

The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues

by Jamal Abedi

There are major issues involved with the disaggregated No Child Left Behind (NCLB) Act in terms of its adequate yearly progress reporting for students with limited English proficiency (LEP). Inconsistent LEP classification, as well as the sparse population of LEP students in many states, threatens the validity of adequate yearly progress reporting. The LEP subgroup's lack of stability also threatens accountability, since students attaining English proficiency move out of the subgroup. The linguistic complexity of assessment tools may lower LEP student performance in areas with greater language demand. Finally, schools with larger numbers of LEP students with lower baselines may require greater gains. Thus, NCLB's mandates may unintentionally place undue pressure on schools with high numbers of LEP students. Continuing efforts to remedy these issues should bring more fair assessment and accountability.

The No Child Left Behind Act (NCLB; Public Law No. 107-110, 115 Stat. 1425, 2002), the most recent reauthorization of the Elementary and Secondary Act of 1965, holds states using federal funds accountable for student academic achievement. States are required to develop a set of high-quality, yearly student academic assessments that include, at a minimum, assessments in reading/language arts, mathematics, and science. Each year they must report student progress in terms of percentage of students scoring at the "proficient" level or higher. This reporting is referred to as adequate yearly progress (AYP). A state's definition of AYP should also include high school graduation rates and an additional indicator for middle and elementary schools. Each state establishes a timeline for all students to reach the "proficient" level or higher, which must be no more than 12 years after the start date of the 2001–2002 school year, provided that the first increase occurs within the first 2 years.

AYP will be reported for schools, school districts, and the state for all students. In addition, AYP must be reported for the following subgroup categories of students: (a) economically disadvantaged students, (b) students from major racial and ethnic groups,¹ (c) students with disabilities, and (d) students with limited English proficiency (LEP). Students in the LEP² subgroup provide a useful focus for discussing critical issues regarding AYP subgroup reports. Students in the other three subgroup categories share some of the issues pertinent to assessing LEP stu-

dents, and many LEP students are also members of at least one other subgroup category.

Technical issues relating to the testing of LEP students merit discussion. However, a thorough discussion of issues related to the education and testing of LEP students is beyond the scope of this article. The focus on AYP reporting for LEP students at this juncture is important because, although issues concerning their assessment have received attention for many years, educational inequity issues have yet to be resolved. This is especially pertinent as this population continues to increase rapidly in size, with particularly high concentrations in a few states. According to the most recent educational statistics (i.e., those for the 2000–2001 school year), the total number of students labeled as LEP in the nation's public schools is more than 4.5 million (or 9.6% of total enrollment; National Center for Education Statistics [NCES], 2002). This article discusses six LEP assessment issues as they relate to AYP reporting:

1. *Inconsistency in LEP classification across and within states.* Different states and even different districts and schools within a state use different LEP classification criteria, thus causing inconsistencies in LEP classification/reclassification across different educational agencies. This directly affects the accuracy of AYP reporting for LEP students.
2. *Sparse LEP population.* The number of LEP students varies across the nation, and, in the case of a large number of states and districts, the number of LEP students is not enough for any meaningful analyses. This might skew some states' accountability and adversely affect state and federal policy decisions.
3. *Lack of LEP subgroup stability.* A student's LEP status is not stable over time, and a school's LEP population is a moving target. When a student's level of English proficiency has improved to a level considered "proficient," that student is moved out of the LEP subgroup. Those who remain are low performing, and new students with even lower levels of language proficiency may also move into the subgroup. Therefore, even with the best resources, there is not much chance for improving the AYP indicator of the LEP subgroup over time.
4. *Measurement quality of AYP instruments for LEP students.* Students' yearly progress is measured by their performance on state-defined academic achievement tests, but studies have shown that academic achievement tests that are constructed and normed for native English speakers have lower reliability and validity for LEP populations (Abedi, Leon, & Mirocha, 2003). Therefore, results of these tests should not be interpreted for LEP students as they are for non-LEP students.

5. *LEP baseline scores.* Schools with high numbers of LEP students have lower baseline scores, which have year-to-year progress goals that are much more challenging and might be considered unrealistic, considering that their students may continue to struggle with the same academic disadvantages and limited school resources as before.
6. *LEP cutoff points.* Earlier legislation adopted a compensatory model in which students' higher scores in content areas with less language demand (such as math) could compensate for their scores in areas (such as reading) with higher language demands. NCLB, however, is based on a conjunctive model in which students should score at a "proficient" level in all content areas required for AYP reporting. This makes the AYP requirement more difficult for schools with many LEP students.

While it is quite clear that the NCLB legislators' intention is to improve the performance of subgroups of students who have lagged behind for many years, it might unintentionally place undue test performance pressure on schools with large numbers of targeted students. This is especially unrealistic when schools may still struggle with the same limited school resources as before. Test performance pressure may still be a reality in spite of any extra resources NCLB may provide to prevent achievement lag (as part of both Titles I and III). The situation might also create divisiveness between parents and even students. For example, students in poor-performing subgroups might be blamed for a school's poor performance rating. Parents of other students might make the AYP situation worse by moving their children to other schools. Teachers might blame students if the school receives sanctions. The following elaborates on these points.

Inconsistency in LEP Classifications Across and Within States

To begin discussing LEP students' AYP, we need to define the LEP population. The NCLB defines LEP students as (a) being 3 to 21 years of age, (b) enrolled or preparing to enroll in elementary or secondary school, (c) either not born in the United States or speaking a language other than English, and (d) owing to difficulty in speaking, reading, writing, or understanding English, not meeting the state's proficient level of achievement to successfully achieve in English-only classrooms.

The operational definition of LEP varies considerably across schools, districts, and states. Among the many different criteria introduced by NCLB and states for classification of LEP, the most important are (a) being a nonnative speaker of English and (b) scoring low on English proficiency tests. In school districts in several states, the first criterion, being a nonnative English speaker, is based on information garnered from a home language survey. Unfortunately, the validity of this survey is threatened by parents' concerns over equity of opportunity for their children, citizenship issues, and parents' literacy level (Abedi, 2003).

Abedi, Lord, and Plummer (1997) found significant discrepancies between student reporting and the school records of students speaking a language other than English at home. The school record of the number of students speaking a language other than English at home was significantly lower than what the students themselves reported. Another study (Abedi, 2003) showed a low level of relationship between language proficiency

test scores and the LEP classification code. This study reported an average correlation of .223 between scores on the Language Assessment Scales and LEP classification codes across grades, explaining less than 5% of the common variance. The relationship between standardized achievement test (Stanford 9, Iowa Test of Basic Skills) results and LEP classification codes reported in this study was also weak. For example, analyses of data showed that the correlation coefficient between Stanford 9 math concepts and students' LEP code ranged from .045 ($n = 35,981$) to .168 ($n = 25,336$), with an average correlation of .122 (explaining 1.5% of the variance). The correlation between math computation and LEP code ranged from .028 ($n = 36,000$) to .099 ($n = 25,342$), with an average correlation of .069 (explaining less than 0.5% of the common variance between the two variables).

Another issue concerning the LEP subgroup is its heterogeneity. LEP students exhibit differences in level of performance, language proficiency, and family and cultural background characteristics. For example, the results of a study of fourth- and eighth-grade LEP and non-LEP students suggested that parent education is highly related to student performance (Abedi, Leon, & Mirocha, 2003). LEP students of parents with less than a high school education had a mean reading score of 25.23 ($n = 30,091$, $SD = 14.10$), as compared with a mean of 40.35 ($n = 1,649$, $SD = 19.56$) for LEP students of parents with a postgraduate education. It is interesting to note that mean reading scores for some LEP students with higher levels of parent education were higher than mean reading scores for non-LEP students with lower levels of parent education. For example, the mean reading score for LEP students whose parents had a postgraduate education ($M = 40.35$, $SD = 19.56$, $n = 1,649$) was higher than the mean for non-LEP students whose parents had less than a high school education ($M = 37.08$, $SD = 17.84$, $n = 16,806$). A similar trend was seen for Grade 8 reading scores, as well as for math content areas (Abedi, Leon, & Mirocha, 2003).

Once again, these data suggest that students labeled as LEP differ substantially in many aspects, including family characteristics, cultural and language background, and level of English language proficiency. Thus, the LEP subgroup is not a well-defined, homogeneous group of students. However, the present discussion of the issues concerning AYP theory and practice for the LEP subgroup continues based on the existing classification of LEP students.

Sparse LEP Population

A serious consideration in valid and reliable AYP reporting is subgroup size. If there are not enough students in a subgroup category to provide *statistically reliable* data, then schools, districts, or states will not be required to provide disaggregated reports for this subgroup category. Linn, Baker, and Herman (2002) explained the technical aspects of disaggregated reporting and the sample size necessary to compile statistically reliable reports on subgroup categories. They indicated that for statewide and large-district reporting, the number of students in these subgroup categories might not be an issue, since there are large enough numbers of students in each subgroup. However, they warned that small districts and individual schools might not be able to report statistically reliable data because of small numbers of students in each subgroup.

To illustrate this issue, Linn et al. (2002) included a table of standard errors of the differences between two independent sample

percentages as a function of number of students. In this table, the standard error of differences in the percentages ranged from 7.1 (with 100 observations per group) to 22.4 (with 10 observations per group). These data suggest that the higher the number of students, the smaller the standard error of difference in percentages. Linn et al. acknowledged the trade-offs between disaggregated reporting and protecting against mistakenly identifying schools for improvement as a result of low statistical reliability. As a conclusion, they suggested a minimum group size of 25 students, which is large enough to provide reasonably statistically reliable results and detailed enough to permit subgroup reporting. However, in order to detect a moderate level of change (e.g., 5 to 6 percentage points), several hundred subjects would be needed (Hill & DePascale, 2003).

Different states have different numbers of LEP students with different backgrounds. The number of LEP students across the states in the 2000–2001 school year ranged from less than 1,000 in Vermont (1% of the total student population) to more than 1.5 million in California (25% of the total student population). In 31 of the 50 states, LEP students account for less than 5% of the state’s total student population, and in 13 states LEP students account for less than 1% of the student population (NCES, 2002). Dividing the already small number of students in these states across district and student background variables reduces the total LEP enrollment in some districts to a level that might not be sufficient to perform any meaningful statistical analyses. On the other hand, the consequences of excluding LEP students from AYP reporting would be grave, because LEP students’ test results might then be excluded from subgroup accountability determinations and from state and federal policy decisions.

Furthermore, since LEP student populations in different parts of the country are of different cultural and language backgrounds, excluding LEP students in smaller communities from AYP reporting may give more weight to the results obtained for LEP students in larger communities. For example, the majority of LEP students in the nation (more than 76%) have Spanish as their home language (NCES, 2002). The other 24% of LEP students come from varying language, economic, and cultural backgrounds that might produce different academic performance

results. However, owing to smaller numbers and sometimes suburban locations, they might be excluded from AYP reporting, while the results for LEP students in larger communities might be overgeneralized.

Lack of LEP Subgroup Stability

A major problem in AYP reporting for LEP students is the lack of stability of the LEP subgroup. This lack of stability is due to systematic rather than random fluctuation. The LEP subgroup is the least stable among the four subgroup categories targeted for reporting by NCLB. When an LEP student makes significant progress in math and reading (the main subject area focuses of NCLB), he or she will be reclassified as fluent English proficient (FEP) and will no longer be part of the LEP subgroup. Therefore, members of the LEP subgroup, by definition, will almost always be among the low-performing group of students and will hardly make substantial progress. In addition, new students who continually move into schools at lower levels of language proficiency will contribute to the situation of instability. Thus, schools with large numbers of LEP students will continue to remain in the “in need of improvement” category.

In response to this risk caused by the revolving-door nature of LEP populations, several states have proposed plans that approach a “once LEP, always LEP” classification policy for AYP reporting. These states will include “exited” LEP students in the LEP subgroup by expanding exit criteria to include years in which the students’ progress is monitored (Erpenbach, Fortefast, & Potts, 2003). As a means of illustrating the effect of LEP subgroup instability on test scores, a cohort³ of about 14,000 LEP students was followed for a period of seven semesters, from Grade 9 (in fall 1996) to Grade 12 (in fall 1999). Students who were reclassified as non-LEP were compared with those who remained in the LEP category. For these comparisons, median percentile scores in reading and math were used. Table 1 presents the results.

As Table 1 shows, at the starting point (fall 1996), all students in the cohort had been classified as LEP. Median percentile scores for this group were 12 ($n = 13,989$) in reading and 21 ($n = 14,151$) in math. After each semester, some of these students who had

Table 1
Grade 9 Fall 1996 LEP Cohort SAT 9 Percentile Rank Medians

	Reading SAT 9 (<i>n</i>)		Math SAT 9 (<i>n</i>)	
	LEP	FEP	LEP	FEP
Grade 9, fall 1996	12 (13,989)	NA (0)	21 (14,151)	NA (0)
Grade 9, spring 1997	12 (13,255)	21 (659)	20 (13,402)	32 (674)
Grade 10, fall 1997	8 (8,300)	15 (1,313)	21 (8,456)	30 (1,324)
Grade 10, spring 1998	8 (7,549)	14 (1,987)	19 (7,694)	28 (2,009)
Grade 11, fall 1998	6 (5,435)	13 (2,447)	19 (5,523)	26 (2,463)
Grade 11, spring 1999	7 (4,701)	19 (3,217)	20 (4,807)	30 (3,242)
Grade 12, fall 1999	7 (3,809)	18 (3,685)	20 (3,885)	31 (3,712)

Note: NA = not applicable.

made progress were reclassified as FEP. For example, in spring 1997, about 5% of the LEP students were classified as FEP. The median percentile scores of LEP students remained about the same in both reading ($Mdn = 12$, $n = 13,255$) and math ($Mdn = 20$, $n = 13,402$), but the FEP students showed substantially higher performance in reading ($Mdn = 21$, $n = 659$) than those who continued to be classified as LEP.

Major differences between the LEP and FEP students were also observed in the subsequent semesters. The median percentile score of LEP students in reading decreased from 12 in spring 1997 to 8 in fall 1997, and the median score of FEP students decreased from 21 to 15. In math, however, performance remained almost unchanged from spring 1997 to fall 1997 for both LEP (20 in spring 1997 and 21 in fall 1997) and FEP (32 in spring 1997 and 30 in fall 1997) students. In the subsequent semesters, while the performance of both LEP and FEP students remained the same with minor fluctuations, the gap between the performance of LEP and FEP students became substantial. For example, in the last semester (fall 1999), the median reading percentile score for LEP students was 7 ($n = 3,809$), as compared with a median reading score of 18 ($n = 3,685$) for FEP students. For math, the median percentile score for LEP students was 20 ($n = 3,885$), as compared with a median of 31 ($n = 3,712$) for FEP students. While, as the data suggest, both the LEP and FEP students performed well below their native English-speaking peers, the gap between LEP and FEP students remained high. These data once again suggest that language proficiency is inevitably a strong determiner of test performance, a fact reflected in the difference between the performance of LEP and non-LEP students on linguistically complex content area test items (e.g., see Abedi, Courtney, & Leon, 2003).

Measurement Quality of AYP Instruments: Impact of Language Complexity on LEP Assessment

A concern specific to LEP students is the impact of language factors on their assessments. Because of the confounding of test language comprehension with student demonstration of content knowledge, LEP students may show improvement in content knowledge (such as math) only when their level of academic English proficiency increases (Abedi & Lord, 2001). However, the LEP population is perpetually growing, and students are often assessed in content areas without proper time to develop sufficient English proficiency for valid testing. Thus, schools with larger numbers of LEP students are more likely to be cited as being "in need of improvement" than schools with fewer or no LEP students.

As specified in the NCLB, state-defined achievement tests are used in measuring students' yearly progress. Most states use different kinds of standardized achievement tests. These tests might function well for measuring the academic progress of native English speakers; however, the language complexity of test items in content-based assessments makes the reliability and validity of these tests suspect for LEP students (Abedi, 2002). 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 LEP students. In addition, as a result of the influence of students' language background on their assessment, these tests might underestimate LEP students' performance in content-based areas.

Since most standardized, content-based tests are conducted in English and field tested with mostly native English speakers, they might inadvertently function as English language proficiency tests. LEP students might have trouble demonstrating their content knowledge because they are unfamiliar with the complex linguistic structure of the questions, they might not recognize certain vocabulary forms, or they might mistakenly interpret an item literally (Duran, 1989; Garcia, 1991). Also, they may not perform as well on tests because they read more slowly (Mestre, 1988). In addition, issues related to standardized achievement tests are more profound with norm-referenced tests than criterion-referenced tests. In the case of LEP students, many states still use norm-referenced tests for AYP reporting (Erpenbach et al., 2003).

Research has demonstrated that language background affects students' performance, particularly in content-based assessments (Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000; Solano-Flores & Trumbull, 2003). A student possessing content knowledge, such as in mathematics, science, or history, is not likely to demonstrate this knowledge effectively if she or he cannot interpret the vocabulary and linguistic structures of the test. Minor changes in the wording of content-related test items can raise student performance (Abedi & Lord, 2001; Abedi et al., 1997; Cummins, Kintsch, Reusser, & Weismer, 1988; De Corte, Verschaffel, & DeWin, 1985; Hudson, 1983; Riley, Greeno, & Heller, 1983). Accordingly, one approach to testing LEP students involves rewording test items to minimize construct-irrelevant linguistic complexity.

Recent studies have used the linguistic modification approach as an alternative in the assessment of LEP students. These studies compared student scores on NAEP original test items with tests containing parallel items in which the mathematics task and terminology were retained but noncontent vocabulary and linguistic structures were modified. The results of these studies consistently show higher performance for LEP students on linguistically modified test items (Abedi & Lord, 2001; Abedi et al., 1997, 2000; Kiplinger, Haug, & Abedi, 2000; Maihoff, 2002).

Some linguistic features slow down the reader, make misinterpretation more likely, and add to the reader's cognitive load, thus interfering with concurrent tasks. Indexes of language difficulty include word frequency/familiarity, word length, and sentence length. Other linguistic features that might cause difficulty for readers include passive voice constructions, comparative structures, prepositional phrases, sentence and discourse structure, subordinate clauses, conditional clauses, relative clauses, concrete versus abstract or impersonal presentations, and negation.

To illustrate the impact of language on content-based assessments, a brief discussion is provided of results from analyses of extant data (see Note 3) in which the performance of LEP and non-LEP students was compared on math analytical, math concepts, estimation, problem solving, and math computation involving varying degrees of language demand. Performance differences were estimated in terms of effect sizes (Cohen, 1988; Kirk, 1995, pp. 180–182). There were 996 LEP and 13,054 non-LEP students in the Grade 3 sample, 726 LEP and 12,628 non-LEP students in the Grade 6 sample, and 692 LEP and 11,792 non-LEP students in the Grade 8 sample. Table 2 shows effect sizes along with the numbers of students in Grades 3, 6, and 8 in reading, math calculation, and math analytical. As the data in

Table 2
Numbers of LEP and Non-LEP Students and Effect Size Estimates

Grade	Number of students		Effect size		
	LEP	Non-LEP	Reading	Math calculation	Math analytical
3	996	13,054	.18	.07	.15
6	726	12,628	.24	.09	.18
8	692	11,792	.22	.09	.15

Table 2 show, results were consistent across the three grade levels. For example, reading effect sizes were .18 in Grade 3, .24 in Grade 6, and .22 in Grade 8. The corresponding effect sizes were .07, .09, and .09 for math calculation and .15, .18, and .15 for math analytical.

For reading, the effect sizes across the grade levels ranged from .18 for Grade 3 to .24 for Grade 6. These effect sizes could be considered medium. For math analytical, the effect sizes ranged between .15 for Grades 3 and 8 to .18 for Grade 6; these effect sizes were substantially smaller than those for reading. For math calculation, the effect sizes ranged between .09 for Grades 6 and 8 to .07 for Grade 3. These effect sizes for math calculation were smaller than those for math analytical and much smaller than those for reading. The smaller the effect size, the smaller the performance gap between LEP and non-LEP students.

The results of these analyses suggest that the performance difference between LEP and non-LEP students was the largest in reading (the highest level of language demand) and the smallest in math calculation (the lowest level of language demand). Averaging over the three grades, effect sizes were .213 for reading, .160 for math analytical, and .083 for math calculation. The results of the analyses also show that the effect sizes were relatively smaller for lower grades (Grade 3) and became larger as the grade levels in-

creased. This might also have been due to language factors, since there is greater language demand in higher grade assessments.

Results indicated as well that test items for LEP students, particularly those at the lower end of the English proficiency spectrum, suffer from lower reliability. To illustrate this point, the reliability of Stanford 9 achievement tests for LEP and non-LEP students was estimated by computing internal consistency (see Note 3). Table 3 presents results of internal consistency analyses for math, language, science, and social science items. Internal consistency (alpha) coefficients were computed separately for LEP, FEP, and native speakers (English only).

As the data in Table 3 show, alpha coefficients were highest for the English-only group, lower for the FEP students (who were nonnative English speakers reclassified as fluent), and lowest for the LEP students. The sizes of the alpha coefficients for English-only students were relatively stable across the content areas, ranging from a high of .898 for math to a low of .805 for science and social science. Among LEP students, however, alpha coefficients differed considerably across the content areas. In math, where language factors might not have much influence on performance, the coefficient for LEP students (.802) was slightly lower than the coefficient for English-only students (.898). In language, science, and social science, however, the alpha coefficient gap between

Table 3
Grade 9 Stanford 9 Subscale Reliabilities and Standard Deviations

Subscale (number of items)	English only (approximate N = 180,000)		FEP (approximate N = 38,000)		LEP (approximate N = 53,000)	
	α	SD	α	SD	α	SD
Math						
Total (48)	.898	9.58	.898	9.603	.802	6.941
Language						
Mechanics (24)	.803	5.56	.802	5.469	.686	4.593
Expression (24)	.823	5.78	.804	5.522	.680	4.732
Average	.813	5.67	.803	5.496	.683	4.663
Science						
Total (40)	.805	6.52	.778	6.104	.597	4.694
Social Science						
Total (40)	.805	16.83	.784	15.748	.530	12.777

English-only and LEP students was large. Averaging over language, science, and social science results, the alpha coefficient for English-only students was .808, as compared with an average coefficient of .603 for LEP students. Thus, language factors introduce another source of measurement error in LEP student test results that might not have much impact on native/fluent speakers of English (see also Abedi, 2002; Solano-Flores & Trumbull, 2003).

The results also showed that the correlation between standardized achievement test scores and other valid achievement indicators was significantly larger for the non-LEP than the LEP population. Structural models for LEP students demonstrated lower statistical fits. Factor loadings were generally lower for LEP students, and the correlations between the latent content-based variables were weaker for these students. Results suggested that language factors might cause such differences between LEP and non-LEP groups by creating a restricted range distribution of scores. Thus, language factors act as construct-irrelevant sources (Messick, 1994).

The data just summarized on the impact of language on the performance of LEP students and on LEP/non-LEP differences in psychometric characteristics of tests clearly suggest that assessment results are not directly comparable across the LEP and non-LEP groups. The data also show that, as a result of confounding of language and content, the performance of LEP students may be underestimated; thus, schools, districts, and states with larger numbers of LEP students must expend a substantially higher level of effort to satisfy the NCLB requirement of performance increases by the target date of no later than 2014.

LEP Baseline Scores

Obviously, schools differ in terms of their resources, students' opportunity to learn, students' socioeconomic status, and education levels of students' parents. Some of these differences have been shown to correlate with students' performance on standardized achievement test scores (Abedi, Leon, & Mirocha, 2003). Schools are required to define a starting point or baseline for AYP based on scores from a state-defined achievement test administered during the 2001–2002 school year. Consequently, schools enter into the NCLB race at very different starting points. In general, schools with larger numbers of students in the LEP category will start with lower baseline scores. It is obvious that schools with lower baseline scores will have to spend more time and resources—significantly more than schools with higher baseline scores—in order to reach the level of proficiency by their target year (i.e., no later than 2014).

As an example, consider two schools with two different starting points. At School A, 78% of students are categorized as proficient or higher based on a 2001–2002 measure of academic achievement in reading/language arts and math. At School B, however, the starting point is 25%. At School A, annual performance needs to increase by less than 2% [$(100 - 78)/12 = 1.83\%$], while, in order for School B to satisfy the AYP requirement, it must have a yearly increase of more than 6%.

Thus, schools with LEP students have double duty. Not only must they excel in helping students learn more in content-based areas such as math, but they must also help them become more proficient in English so that they can better follow instructions

and understand test questions. Schools not making adequate yearly progress will be deemed as “in need of improvement” and might receive sanctions. For example, schools failing to make AYP for 4 consecutive years might be required to replace staff, fully implement a new curriculum, continue to offer public school choice, and provide supplemental services. The district will take these corrective actions even if a single subgroup of students fails to show sufficient progress. However, various economic, social, cultural, physical, and/or linguistic factors are impediments to academic progress as well as to the valid and reliable measurement of the progress of the targeted subgroups. For these students, making progress has always required extraordinary school resources, and measuring such progress often requires improved testing tools and/or procedures.

Multiple Criteria and Cutoff Points in AYP

As mentioned, the NCLB is the most recent reauthorization of the Elementary and Secondary Education Act of 1965. The Council of Chief State School Officers (CCSSO; 2002) has elaborated on the accountability requirement differences between the 1994 reauthorization, known as the Improving America's Schools Act (IASA), and the 2001 reauthorization. Among the major differences are changes in the direction and emphasis of accountability. The IASA applies a *compensatory* model for accountability purposes. In this model, higher performance in one subject area will compensate for lower performance in another subject area. For example, higher performance in math may compensate for lower performance in reading/language arts. In contrast, NCLB applies a *conjunctive* model in which scores on all of the measures that are required for AYP must be above the criterion point or cut scores (CCSSO, 2002).

These two approaches may lead to different outcomes. As a means of illustrating this point, a comparison was made of compensatory and conjunctive models using data from a state with a large number of LEP students (see Note 3). This comparison involved the use of the cutoff point of the 36th percentile score established and used by the state. Based on the compensatory model, a student can be reclassified as non-LEP or be placed in the “pass” category if a higher score in one area can compensate for a lower score in another. For example, if a student obtains percentile scores of 29 in reading and 43 in math, then the higher math score (7 percentile points higher) will help compensate for the lower reading score (7 percentile points lower). However, if the conjunctive model is used, this student will “fail” since her reading score is below the cutoff point of the 36th percentile score, regardless of her math score. Table 4 presents the results of the analyses comparing the two models.

As the data in Table 4 suggest, the two models produce very different results. The conjunctive model is more conservative than the compensatory model in recognizing students' progress. For example, among Grade 4 students, 2,227 or 10% of LEP students were placed in the “pass” category under the conjunctive model; in contrast, 20% of these students were placed in the “pass” category based on the compensatory model. In Grades 7 and 11, smaller percentages of LEP students than in Grade 4 were placed in the “pass” category based on both models. However, the difference between outcomes based on the two models was large. In Grade 7, 3.4% of LEP students were placed in the

Table 4
Comparison of Conjunctive and Compensatory Methods, 1999–2000

		Conjunctive method		Compensatory method		Total
		Fail	Pass	Fail	Pass	
LEP students						
Grade 4	<i>N</i>	20,003	2,227	17,784	4,446	22,230
	%	90.0	10.0	80.0	20.0	100.0
Grade 7	<i>N</i>	10,455	363	9,979	839	10,818
	%	96.6	3.4	92.2	7.8	100.0
Grade 11	<i>N</i>	3,527	119	3,170	476	3,646
	%	96.7	3.3	86.9	13.1	100.0
Non-LEP students						
Grade 4	<i>N</i>	14,642	14,602	10,787	18,457	29,244
	%	50.1	49.9	36.9	63.1	100.0
Grade 7	<i>N</i>	18,457	12,167	14,885	15,739	30,624
	%	60.3	39.7	48.6	51.4	100.0
Grade 11	<i>N</i>	12,998	8,271	9,732	11,537	21,269
	%	61.1	38.9	45.8	54.2	100.0

“pass” category in the conjunctive model, as compared with 7.8% in the compensatory model. Similarly, in Grade 11, 3.3% of students were placed in the “pass” category based on the conjunctive model, as compared with 13.1% based on the compensatory model. Among non-LEP students, the difference between the conjunctive and compensatory models was also larger (see Table 4).

Based on these data, it is quite clear that NCLB is more strict in terms of criteria to judge students’ performance. The issues of compensatory versus conjunctive cutoff points are more pronounced for LEP students. As explained earlier, as a result of the impact of linguistic factors on assessment, LEP students have more difficulty with content areas high in language demand. For example, it has been demonstrated that LEP students have more difficulty in reading than in math. Even within the math content, they have more difficulty with items that are more linguistically demanding, such as problem solving. In general, there is a much larger gap between LEP and non-LEP students in reading than in math. Therefore, LEP students are more likely to stay in the “fail” category for a substantial period of time owing to their low scores in reading.

Other Factors Affecting AYP

The AYP measurement of LEP students is also affected by other factors, such as students’ current capacity to understand instruction. As a result of English language barriers, LEP students may not benefit from teacher instruction at the same level as their non-LEP peers. Even when schools provide “sheltered English” classes in content subjects, LEP students may not attain content mastery. Results of a recent study (Abedi, Herman, Courtney, Leon, & Kao, 2004) involving more than 600 Grade 8 LEP and non-LEP students in math revealed that LEP students reported significantly less opportunity to learn than their non-LEP peers. Interestingly, in the observation phase of this study, the results

showed that LEP students were less outwardly involved in classroom activities. They raised their hands less often than non-LEP students, and, when they did, teachers did not call on them as often as the non-LEP students. If LEP students require more time and practice to attain mastery in their language and content studies because of language and/or cultural factors, then their need for a higher level of opportunity to learn may directly affect their achievement measure results and reflect poorly on schools that are actually performing well. More research needs to be done to explore the influence of other factors on the validity of AYP reporting for LEP students.

Discussion

The disaggregated progress reports by subgroup mandated by the NCLB legislation will monitor the nation’s goal of having “no child left behind.” However, there are major issues in this disaggregated reporting among different subgroup categories (students who are economically disadvantaged, students from major racial and ethnic groups, students with disabilities, and LEP students). The NCLB requirement for subgroup reporting may give the impression that students in the subgroup categories start the achievement race at about the same level and can progress with other students at about the same rate. This might be an overly optimistic view of the situation of less advantaged learners. By focusing this discussion on the consequences for schools enrolling LEP students, we see how putting into practice the policy may produce invalid assessment and unreliable reporting while exacerbating the burdens of current educators. Following is a discussion of some challenges in AYP measurement and reporting for LEP students.

The results of research on the assessment of LEP students reported in this article and elsewhere suggest a strong confounding of language and performance. LEP students exhibit substantially lower performance than non-LEP students in subject areas high

in language demand. The study findings suggest that the large performance gap between LEP and non-LEP may *not* be due mainly to lack of content knowledge. LEP students may possess the content knowledge but may not be at the level of English language proficiency necessary to understand the linguistic structure of assessment tools. The strong confounding of language factors and content-based knowledge makes assessment and accountability complex for LEP students and, very likely, students in other targeted subgroups.

Because of the strong effect of language factors on the instruction and assessment of LEP students, they lag far behind native English speakers. This leads to huge initial differences. That is, LEP students start with substantially lower *baseline* scores. More important, unless LEP students' English language proficiency is improved to the level of native English speakers—which is not an easy task—they will not be able to move at the same rate on the AYP progress line as do native English speakers.

It is clear that NCLB cannot have much of an effect on the initial performance differences between LEP and non-LEP students. A more sensible question here is whether or not NCLB can provide enough resources to schools with large numbers of LEP students to help them increase these students' language proficiency to a sufficient extent that they can progress with their native English speaker peers in both instruction and assessment.

Inconsistency in LEP classification across and within states makes AYP reporting for LEP students even more complex. If students are not correctly identified as LEP, how can their AYP be reliably reported at a subgroup level? Although NCLB attempts to resolve this issue by providing a definition for this group, its criteria for classifying LEP students may face the same problems as the existing classification system (Abedi, 2003; Zehler, Hopstock, Fleischman, & Greniuk, 1994).

Inconsistency in the classification of LEP students may lead to more heterogeneity in the LEP subgroup. With a more heterogeneous population, larger numbers of students are needed to provide the statistically reliable results required by NCLB. However, as elaborated here, the population of LEP students in many districts and states is sparse. In many states, there may not be enough students in a district or school to satisfy even the minimum number of 25 students suggested in the literature (Linn et al., 2002). As indicated earlier, other researchers have argued that even 25 students may not be enough to provide statistically reliable results and have proposed a minimum group size of 100 students (Hill & DePascale, 2003). Considering the small number of LEP students in many districts and states, the small group size for LEP reporting would be another obstacle in regard to reliable AYP reporting.

The LEP subgroup suffers from yet another major problem related to AYP reporting: the lack of stability of this group. In many states and districts across the nation, LEP students' level of English proficiency is reevaluated regularly, and if they reach a proficient level of English proficiency, they move out of the LEP subgroup. While this helps the more English-proficient students receive more appropriate instruction and assessment, it results in the LEP subgroup continuing to be low performing. Thus, the students in this group will always be labeled as underachievers, and schools with large numbers of LEP students will be stuck in the "need for improvement" category.

Some states with substantial numbers of LEP students have expressed concern over this issue. They have proposed ideas and negotiated with the federal government to ease the level of possible negative impact that this situation may have on school, district, and state accountability. For example, Indiana and Delaware will continue to include exited LEP students in the LEP subgroup for 2 years after they have been determined to be proficient in English. Georgia plans to include LEP students as long as they still receive services through the English for Speakers of Other Languages program, even if they have met exit criteria (Erpenbach et al., 2003). In California, students redesignated as FEP will remain in the LEP category until they reach the proficient or above level on the California Standards Test in English-language arts for 3 consecutive years (California Department of Education, 2003); however, the question of whether this policy will provide a long-term solution to the problem of LEP subgroup instability or serve only as temporary relief remains unanswered.

Thus, measurement of the academic achievement of LEP students is much more complex than what the NCLB legislation conceives. A fair assessment of students in the four targeted subgroup categories requires much more serious consideration than is outlined in the law. Despite attempting to solve the age-old problem of heterogeneity among LEP students, the NCLB seems to perpetuate it, thereby leaving more room for children to be left behind.

On the other hand, I believe that the NCLB's attention to students in the four subgroup categories in general and to the LEP population in particular is a step in the right direction. It is promising, for example, to see that Title III of NCLB requires assessment of LEP students' English proficiency on an annual basis and is providing support to states to develop reliable and valid measures of students' proficiency. However, I believe that any decisions concerning assessment for all subgroups, particularly LEP students, must be informed by results of research and experience in the education community. I elaborate this point by discussing issues concerning states' development of English language proficiency measures. This may provide a good example of how recommendations provided in the NCLB might be implemented.

There are many existing tests for measuring students' level of English language proficiency. Some of these tests have been used frequently and over many years by different states and districts. In spite of the existence of such tests, states are developing new English language proficiency tests with funding through the NCLB's Enhanced Assessment Instruments. A reasonable explanation for this might be that states did not find that the existing tests provided reliable and valid measures of students' level of English language proficiency as required by NCLB. If this is the reason for the development of new tests, then the test developers should be aware of problems in the existing tests to avoid the same problems in the new tests.

For example, a careful review of some of the most commonly used language proficiency tests concluded that the tests differ considerably in types of tasks and specific item content and are based on different theoretical emphases prevalent at the time of their development (Zehler et al., 1994). This suggests that, in the case of some of the existing tests, the English language proficiency domain was not operationally defined before the test development process. This and similar studies and reviews should inform the develop-

ment process of new tests. For example, it is imperative that before any effort in developing an 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 should also be considered (see Bailey & Butler, 2003).

Furthermore, in analyzing data from the administration of existing language proficiency tests, researchers have expressed concerns with the reliability and validity of these tests, the adequacy of the scoring directions, and the limited populations on which test norms are based. For example, analyses of several large data sets from different locations across the nation have shown validity problems in predicting LEP classification and lack of power in identifying different levels of English language proficiency among the LEP student population (Abedi, 2003; Abedi, Leon, & Mirocha, 2003). Those involved in the development of new English language proficiency tests should learn from such research and should conduct more analyses on the wealth of data that exist in this area. To be considered valid and reliable measures of English language proficiency, as outlined in the NCLB, new tests must first go through a rigorous validation process. Otherwise, there may not be a reasonable justification to spend the limited NCLB resources on English language proficiency test development.

As a final thought, assessment and accountability of LEP students cannot be pursued in isolation of other important factors. An effective education system for LEP students that may lead to a successful AYP outcome should include at least three interactive components (see Figure 1): (a) classification, (b) instruction, and (c) assessment. A problem in any one of these components may affect the other two. For example, a student misclassified as an LEP student may be assigned a different curriculum and thus receive inappropriate instruction. Alternately, inappropriate instruction may result in low performance, which may in turn result in misclassification. While each component has its unique role, they share common ground: the effect of language factors or barriers. For example, as explained earlier, unnecessary linguistic

complexity of assessment may threaten the validity and equitability of assessment among LEP students. Complex linguistic structure of instruction may negatively affect LEP students' ability to understand classroom instruction, and invalid assessment of students' level of English proficiency may result in misclassification. In a positive light, valid assessment may provide diagnostic information that can inform instruction and classification.

An effective way to help LEP students reach proficiency in the AYP model is to consider the broader picture using the interactive model just described. The following are a few critical needs.

1. *Improve current LEP classification and assessment.* There is a need to establish a common definition of English language proficiency and substantially improve the validity of LEP instruments. Among other things, validity of LEP assessment can be enhanced by avoiding cultural biases and reducing unnecessary linguistic complexity of assessments.
2. *Improve monitoring of progress.* Schools need effective and valid data collection methods that can be used to monitor LEP progress at every stage of a student's education. Weaknesses must be quickly addressed with appropriate instructional strategies.
3. *Improve teacher capacity.* LEP students need teachers who are well qualified in both language development and content, each of which plays a crucial role in LEP student achievement. The federal government can play a key role in this process by funding and encouraging programs that improve teacher capacity in this dual role. Teachers of LEP students should receive training in content delivery, language sheltering, and the teaching of academic language.
4. *Consider redesignated LEP students as part of the LEP subgroup that established the baseline score.* State plans allowing redesignated students to remain in the LEP subgroup for only a limited time are temporary fixes. While new LEP students are added to the subgroup, redesignated students should also be retained for AYP reporting. This "semicohort" approach to tracking LEP students allows the progress of redesignated students to be counted toward subgroup AYP progress.

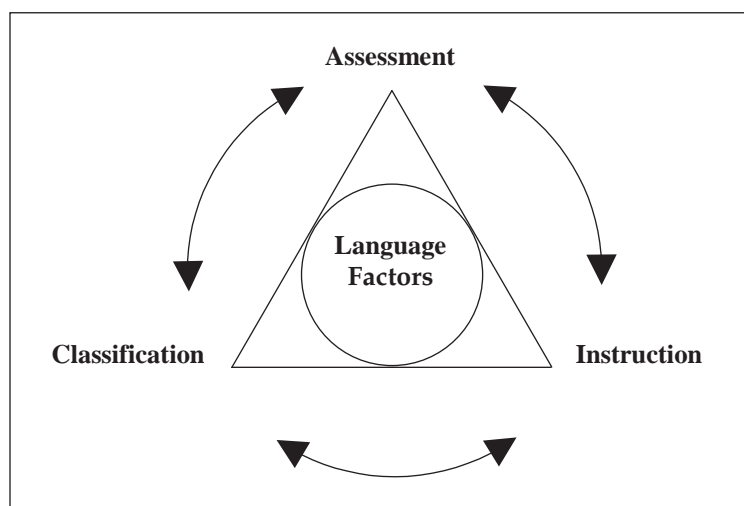


FIGURE 1. *Interactive school achievement model for LEP students.*

The academic progress of subgroups, especially LEP students, is much more complex than conceived by the NCLB. No facet of the challenge should be overlooked. We must continue to explore the many complex interrelationships among the factors that have the greatest influences on LEP achievement.

The purpose of this article has been to raise and discuss issues concerning accountability for LEP student achievement. It is hoped that policymakers will seriously consider these observations when making decisions on the assessment and accountability of LEP students. Based on the results of research presented here and elsewhere, policymakers, lawmakers, and decision makers are urged to take appropriate action to correct the inequities resulting from the NCLB in regard to the subgroups targeted by the legislation, particularly the LEP student subgroup. It is, however, encouraging that states, in collaboration with the federal government, are taking steps to remedy some of these issues. The hope is that these continued efforts will bring more fairness into the assessment of and accountability for LEP students.

NOTES

The work reported here was supported in part under a grant (R305B960002) from the U.S. Department of Education, Office of Educational Research and Improvement. The findings and opinions expressed do not reflect the positions or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

The author acknowledges valuable contribution of colleagues in preparation of this article. Joan Herman, Kathleen Leos, and Ron Dietel contributed to the article with their helpful comments and suggestions. Robert Linn contributed with his review of the article and helpful comments. Mary Courtney and Jenny Kao contributed substantially with comments and assistance in structuring and revising the article. Seth Leon provided valuable assistance with the data analyses. The author is also grateful to Eva Baker and Joan Herman for their support of this work.

¹ The second subgroup category (students from major racial and ethnic groups) is not treated as a single aggregated group under NCLB. Rather, it consists of separate groups (e.g., African American/Black, Hispanic/Latino) as determined by states.

² The author acknowledges the term "English language learner" as an alternative to "LEP." Both refer to students who may be in need of English language instruction, a category that 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 "LEP," which some regard has having a negative connotation (August & Hakuta, 1998). However, in this article, the term LEP is used more often since it is more commonly used in research and practice.

³ Data were obtained from four different U.S. locations. One site was a large public urban school district in which Grades 3, 6, and 8 data were analyzed for the 1998–1999 school year. More than 89,000 students were enrolled in those grades during that school year, and about 14% were characterized as receiving bilingual services. Another site was a state with more than 1 million students enrolled in Grades 2, 7, and 9 during the 1997–1998 school year, of which 17% were LEP students. A third site was an urban school district with more than 22,000 students in Grades 10 and 11 during the 1997–1998 school year, of which 3.4% were LEP students. The fourth site was a state with more than 39,000 students enrolled in Grades 3, 6, and 8 during the 1997–1998 school year, of which 6.8% were LEP students. For further detail regarding these sites, see Abedi, Leon, and Mirocha (2003). In addition to the data

sets just described, data from several filed studies conducted by Abedi and colleagues were used. For reports of these studies, visit the UCLA/CSE Web site at www.cse.ucla.edu.

REFERENCES

- Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment*, 8(3), 231–257.
- Abedi, J. (2003). *The validity of the classification system for students with limited English proficiency: A criterion-related approach*. Manuscript submitted for publication.
- Abedi, J., Courtney, M., & Leon, S. (2003). *Research-supported accommodation for English language learners in NAEP* (CSE Tech. Rep. No. 586). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Herman, J., Courtney, M., Leon, S., & Kao, J. C. (2004). *English language learners and math achievement: A study on classroom level opportunity to learn*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of students' language background on content-based assessment: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14, 219–234.
- 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). *Improving schooling for language-minority children: A research agenda*. 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, National Center for Research on Evaluation, Standards, and Student Testing.
- California Department of Education. (2003). *2002 base adequate yearly progress report* [information guide]. Retrieved July 21, 2003, from <http://www.cde.ca.gov/ayp/2002/aypinfo.pdf>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Council of Chief State School Officers. (2002). *Making valid and reliable decisions in determining adequate yearly progress*. Washington, DC: Author.
- Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, 20, 405–438.
- De Corte, E., Verschaffel, L., & DeWin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology*, 77, 460–470.
- Duran, R. P. (1989). Assessment and instruction of at-risk Hispanic students. *Exceptional Children*, 56, 154–158.
- Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB*. Washington, DC: Council of Chief State School Officers.
- Garcia, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly*, 26, 371–391.

- Hill, R. K., & DePascale, C. A. (2003). Reliability of no child left behind accountability designs. *Educational Measurement: Issues and Practice*, 22(3), 12–20.
- Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development*, 54, 84–90.
- Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). *Measuring math—not reading—on a math assessment: A language accommodations study of English language learners and other special populations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Boston: Brooks/Cole.
- 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, 55–75.
- Linn, R. L., Baker, E. L., & Herman, J. L. (2002). *Minimum group size for measuring adequate yearly progress: The CRESST line*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.
- Maihoff, N. A. (2002, June). *Using Delaware data in making decisions regarding the education of LEP students*. Paper presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, Palm Desert, CA.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13–23.
- Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 200–220). Hillsdale, NJ: Erlbaum.
- National Center for Education Statistics. (2002). *Public school student, staff, and graduate counts by state: School year 2000–01* (NCES Publication 2002-348). Washington, DC: Author.
- Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). New York: Academic Press.
- 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, 3–13.
- 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.

AUTHOR

JAMAL ABEDI is a faculty member at the UCLA Graduate School of Education and Information Studies and a senior researcher at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), 300 Charles E. Young Drive North, GSEIS Building, 3rd Floor, Los Angeles, CA 90095-1522; jabedi@cse.ucla.edu. His recent studies have focused on the effects of linguistic factors and accommodations on English language learners.

Manuscript received April 1, 2003

Revisions received October 2, 2003

Accepted October 16, 2003

Airline Travel Information

AERA is pleased to announce special airline offers for the 2004 AERA Annual Conference, April 12–16, in San Diego, CA:

Delta Air Lines:

Special rates allow you a 5% discount on Delta's published round-trip fares within the continental United States. Call **1-800-241-6760** and talk to a Delta representative. Make sure you mention the AERA File Number: **202389A** to receive your discounts.

United Air Lines:

Special rates are also available from United! Call the United Travel desk at **1-800-521-4041** for 5–10% discounts on select flights within the continental United States. Refer to the special AERA Meeting Plus ID code: **597BT**.