Inclusion of Students with Limited English Proficiency in NAEP: Classification and Measurement Issues

CSE Report 629

Jamal Abedi
CRESST/ University of California, Los Angeles

May, 2004

Center for the Study of Evaluation (CSE)
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education and Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 4.2: Validity of Assessment and Accommodations for English Language Learners
Jamal Abedi, Project Director, UCLA.

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, the U.S. Department of Education, or the McDonnell Foundation.

INCLUSION OF STUDENTS WITH LIMITED ENGLISH PROFICIENCY IN NAEP: CLASSIFICATION AND MEASUREMENT ISSUES

Jamal Abedi

CRESST/University of California, Los Angeles

Abstract

Research reports major concerns over classification and measurement for students with limited English proficiency (LEP). A poor operational definition of the English language proficiency construct and validity concerns about existing language proficiency tests are among these issues. Decisions on including LEP students in large-scale assessments such as the National Assessment of Educational Progress (NAEP) may be directly influenced by some of these factors. Poor relationships between the existing LEP classification codes with English proficiency and achievement test scores raise concern over the validity of the LEP classification system. These factors have contributed to inconsistencies in LEP classification across districts and states. Criteria used for the inclusion of LEP students in NAEP need to be more objectively defined. Based on the recommendations of existing research, the appropriate levels of English language proficiency for participation in NAEP should be determined by reliable and valid English language proficiency measures. With funding through a competitive bidding process authorized under the No Child Left Behind section on *Enhanced Assessment Instruments*, there are national efforts currently underway in developing English proficiency tests that can be used to provide valid measures of students' level of English proficiency. These efforts should be guided by the relevant theory and research findings, otherwise past problems relating to the validity of English proficiency tests may recur. Multiple criteria including valid and reliable measures of students' level of English proficiency could help with a more consistent decision-making process for the inclusion of LEP students.

Perspectives

There is growing concern over the validity of assessment for English language learners (ELLs). Usually referred to as students with limited English proficiency (LEP), they are the fastest growing school-age population in the United States. Between 1990 and 1997, the number of U.S. residents not born in the United States increased by 30 percent, from 19.8 million to 25.8 million (Hakuta & Beatty, 2000). According to the survey of the states' LEP students,

over 4.5 million LEP students were enrolled in public schools during the 2000-2001 school year (Kindler, 2002). The LEP population has grown by over 100% since the 1990-91 school year, during which the general school population has grown by only 12%.

The rapid growth of LEP students demands consistent and accurate measurement of their academic progress and the determination of areas in which they need most assistance. Accordingly, legislation in the last decade, such as the Improving America's School Act of 1994 and the No Child Left Behind (NCLB) Act of 2001, have mandated inclusion of these students in national and state assessments using reliable and valid measures (Abedi, 2004, Erpenbach, Forte-Fast, & Potts, 2003; Mazzeo, Carlson, Voelkl, & Lutkus, 2000; NCLB, 2002).

The goal of NAEP as the *Nation's Report Card* is to provide accurate and fair assessment to *all* students, including LEP students. Thus, NAEP's policy, particularly in recent years, is to include LEP students to the extent possible. However, before developing an assessment system that provides valid and reliable inferences of what LEP students know and can do, a well-defined, objective definition of the term "LEP" is needed. Who are these students? What are their academic and background characteristics? How consistent are the NAEP and states in their LEP definitions and inclusion decisions?

This paper will discuss issues concerning the classification of ELL students and will elaborate on factors that impact decisions to include ELL students in NAEP assessments. The paper will:

- examine the validity of LEP classification codes assigned to these students;
- describe current and proposed assessments used to measure English language proficiency;
- discuss the relationship between language proficiency and achievement test scores with LEP classification codes;
- address issues related to the inconsistency of LEP classification across districts and states;
- speculate on appropriate levels of English language proficiency for participation in NAEP given the current status of testing and classifying ELLs; and
- explore ways to increase consistency in inclusion decisions made for NAEP.

Validity of LEP Classification

For LEP students, assessment is especially complex and important because it is not only used for accountability purposes, student learning, and growth, but also identifies which students qualify for special LEP-related services, such as bilingual education. According to Zehler, Hopstock, Fleischman, and Greniuk (1994) for LEP students, "Assessment also is used to determine student placement, to determine eligibility for other special programs, to reclassify students from LEP status, and to make decisions regarding promotion or graduation" (p. 4). Unfortunately, criteria for identifying LEP students are not used uniformly across the nation. This could be problematic, since as August (1994) indicated, the lack of a national or even a state definition of limited English proficiency can "result in differential exclusion rates across states making state by state comparison problematic" (p. 2).

The existing LEP definition is based on many different criteria at different locations in the nation. Kindler (2002, Table 2) identifies seven methods and four categories of tests (1. language proficiency tests, 2. achievement tests, 3. criterion reference tests, and 4. other tests) used for identifying LEP students. Among the most important criteria for identifying LEP students are being a speaker of a language other than English and scoring low on the English proficiency tests. The first criterion, i.e., being a non-native English speaker, is defined in many areas nationwide based on the information from the Home Language Survey (HLS). The second criterion, student's proficiency in English, is obtained based on scores on English proficiency tests and achievement tests.

HLS is used in many locations nationwide to determine which students should undergo English Language Assessment. This survey is administered to the families of students before matriculation and the information provided by the survey is included in each student's permanent file. The main purpose of the survey is to determine if a language other than English is being spoken in the home and to identify that language. For many schools, the HLS is the only source of information used to determine the need for a student to be tested for English proficiency. Recent dialogue over the effectiveness of bilingual instruction has led to an increase in the number of parents filling out the HLS inaccurately for the purpose of assuring that their children be treated no differently than other students. Other concerns for the student whose parents may have citizenship

issues have led to a more relaxed treatment of the home surveys than the district would prefer. Questions have been raised about the accuracy of surveys completed by parents who are illiterate or who have no familiarity with written English.

In a study by Abedi, Lord, and Plummer (1995), the accuracy of information provided by the HLS was examined by developing a Language Background Questionnaire which was administered to a group of about 1,500 Grade 8 students in math. Data culled from the students' responses to this Language Background Questionnaire were compared with school rosters reporting the students' official primary languages, as reported by the parents on the district's HLS, and their ESL status where appropriate. Significant discrepancies were revealed which led researchers to question whether or not the schools were always cognizant of the language backgrounds of their students. In the case of most schools, the school's record of the number of students who spoke a language other than English at home, regardless of ESL classification, was significantly lower than what the students themselves reported. Similar discrepancies were found in other language background studies conducted at CRESST (Abedi, Lord, & Hofstetter, 1997a; Abedi, Lord, & Hofstetter, 1997b).

Educational policy makers believe that there should be only one scale upon which the many different measures of the English proficiency of students with different language backgrounds are weighed. Every school surveyed in the UCLA/CRESST studies used some form of the LEP classification codes. The tacit assumption, however, is that different schools apply the LEP classifications uniformly, meaning that a student who is classified with a specific level at one school would be similarly classified if that student attended a different school. Unfortunately, a close look at the samples in the CRESST studies reveals that this may not be the case.

The problem of a lacking operational definition of LEP discussed by CRESST researchers exists elsewhere in the nation. NAEP does not provide a definition of the LEP population—instead, it presents criteria for the inclusion of LEP students. For LEP students, NAEP inclusion criteria indicate that:

> A student who is identified on the Administration Schedule as limited English proficient (LEP) and who is a native speaker of a language other than English should be included in the NAEP assessment unless:
>
> The student has received reading or mathematics instruction primarily in English for less than 3 school years including the current year, and
>
> The student cannot demonstrate his or her knowledge of reading or mathematics in English even with an accommodation permitted by NAEP (NCES, 2001).

While this definition is not based on students' level of English language proficiency, it relies on some other information such as the number of years in English-only classes and judgment on LEP students' ability to demonstrate their knowledge in reading and math. Unfortunately due to a lack of reliable data and because of the subjectivity of some of these criteria, the validity of such information is also questionable.

Based on the inclusion instructions described above, NAEP excludes students who have received reading or mathematics instruction primarily in English for less than 3 school years and cannot demonstrate their knowledge of reading or mathematics in English even with accommodations permitted by NAEP. This type of information can only be obtained from individual schools, which means that NAEP must rely on school records. However, in many different parts of the country, schools are unable to provide accurate information on these areas that NAEP uses as criteria for including ELL students. A high rate of transience in schools with large numbers of LEP students may cause inaccuracy in reporting the number of years in English only classes. Similarly, school's judgment on students' ability to demonstrate their knowledge may not be accurate since it may be subjective.

The No Child Left Behind Act and ELL Classification.

NCLB, the most recent reauthorization of the Elementary and Secondary Act (ESEA) of 1965, requires states to report Adequate Yearly Progress (AYP) for all students and for subgroups including LEP students (Abedi, 2004). Due to the importance of LEP subgroups in NCLB accountability and reporting, NCLB provides an operational definition of LEP (NCLB, 2002). According to this definition:

The term 'limited English proficient', when used with respect to an individual, means an individual

(A) who is aged 3 through 21;

(B) who is enrolled or preparing to enroll in an elementary school or secondary school;

(C)(i) who was not born in the United States or whose native language is a language other than English;

(ii) (I) who is a Native American or Alaska Native, or native resident of the outlying areas; and

(II) who comes from an environment where a language other than English has had a significant impact on the individual's level of English language proficiency; or

(iii) who is migratory, whose native language is a language other than English, and who comes from an environment were a language other than English is dominant; and

(D) whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual

(i) the ability to meet the State's proficient level of achievement on State assessments described in section 111(b) (3);

(ii) the ability to successfully achieve in classrooms where the language on instruction in English; or

(iii) the opportunity to participate fully in society.

While the NCLB's definition of LEP seems to be operationally defined, different states, districts, and schools may interpret these criteria quite differently (Abedi, 2004). For example, as with the other definition of LEP presented earlier in this paper, in the NCLB definition, English language proficiency test scores are among the most important criterion for LEP classification. We will discuss the major issues concerning the validity of existing English language proficiency tests in the next section. Because of such concerns with the existing English proficiency tests, new instruments for measuring English proficiency are under development. However, no information is available yet to evaluate the quality of the new tests. Assuming the new tests are of a high psychometric and content

quality, states must be given enough time to collect the necessary data for establishing new classification standards.

Current and Proposed Assessments of English Language Proficiency

Language proficiency and achievement tests in English are commonly used for the identification and assessment of LEP students. In a report of survey results of the states' LEP students, Kindler (2002) indicated while a very high proportion of State Educational Agencies (SEAs) provided information on LEP reading assessment in English (54 out of 58), the data do not present a clear picture of LEP reading level. Of the 41 SEAs who reported on both the participation and performance of LEP, 25 (61%) indicated that state-designed achievement tests were used to measure LEP students' level of English reading comprehension. Language proficiency tests such as the Language Assessment Scales (LAS) were used in 15 states, and Terra Nova in 11 states (Kindler, 2002). By reviewing the data on LEP English assessments, Kindler indicated that:

> Meaningful interpretation of the available data is challenging for several reasons. The assessment instruments used – as well as testing policies and cut-off scores – vary from state to state and even among districts within a state; therefore, results across jurisdictions are not strictly comparable (pp. 12-13).

In an earlier study, Hopstock, Bucaro, Fleischman, Zehler, and Eu (1993) indicated that 83 percent of school districts use English language proficiency test scores alone or with other measures to decide if a student is LEP. These tests are also used for assigning LEP students to specific instructional programs by 64 percent of school districts and for reclassifying students from LEP status by 74 percent of schools. According to Hopstock et al., the English proficiency tests used frequently for such purposes are the Bilingual Syntax Measure (BSM), the Idea Proficiency Test (IPT), the Language Assessment Battery (LAB), the Language Assessment Scales (LAS), the Maculaitis Assessment Program (MAC), and the Peabody Picture Vocabulary Test (PPVT).

Zehler et al. (1994) did a comprehensive and thorough review of these tests. They compared these tests by their content and structure (productive skills, receptive skills, and reading skills); test administration procedures; theoretical basis of the tests; and issues related to the validity and reliability of the tests. They found major differences in all the areas in which the tests were compared. For example, by comparing the content of the tests, they indicated that:

> The content comparison of the six language proficiency tests showed that the tests differ considerably in types of tasks and specific item content. Even where two tests appear to require the same type of response and similar item content, the scoring criteria may focus on totally separate aspects of the response. As a result, the items are actually assessing totally different skills (Zehler et al., 1994, p. 13, see also Table 2 and Table 3 in this report, pp. 14-20).

By comparing the theoretical basis of the tests, they reported that:

> The MAC, LAS, LAB, and IPT do utilize broader ranges of language skills to assess language proficiency. The LAB and MAC include items that tap literacy skills as well as oral proficiency skills" (Zehler et al., 1994, p. 22).

In comparing the six tests on issues related to content validity and reliability, Zehler et al. (1994) also found major differences between the tests.

> ...past reviewers of the language proficiency tests have noted concerns with reliability and/or validity of the tests, the adequacy of the scoring directions, the limited populations on which test norms are based, and the availability of the conditions needed for administration of the measure (Zehler et al., 1994, p. 23).

Based on Hopstock et al., achievement tests in English are also used by approximately 52 percent of school districts and schools in the nation to help identify LEP students, assign them to school services, and reclassify them from LEP status. About 40 percent of districts and schools use achievement tests for assigning LEP students to specific instructional services within a school, and over 70 percent of districts and schools use achievement tests to reclassify students from LEP status (Zehler et al., 1994). The achievement test batteries that are most frequently used by school districts for identifying and reclassifying LEP students are the California Achievement Test (CAT), Iowa Test of Basic Skills (ITBS), Metropolitan Achievement Test (MAT), Stanford Achievement Test (SAT), and the Comprehensive Test of Basic Skills (CTBS) (Hopstock et al., 1993).

Zehler et al. (1994) compared the reading tasks and skills of the most frequently used achievement tests in the area of reading, math, language, study skills, listening, and science/social science. On these tests, they also found major differences in those areas (see Zehler et al., 1994, Table 4, pp. 26-31).

Current Efforts in the Development of English Language Proficiency Tests

With funding through a competitive bidding process authorized under NCLB, Title VI, Subpart 1, Section 6112: *Enhanced Assessment Instruments*,

currently there are six efforts, some in the form of consortia of states, concerning the development of new English language proficiency tests. The new tests aim to measure four English proficiency domains: speaking, listening, reading, and writing as required under the No Child Left Behind Act of 2001. The main reason for the development of new instruments for assessing students' level of English proficiency was because of the possible shortcomings of existing tests. Earlier, we discussed some of the conceptual and psychometric problems of existing tests. Among the important issues was the lack of evidence for the alignment of content in existing tests with English language proficiency content standards. Bailey and Butler (2003) indicate that "the content of currently available commercial language proficiency tests is not adequate to measure the level of language proficiency necessary for taking standardized achievement tests and for full participation in the mainstream classroom" (p. 14).

Bailey and Butler (2003) discussed several sources of content standards: (a) the national content standards that are defined by national organizations such as the National Science Education Standards published by the National Research Council (1996); (b) State standards by some states with a large number of LEP students such as California, Florida, New York and Texas. Based on these state content standards students should be able to analyze, compare, describe, observe, and record academic information; (c) ESL standards, such as standards by Teachers of English to Speakers of Other Languages (TESOL, 1997). The TESOL standards differentiate between language for social and personal interactions and standards for academic use.

To provide valid and consistent test scores that could inform classification and inclusion decisions, the new tests must adhere closely to these standards. The consortia assigned to this important national task should carefully review these standards and apply them to their English proficiency test development process. They should also communicate with each other so that they all use a common set of English proficiency standards to the extent possible. Using a common set of standards makes LEP classification and NAEP inclusion decisions more consistent across the nation.

Unfortunately, there is not much information from these consortia on the process they use in developing English language proficiency tests. There is also no evidence of communication between these consortia. In fact, based on the

information the author obtained from the consortia's websites for this paper, some of the consortia have not even begun writing the test items.

The main concern is that the information regarding the consortia's process of new English language proficiency test development could be released at a time when it may be too late to make any adjustments. We believe this is a great opportunity for the nation to develop tests that provide more reliable and valid measures of English proficiency without having the limitations and problems discovered in some of the existing tests. However, to reach the point of having more reliable and valid tests, we must do whatever we can and learn from our past experiences to avoid similar problems in some of the existing English proficiency tests.

Relationship Between Language Proficiency and Achievement Test Scores with LEP Classification Codes

LEP classification codes (i.e., LEP, not LEP, etc.) reflect the level of proficiency in English. Consequently, scores of English language proficiency tests and the English language arts section of academic achievement tests should serve as valid and relevant criteria for LEP classification, and therefore be reflected in the code assigned to students. To examine the relationships between the language proficiency and standardized achievement test scores and the LEP classification codes, we computed correlation coefficients between test scores and LEP codes using data from four locations nationwide (Abedi, 2003). These locations, referred to as Sites 1 thru 4 for anonymity, provided comprehensive data on language proficiency tests such as the Language Assessment Scales (LAS), standardized achievement tests such as Stanford 9 and ITBS, and LEP classification codes. Results of the correlation between LAS levels (and LAS scores when available) and LEP classification codes indicated a weak relationship. For example, the correlation coefficients between the LAS and LEP codes computed for Grades 2 through 12 in Site 4 ranged between .176 (n=836) for Grade 12 to .304 (n=945) for Grade 10. The average correlation between the LAS level and LEP classification code across the 11 grade levels was .223, which explains less than 5% of the common variance (for a detailed description of the results, see Abedi, 2003).

Results of the analyses on data from Site 2 showed that LAS has a negatively skewed distribution, which suggests that the test did not have enough discrimination power on the higher end of the distribution. For example, of the 410 LEP students in one district, 139 or 34% were in the highest score level, obtaining reading scores between 90-100. This may be consistent with the aim of the test as a criterion-referenced test, which is to establish a mastery point beyond which students can be classified as English proficient. If this were the case, then, for example, the large group of students who had scores equal or over 80 (242 or 59%) and were classified as LEP should have been re-designated as English fluent. The markedly skewed distribution is indicative of the restriction of range problem that seriously affects the correlation size between LAS and any other variable (Allen & Yen, 1979, pp. 34, 39). LAS writing scores for Grade 3 LEP students within a district in Site 2, like the LAS reading scores presented above, also had a negatively skewed distribution. That is, the restriction of range is still strong in this case and impacts the correlation coefficients (see Butler & Castellon-Wellington, 2000).

In addition to using English language proficiency test scores, many schools nationwide use scores from standardized achievement tests for the classification and reclassification of LEP students. Though reading/language arts subscale scores are frequently used as criterion for LEP classification, some content-based subscale scores, such as math, are also used. Since these subscales measure language proficiency to a certain extent, one would also expect a relatively high correlation between the subscale scores and LEP classification codes.

However, results of our analyses suggest that the relationship of the reading/language arts and math subscale scores of standardized achievement tests with the LEP classification codes is not strong. Two different standardized achievement tests were used among the four data sites: ITBS by one site and SAT-9 by the other three. Correlations between ITBS test scores and LEP classification codes were all statistically significant, but were very low. For Site 1, correlation coefficients between ITBS reading and students' bilingual status ranged from .160 (n=36,006) in Grade 3 to .257 (n=25,362) in Grade 8 with an average correlation of .224, suggesting the two variables share less than 5% of the variance. For this site, correlation coefficients between math concepts and bilingual status ranged from .045 (n=35,981) to .168 (n=25,336) with an average correlation of .122 (1.5% of the variance), and correlation coefficients between

math computation and bilingual status ranged from .028 (n=36,000) to .099 (n=25,342) with an average correlation of .069 (less than half a percent of the variance).

In Site 2, correlation between the reading section of the SAT-9 and LEP classification code ranged from .387 (n=225,113) in Grade 11 to .450 (n=336,309) in Grade 7 with an average correlation of .422 (18% of the variance of the two distributions). The correlation coefficients between the science and LEP codes ranged from .295 (n=225,671) in Grade 11 to .363 (n=102,595) for Grade 7 with an average correlation of .323 (10% of the variance). For math, the correlations ranged from .225 (n=227,217) in Grade 11 to .329 (n=370,435) in Grade 5 with an average correlation of .287 (8% of the variance).

Correlations between SAT-9 reading test scores and the LEP classification code for Site 3 ranged from .131 (n=11,158) to .140 (n=8,740) with an average correlation of .136 (1.8% of the variance). The correlation between SAT-9 science scores and the LEP classification codes ranged from .088 (n=10,231) to .095 (n=7,900) with an average correlation of .092 (less than 1 percent of the variance), and correlation between SAT-9 math scores and the LEP classification codes ranged from .005 (n=8,040) to .029 (n=10,301) with an average correlation of .017 (almost 0% of the variance).

Similar trends were found with correlation coefficients between SAT-9 reading scores and the LEP classification code for Site 4. These correlations ranged from .178 (n=14,050) to .252 (n=9,499) with an average correlation of .223 (about 5% of the variance). For this site, the correlation coefficients between math computations and the LEP classification code ranged from .067 (n=14,282) to .088 (n=12,579) with an average correlation of .081 (about half a percent of the variance).

As a caveat in our discussion, we must indicate that the size of point-biserial correlation which was used to compute correlations between LEP status (0,1) as a dichotomous variable and test scores (English proficiency and achievement tests) as a continuous variable can be severely restricted by the unequal proportions in the LEP categories (LEP versus non-LEP). For details of this restriction of the correlation size, see Allen and Yen (1979).

The data presented above clearly suggest that English language proficiency and achievement test scores, including reading and language arts subsections of

these tests, did not show enough power in predicting levels of the LEP classification codes. This may be a clear indication that other factors/variables influence LEP classification.

Consistency of LEP Classification Across Districts and States

Many school districts around the nation use standardized achievement tests for the reclassification of LEP students as Reclassified Fluent English Proficient (RFEP). In Site 2, to be reclassified as RFEP in 1997-1998, LEP students had to score above the 36th percentile on the SAT-9 reading comprehension test with some discretion allowed. The results of analyses on the agreement between the SAT-9 reading levels and LEP classification codes for Grade 6 students in Site 4 indicated that 90.3% of students scoring below the 36th percentile were designated as LEP. However, there is less agreement for the students scoring above the 36th percentile, as only 47% of these students had been reclassified as RFEP and 53% were still classified as LEP. A *kappa* coefficient of .403 shows only a moderate level of agreement between the SAT-9 reading scores and LEP classification code.

Variation among classification patterns was examined for those students scoring below the 36th percentile in reading, in those districts with at least 200 Grade 3 students. The results of these analyses suggested that low-scoring students in Grades 3 thru 5 tended to remain classified as LEP, while low-scoring students in later grades were more likely to be reclassified as RFEP. As grade level increased, however, the variation in agreement among districts also increased. This district-level variation is especially pronounced in Grades 9 through 11 as some districts reclassify a large portion of these low-performing students, while other districts keep the majority of these students classified as LEP (for details of these analyses, see Abedi, 2003).

The results of analyses also showed variation among district classification patterns for students scoring above the 36th percentile in reading and reclassified as RFEP. In the early grades, the majority of students who score above the 36th percentile remained classified as LEP. As grade level increased, there was more agreement between performance and classification. Students who scored above the 36th percentile in later grades were more likely than those who did so in earlier grades to be reclassified as RFEP. There was a great deal of variation between districts' classification patterns for students scoring above the 36th

percentile. In other words, some districts were consistently more likely to reclassify students scoring above the 36th percentile as compared to other districts.

The results of analyses also indicated that with increasing grade level, agreement between classification and performance decreased for low-performing students but increased for students who scored above the 36th percentile. In order to better understand the overall agreement between classification and SAT-9 performance, we examined the *kappa* coefficient. There was a wide variation in overall agreement among these districts in Site 2. Two districts were very close to the *kappa* = 0 line which indicates little or no agreement beyond chance. In comparison, 3 districts had *kappa* > 0.50 in middle school. As grade level increased from early elementary to middle school, *kappa* tended to increase. However, this trend reversed as students enter high school.

Appropriate Levels of English Language Proficiency for Participation in NAEP

To suggest an appropriate level of English language proficiency for participation in NAEP, one must have access to a reliable and valid measure(s) of English language proficiency. Research results (as partly discussed above) do not suggest any single measure that is adequately reliable and valid to be used for this purpose. That is, there may not be any single criterion that highly correlates with the LEP classification code. This may be due to psychometric characteristics of the measures (criteria), or to issues regarding the validity of LEP classifications or, a combination of both, or to the weight of other (non-testing) reclassification criteria. Using multiple criteria may help increase the validity and reliability of measures in assessment and classification of LEP students.

First, we present research outcome that supports the use of multiple criteria. We will then provide suggestions on how to use multiple measures to create a more reliable and valid LEP classification code.

The validity of multiple criteria versus single criteria in assessing students' level of English proficiency was investigated using scores from a group of 391 LEP students (Abedi et al., 2003). To test the level of improvement in the validity of LEP classification using multiple criteria, we created a measured composite score of three different language measures: (a) a 10-item subsection of the LAS, (b) a 10-item subsection of NAEP and (c) a 60-word word recognition test.

The measured composite score was created by adding up the scores of the three instruments. To adjust for scale differences, standardized scores (z scores) were computed based on the mean and standard deviation from previous data (Abedi et al., 2003). We first correlated the scores of each of the three instruments with the LEP classification code. We then correlated the measured composite with the LEP classification code. As a single criterion, the 10-item LAS section had a correlation of .50 with LEP classification code. Similarly, the correlations of the NAEP subsection and the word recognition test were .36 and .41, respectively, with the LEP classification codes. The measured composite had a correlation of .46 with the LEP classification codes.

We also created a latent composite to correlate with the LEP classification codes. A latent composite score was created through a confirmatory factor analytic approach using the three English language measures as measured variables. We found a correlation of .59 between this latent variable (language) and the LEP classification code. Using multiple criteria in classifying students would help to further strengthen the relationship between English proficiency and LEP classification.

We highly recommend using multiple criteria for the identification of LEP students and for decisions regarding the inclusion of these students in NAEP. Multiple criteria can be used in different ways. They can either be combined using a single cutoff point or used separately with multiple cutoff points. Multiple criteria should include multiple measures of students' level of English proficiency, the teacher's rating of a student's level of English proficiency, and scores of the reading/language arts portions of standardized achievement tests. In addition to test scores and teacher's ratings, some of students' background variables may help to increase validity of criterion for LEP classification. Among these variables, the number of years in the U.S. and the number of English-only classes a student has been attending can be mentioned.

Ways to Increase Consistency in Inclusion Decisions for NAEP

Inconsistencies in inclusion decisions for LEP may occur within NAEP across assessment years and/or in a given assessment across states. These inconsistencies, whether within NAEP assessment years or between states on a given year, may be caused by the lack of an operational definition and of objective criteria used for inclusion decisions. As indicated earlier, in its inclusion

decision, NAEP relies on schools to determine how many years a student has in English-only classes, and whether or not the student is able to demonstrate his/her English language knowledge in reading and math. We discussed schools' limitations in providing accurate information on the number of years students study in English-only classes. We also indicated that school judgment on students' ability to participate in assessment is not usually based on objective criteria; therefore, the validity of these criteria may be questionable.

Since the main concern for including LEP students is their possible English language limitations, a set of reliable and valid measures of English proficiency must be devised and decisions about LEP students should be based on such objective criteria. At this juncture, when NCLB is supporting the development of English proficiency tests, it is imperative that prior to any effort in developing any English language proficiency test, this domain be operationally defined. The definition should be based on current developments in the areas of psycholinguistics, developmental psychology, education, linguistics, and psychometrics. Content standards for English for speakers of other languages (ESOL) should also be considered (see Bailey & Butler, 2003). The new test development should also be informed by the wealth of experience from the administration of current or old language proficiency tests.

It is also essential that the different consortia developing the new English proficiency tests communicate with each other so that the test contents are common across the different consortia. The new English proficiency tests that are developed based on solid theory and experience can then be used as the main criteria for inclusion decisions.

Discussion

Non-native English-speaking students who have difficulty in reading, writing, speaking, and understanding the English language—usually referred to as LEP students—have been traditionally excluded from large-scale state and national assessments. However, in response to recent legislation mandating inclusion of all students in assessments, NAEP has modified its policy to incorporate all students including English language learners. The policy of inclusion of all students should be implemented by providing reliable, valid, and fair assessments for all. However, new studies on the assessment, classification, and accommodation of ELL students cast doubt on the assessment quality and

policy for these students. Among the major concerns related to the inclusion and assessment of LEP students are content coverage and psychometric characteristics of achievement and English language proficiency tests. The validity of measures has directly impacted the identification and classification of these students. Thus, due to the lack of a commonly accepted operational definition of LEP and problems with the existing achievement and English language measures for LEP students, there are major concerns about the validity of classification of LEP students. If the population of LEP students is not well defined due to these measurement problems, then decisions on the inclusion of these students will be inconsistent.

The principal theme of this paper focuses on the validity of the classification of LEP students and how problems in classifications could jeopardize sound decisions for the inclusion of LEP students in NAEP. We have presented results of studies on six different areas that are all directly, or indirectly, related to sound decisions for the inclusion of LEP students in NAEP.

Results of the studies presented earlier in this paper suggest major flaws in the criteria used for classifying students as either LEP or non-LEP. Among the most commonly used criteria for LEP classification are Home Language Survey (HLS) results, achievement, and English proficiency test scores. The research results presented above cast doubt on the validity of these criteria. There is reason to believe that HLS results may not be valid due to parents' concern over equity in education for their children, parents' citizenship issues, and communication problems.

Current English proficiency tests have major shortcomings. There is little evidence on the alignment of existing English proficiency test contents with English language proficiency standards (national, state, or ESOL standards). Studies comparing existing English proficiency tests found substantial differences in the content of these tests. Therefore, decisions based on the results of these tests could be very different depending on which test is used.

Results of the recent studies did not show a strong relationship between the LEP classification categories and the English proficiency and achievement test scores. This is a great concern since LEP classification should be defined based on students' level of language proficiency and their test scores in reading/language arts and mathematics. The lack of a strong relationship between these variables

suggests that variables other than students' level of English proficiency and achievement may determine their LEP status.

The influence of factors other than test scores on LEP classification causes inconsistency in the classification of LEP students. Results of the studies presented above clearly show these inconsistencies. Kappa coefficient, an index of the level of consistency, ranged from zero (no consistency between schools in the districts) for some districts to kappa greater than .50 (suggesting a relatively high level of consistency between schools) within some other districts. We believe districts with a higher level of consistency in their LEP classification policies have more operationally defined criteria for LEP classification.

NAEP's current policy of the inclusion of LEP students relies more on information from schools rather than test scores and other objective criteria. Information such as the number of years a student has attended English-only classes and the school's judgment on whether a student can meaningfully participate in the assessment are used by NAEP to decide whether to include LEP students with lower levels of English proficiency. As we discussed earlier, the validity of these data may be questionable due to threats by factors outside of school controls. For example, most of the schools with large numbers of LEP students in the nation may not have a good tracking system of the number of years a student participates in English-only classes. Schools' judgments on a student's ability to demonstrate knowledge in reading and math may also be subjective.

The results of research on the classification and inclusion of LEP students in large-scale assessments such as NAEP suggest that relying more on objective test scores (English proficiency and achievement tests) would provide more consistent results. However, the English proficiency and achievement tests must be based on sound psychometric standards and must provide evidence of alignment with English proficiency and test content standards. These standards should be consistent across the nation.

Cut scores from these tests, which are obtained using sound methodology, can then inform the classification of LEP students and their inclusion in NAEP. Using similar cutscores for local, state, and NAEP would lead to a higher level of consistency in inclusion decisions. On the other hand, using different criteria for classification and inclusion decisions, as is the case in many states across the

country, may cause more inconsistencies. The issues related to consistencies in inclusion decisions are of paramount importance in the NCLB accountability system. In addition to being the *Nation's Report Card*, NAEP will be used as national criteria to compare state achievement levels across the nation. Lack of consistency in inclusion decisions may jeopardize this very important recent NAEP mission, which is vital to understanding the trend of students' performance in the nation.

REFERENCES

Abedi, J. (2004, In Press) The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues. *Educational Researcher*, 33(1).

Abedi, J. (2003). The validity of the classification system for students with limited English proficiency: A criterion-related approach. (submitted for publication).

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

Abedi, J., Lord, C., & Hofstetter, C. (1997a). *The impact of different types of accommodations on students with the limited language proficiency*. Los Angeles: University of California, Center for the Study of Evaluation.

Abedi, J., Lord, C., & Hofstetter, C. (1997b). *The impact of students' language background variables on their NAEP mathematics performance*. Los Angeles: University of California, Center for the Study of Evaluation.

Abedi, J., Lord, C., & Plummer, J. (1995). *Language background as a variable in NAEP mathematics performance: NAEP TRP Task 3D: Language background study*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

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

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

August, D. (1994). Overview, inclusion guidelines and accommodations for limited English proficient students in the National Assessment of Educational Progress. In *Briefing book, conference on inclusion guidelines and accommodations for limited English proficient students in the National Assessment of Educational Progress*. Washington DC: National Center for Education Statistics.

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

Butler, F. A., & Castellon-Wellington, M. (2000). *Students' concurrent performance on tests of English language proficiency and academic achievement* (Final Deliverable to OERI/OBEMLA, Contract No. R305B60002, pp. 51-83). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

Hopstock, P. J., Bucaro, B. J., Fleischman, H. L., Zehler, A. M., & Eu, H. (1993). *Descriptive study of services to limited English proficient students. Vol. II: Survey results* (Rep. to the U.S. Department of Education, Office of Policy and Planning). Arlington, VA: Development Associates.

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

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

NCES (2001). The NAEP 1998 Technical Report. Washington, DC: National Center for Education Statistics.

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

Teachers of English to Speakers of Other Languages (TESOL). (1997). *ESL standards for pre-K-12 students*. Alexandria, VA: TESOL

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.