELP, and achievement test scores) may have serious limitations at least in two ways: (1) they do not include the major variables that the literature suggests for increasing the validity of ELL classification (such as number of years in the United States, number of English only classes ELL students take and their proficiency in their native language), and (2) there are serious technical flaws in many of these assessments. However, we want to demonstrate that some improvements to the ELL classification system can be made even under such less than desirable conditions.

The process starts with data from the HLS as one of the most commonly used criteria for identifying linguistic minority status and establishing the need to assess for possible ELL classification. This will establish an initial potential ELL cohort that will then be augmented by additional criteria such as English proficiency and academic achievement test scores. A potential shortcoming at this first level of augmentation, however, is under- or over-identification of potential ELL students based on the HLS data. As an example of over-identification, a native English speaker who does not speak another language may be placed in the potential ELL cohort simply because a family member living with the child sometimes speaks a language other than English. This issue can be avoided by categorizing them as linguistic minority until the ELP assessment identifies them as ELL. On the other hand, under-identification may occur when parents declare the child as English-only because they speak only English at home for practice.

For lack of a better term, we call this process the *Augmented-Classification* (AC) approach. As Figure 3 shows, AC starts with the information from the Home Language Survey. All students who are identified as born outside the United States or who speak a language other than or in addition to English at home will establish the initial potential cohort of language minority students. Information from the next levels of the augmentation (ELP and achievement test scores) will help identify students who are not ELL, despite not being born in the United States and speaking a language other than English at home.

In using test scores (English language proficiency and achievement test scores) as the next levels of augmentation, we need to distinguish



Note: Non-ELL includes FEP and RFEP.

FIGURE 3. Diagram of the Augmented-Classification approach.

between norm-referenced test (NRT) and criterion-referenced test (CRT) data as the two types of data may lead to different classification outcomes. The NRT-based classification system was more common prior to the implementation of NCLB. After NCLB and particularly after development of the new generation of ELP assessments, the CRT-based system was more commonly used. However, since there are still some NRT-based classification practices, we present our proposed AC system based on both alternatives (NRT and CRT).

NRT-Based Augmentation

At this level of augmentation under the NRT-based model, percentile scores are often used. The major flaw with the NRT-based augmentation is the inconsistencies in the cutscore based on which the decision is made to classify/reclassify a student as ELL or non-ELL. Different states adopt different cutscores for identifying their ELL students based on their test scores.

ELP percentile scores are used as the criterion for the second augmentation level. Students who are at or above a given percentile point (to be determined by states based on empirical data) on the ELP tests may exit the ELL cohort. The remaining students in the cohort can then be considered ELL with a higher level of confidence than those identified as ELL based on HLS data alone with teacher's discretion to add an achievement test criterion as another layer of confidence. However, there might be a slight possibility that ELP assessments may have underidentified some students. To control for the possibility of underidentification, achievement test scores can be used as the third level of augmentation. While it may seem unproductive to test ELL students at the lower level of English proficiency in content-based assessment in English, there is evidence that points to such practices. For example, based on the data presented in Table 5, of the total 75,267 ELL students for whom we had score in English language arts, 19,172 or 25.5% of them scored above the 36th percentile on the NRT Reading. Based on these data used as the third level of augmentation, these students can also be reclassified as RFEP.

In the *AC* approach, one can use single or multiple measures at each stage of augmentation. Districts and schools may have multiple measures of English content and multiple measures of English proficiency. To test the improvement level in the validity of ELL classification using multiple criteria at each stage, data from a group of 916 ELL/non-ELL third-grade students were used (Staley, 2005). These students were from a single district within Site 2. For each student the data included four measures of ELP, three standardized English language arts achievement measures, and six standardized mathematics achievement measures. Of the 916 students, 602 had complete data on all the measures used in the analyses.

Composite scores of the four English proficiency scores and three English language scores were created using both latent composite and simple composite approaches. In the latent composite approach, we created a single-factor confirmatory factor model of the individual measures and used the factor score as a latent composite. The simple composite score was computed by converting the measures to standard Normal Curve Equivalent (NCE) scores with a mean of 50 and a standard deviation of 21.06 (Linn & Gronlund, 1995) and then averaging over all the individual measures.

We applied the proposed *AC* approach to the data discussed above by starting with the ELL cohort identified by the district according to their classification policy. Table 7 summarizes the results of these analyses. As the data in Table 7 show, of the 602 students with complete data, 309 or 51.3% were classified as ELL and 293 or 48.7% as non-ELL by their schools.

In the second phase of augmentation, we used the latent composite of ELP

Table 7. Added Classification Power at Different Augmentation Phases

ELL Status Defined	No. of ELL	No. of Non-ELL	% Moved from ELL to Non-ELL	ω^2 English Measures (Effect Size)
By school	309	293	0	.248 (.57)
Based on English proficiency	182	420	41.1	.325 (.70)
Based on English measures	117	485	35.7	.411 (.84)

measures with a cutscore at the median of the English proficiency score distribution. As the data in Table 7 show, of the 309 students who were identified as ELL based on their ELP scores, 127 or 41.1% had ELP test scores above the median of the distribution and thus were reclassified as non-ELL. In this phase, the strength of association (ω^2) was increased from .248 to .325 (an increase of about 8% on the prediction power).

In the next phase of augmentation, we applied the 36th percentile policy on the composite score of NRT English subscales (comprehension, vocabulary, and language). Of the 182 students who were classified as ELL by their school, 65 or 35.7% had scores above the 36th percentile. We reclassified these students as non-ELL. Thus, the AC model improved the strength of association between the ELL classification code and the criteria used for such classification from .248 (24.8% of the variance of ELL classification explained) to .411 (41.1% of the variance) using existing data. We believe more substantial improvements can be observed if the validity of the current criteria for ELL classification is increased. The results of our analyses based on the post-NCLB English language proficiency tests were consistent with the pre-NCLB analyses showing an even higher trend of improvements in the validity of the ELL classification system due to better ELP assessment quality. With the post-NCLB data, we could improve the strength of association between the ELL classification code and the criteria used for classification from 26.4% to 49.6%.

CRT-Based Augmentation

The proposed CRT-Based augmentation model is very similar to the NRT-based model except that in the CRT-based model achievement levels rather than percentile scores are used. Similar to the NRT-based model, this model uses three levels of augmentation. The first level uses data from HLS, the second level is based on ELP test scores (language proficiency levels) and the third level is based on academic achievement test scores (achievement levels). Since the CRT-based classification system has been implemented only recently, there is not enough data to test its effectiveness over the traditional classification system. However, as discussed earlier, the newly developed ELP assessments show higher power in discriminating ELL students at different levels of proficiency in English; therefore, we expect more improvements in the ELL classification system using the new ELP assessments.

There are several potential risks to the validity of the CRT-based classification system. For example, while the collaboration between states in the form of consortium reduced the variation in the ELP assessment outcomes, different states still continue using different ELP assessments. Additionally, differences in the standard setting procedures by states may also create discrepancies in ELL classification system across the nation.

We propose the AC model for two reasons: first, to convey the fact that there are several major concerns with the current ELL classification system that necessitate an urgent remedy and, second, to demonstrate that even the existing data with the limitations and validity issues can be used to provide a more reasonable means of ELL classification. We acknowledge the limitations of the AC model, but we hope that by demonstrating how small steps such as creating a model based on multiple criteria could potentially *improve* the ELL classification system, we can send a message that a more valid ELL classification system is not as far out of reach as many believe. We understand that states may vary in their application of criteria for making decisions about exiting students from ELL status. The AC model could actually help with making such decisions more consistent across and within states.

Note, however, that even the best designed ELL classification system with the most valid criteria may not produce valid outcomes if teachers and school officials (including the bilingual coordinators) are not knowledgeable about assessment and classification systems for these students. To be successful in this area, teachers must know about ELL students, their background characteristics and how their educational needs might be different from the native speakers of English.

Discussion

Because of inherent background differences between ELL and non-ELL students in terms of ELP, the universal use of curricula and assessments designed for non-ELL students may lead to inappropriate instruction and create invalid inferences about ELL academic achievement. The most important prerequisite to providing appropriate instruction and fair and valid assessment for ELL students is to correctly identify them. Inappropriate classification decisions may place students who are at a higher level of English proficiency into remedial or special education programs and may deprive less-proficient students of appropriate curriculum and assessment. Poor placement decisions may affect promotion and graduation, which consequently affects students' academic progress and self-esteem. Misclassification of ELLs may also impact school, district and state accountability systems resulting in negative repercussions. Delay in the reclassification of students who have reached English proficiency may deny them the opportunity to achieve and may reduce access to courses needed for post-secondary education, while premature reclassification may cause ELL students to lose needed specialized academic language instructional services and be placed at greater risk for educational failure (see Linquanti, 2001; Parrish et al., 2006).

Education and assessment communities have raised concerns over the validity of the current ELL classification system. Lack of a strong theoretical foundation and issues regarding the quality of criteria used for such classification are among these concerns. This paper presented empirical evidence substantiating such concerns and initiated discussion on improving the validity of the classification system. Research findings presented in this paper point to the fact that a remedy for this complex problem is urgently needed if ELL students are truly not to be left behind.

As indicated above, the post-NCLB English language proficiency tests improved the validity of the ELL classification system but this trend was not observed in the post-NCLB academic achievement tests. There are at least two possible explanations for this finding. It might be that the performance-gap between ELL and non-ELL students was reduced due to the positive impact of NCLB; therefore, the achievement tests did not show much power in differentiating the two groups in terms of their content-based performance. More likely, this lack of association might be due to ELL classification issues, i.e., large heterogeneity in the ELL population which suppresses the size of performance difference between ELL and

non-ELL students. This finding suggests that states are still facing major challenges in providing more accessible assessments for ELL students.

This paper also cited other contributing factors to inconsistent classification, some of which are unrelated to a student's level of English proficiency, such as ethnicity and schools' Title I status. The impact of such variables on ELL classification decisions may explain large discrepancies in ELL classification within and between states across the nation. While some of these powerful predictors such as ethnicity are not modifiable, knowledge of their impact can inform decisions on a student's ELL status.

The results of studies presented in this paper also raised concerns over the reclassification trends and policies. For example, results indicate that low-scoring students in lower grades (e.g., Grades 3 through 5) tended to remain classified as ELL, while low-scoring students in higher grades were more likely to be reclassified as proficient. Parrish et al. (2006) found that some districts are using alternative, lower reclassification criteria for ELL students in higher grades. If other independent studies confirm this trend, then investigating the potential causes of this trend is important to provide further insight into ELL classification and reclassification practices.

This is a complex situation and a simple solution may not work. Adding more tests to the states already burdened by testing requirements may not be realistic. On the other hand, using existing data as they are used currently for ELL classification may not produce valid and reliable classification outcomes. Thus, the main question is whether the validity of the ELL classification system can be improved using current information with a reasonable level of effort in enhancing the quality of such information.

To initiate a dialog among researchers, educational policymakers, and practitioners, we proposed a model that utilizes the assessment data available from different sources in the state assessment system. The idea is to augment our knowledge of students' English proficiency levels using information from different sources with a minimal level of redundancy.

The model proposed in this paper is tentative and conditional upon improved quality of ELP and standardized achievement tests for ELL students and including other relevant variables such as student's proficiency in L1 and number of years in the United States. For example, academic achievement test scores would help to improve the quality of the ELL classification system when they are more accessible for ELL students linguistically and culturally. Furthermore, the effectiveness of the proposed model is determined by rigorous validation studies that would need to be conducted nationwide. For example the percentile cutscores are set by many states arbitrarily which may not be based on much empirical evidence. Furthermore, some native English speakers score below these cutscores. How should they be treated? More research may be needed to establish the validity of such cutscores nationally.

More importantly, a valid classification system should be based on the theory of second language acquisition and should clearly identify the level of academic language proficiency that is needed for ELL students to function in academic environments where both instruction and assessment are offered only in English. Therefore, we built this model on the concept of *academic English* since beyond such a practical concept there is not a commonly acceptable theoretical foundation for ELL classification.

Improving the validity of the classification system requires both valid criteria and people who are knowledgeable about assessment and classification systems for ELL students to implement the system. The best and the most comprehensive system of ELL classification may not produce desirable outcomes if the implementation phase is not done properly. Therefore, it is imperative for those who are involved in the classification of ELL students to receive proper training and education about these students.

Unfortunately, ELLs are more likely to be taught by teachers without much knowledge on issues concerning classification and assessment of these students and with less classroom experience than teachers of other students (Rumberger & Gándara, 2004; Gándara et al., 2005). Thus, it is clear that not only issues concerning validity of the criteria used in the ELL classification system could create questionable outcomes; problems in the implementation of the classification system also contribute greatly to inconsistencies and problems in the ELL classification system.

We hope this discussion will initiate a national effort in establishing a valid and reliable ELL classification system. Though much legislation mandates equal educational opportunities for every child—including ELL students—very little can be done to improve the academic life of ELL students unless they are validly identified.

Acknowledgments

The work reported here was supported in part under a grant from the U.S. Department of Education, Institute for Education Sciences. The findings and opinions expressed here do not necessarily reflect the position or policies of the Institute for Education Sciences, or the U.S. Department of Education. The author acknowledges the contribution of several people. Robert Linquanti contributed to the quality of this paper by providing excellent comments and suggestions. Rita Pope and Cassandra Hawley assisted in structuring and revising the paper. Seth Leon provided valuable assistance with the data analyses.

Note

[1]The author acknowledges the terms "English language learner (ELL)" or "English learner" (EL) as alternatives to "limited English proficient (LEP)." All refer to students who may be in need of English language instruction, which encompasses a wide range of learners, including students whose first language is not English, students who are just beginning to learn English, and students who are proficient in English but may need additional assistance in social or academic situations (LaCelle-Peterson & Rivera, 1994). "English language learner" has been used as a more positive alternative to the term "LEP," which some regard has having a negative connotation (August & Hakuta, 1998). In this report, we use the term ELL more often since it is more commonly used in research and practice.

References

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

Abedi, J. (2006a). Language issues in item-development. In S. M. Downing & T. M.

Haladyna (Eds.), *Handbook of test development*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Abedi, J. (2006b). Psychometric issues in the ELL assessment and special education eligibility. *Teacher's College Record*, *108*(11), 2282–2303.

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

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *26*(5), 36–46.

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). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

August, D., & Hakuta, K. (Eds.) (1998). *Educating language-minority children*. Washington, DC: National Academy 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* (CSE Tech. Rep. No. 611). Los Angeles: University of California, CRESST.

Bunch, M. B. (2006). *Final Report on ELDA Standard Setting*. Durham: Measurement Incorporated.

Chomsky, N. (1968). Language and the mind. *Psychology Today*, *1*(9), 48–68.

Cisneros, R., & Leone, B. (Eds.) (1995). The ESL component of bilingual education in practice: Critical descriptions of bilingual classrooms and programs. *Bilingual Research Journal*, *19*(3&4), 353–367.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conteh-Morgan (2002). Connecting the dots: Limited English proficiency, second language learning theories, and information literacy instruction. *The Journal of Academic Librarianship*, *28*(4), 191–196.

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

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

Fast, M., Ferrara, S., & Conrad, D. (2004). *Current efforts in developing English language proficiency measures as required by NCLB: Description of an 18-state collaboration*. Washington, DC: American Institutes for Research.

Figueroa, R. A. (1989). Psychological testing of linguistic-minority students: Knowledge gaps and regulations. *Exceptional Children*, *56*(2), 145–153.

Figueroa, R. A. (1990). Best practices in the assessment of bilingual children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (pp. 93–106). Washington, DC: National Association of School Psychologists.

Francis, D., & Rivera, M. (2007). Principles underlying English language proficiency tests and academic accountability for ELLs. In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 13–32). Davis: University of California.

Gándara, P. (2000). In the aftermath of the storm: English learners in the post-227 era. *Bilingual Research Journal*, *24*(1&2), 1–13.

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 (http://www.cftl.org/documents/2005/listeningforweb.pdf) (accessed May 19, 2005).

GAO (2006). *No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency*. Washington, DC: United States Government Accountability Office.

Giraud, G., Impara, J., & Plake, B. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education*, *18*(3), 223–232.

Grissom, J. B. (2004). Reclassification of English language learners. *Education Policy Analysis Archives*, *12*(36) (http://epaa.asu.edu/epaa/v12n36/) (accessed September 3, 2004).

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

Hakuta, K., Butler, Y. G., & 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.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1994).

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485–511). Washington, DC: American Council on Education.

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.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Krashen, S. D. (1988). *Second language acquisition and second language learning*. New York: Prentice-Hall International.

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

Lightbown, P. M., & Spada, N. (2000). *How languages are learned* (revised ed., pp. 38–40). Oxford, U.K.: Oxford University Press.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Linquanti, R. (2001). *The redesignation dilemma: Challenges and choices in fostering meaningful accountability for English learners*. Policy Report 2001-1. Santa Barbara: University of California, Linguistic Minority Research Institute.

Littlejohn, J. (1998). *Federal control out of control: The Office for Civil Rights' hidden policies on bilingual education*. Sterling, VA: Center for Equal Opportunity (http://www.ceousa.org/READ/ocr2.html) (accessed November 23, 2004).

Loop, C. (2002). *Which tests are commonly used to determine English and/or Spanish language proficiency?* (AskNCELA No. 25). Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Mahoney, K. S., & MacSwan, J. (2005). Reexamining identification and reclassification of English language learners: A critical discussion of select state practices. *Bilingual Research Journal*, *29*(1), 31–42.

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

Miller, R. A., & Zhou, X. (1997). *Survival analysis with long-term survivors*. New York: John Wiley & Sons.

NAEP (2007, September 18), NAEP Inclusion Policy (http://nces.ed.gov/nationsreportcard/about/inclusion.asp) (accessed November 12, 2007).

National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs (2002, December). *Glossary of terms related to the education of linguistically and culturally diverse students* (In Ask NCELA No. 10). Washington, DC: Author (http://www.ncela.gwu.edu/expert/glossary.htm) (accessed May 21, 2004).

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

Parrish, T., Perez, M., Merickel, A., & Linquanti, R. (2006). *Effects of the implementation of Proposition 227 on the education of English learners, K-12: Findings from a five-year evaluation (final report)*. Palo

Alto and San Francisco: American Institutes for Research and WestEd.

Reutzel, D. R., & Cooter, R. B. (2007). *Strategies for reading assessment and instruction helping every child succeed*. Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

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.

Rossell, C. H. (2000). *Different questions, different answers: A critique of the Hakuta, Butler, and Witt report, 'How long does it take English learners to attain proficiency?'* READ Perspectives, 7, The READ Institute (http://www.ceousa.org/READ/hakuta.html) (accessed November 23, 2004).

Rumberger, R., & Gándara, P. (2004). Seeking equity in the education of California's English learners. *Teachers College Record*, *106*, 2031–2055.

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.

Staley, L. E. (2005). *The effects of English language proficiency on students' performance on standardized tests of mathematics achievement*. Unpublished doctoral dissertation, University of California, Los Angeles.

Stefanakis, E. H. (1998). *Whose judgment counts? Assessing bilingual children, K-3*. Portsmouth, NH: Heinemann.

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.

United States General Accounting Office (2001). *Meeting the needs of students with limited English proficiency* (GAO-01-226). Washington, DC: Author.

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

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