Measuring Students’ Level of English Proficiency: Educational Significance and Assessment Requirements

Jamal Abedi
University of California, Davis

This article discusses the status of existing English language proficiency (ELP) tests, and compares the content coverage and psychometric characteristics of ELP assessments that existed prior to the implementation of the No Child Left Behind Act (NCLB) with those developed after NCLB Title III guidelines were introduced. The article argues that the newly developed ELP assessments (post-NCLB) exhibit higher assessment quality when compared with most ELP tests developed prior to NCLB (pre-NCLB). However, more work needs to be done if the new ELP assessments are to inform the instruction and content-based assessment of English language learner students. The results of analyses suggest that ELP assessment outcomes can help inform decisions regarding participation of English language learner students in state content-based assessments.

INTRODUCTION AND PERSPECTIVE

English language proficiency (ELP) assessment is an extremely important aspect of English language learner (ELL) students’ academic careers as the output of such assessment determines and influences their instruction, classification and promotion. Therefore, providing reliable and valid ELP assessments are most important in determining their academic progress. Assessments of ELP based on questionable measures may cause grave academic consequences. ELL students who are inadequately assessed may be misclassified with respect to their level
of proficiency in English and may receive inappropriate instruction. They may even be misclassified as students with learning disabilities, which may greatly impact their academic career (see, e.g., Abedi, 2006b; Artiles, Rueda, Salazar, & Higareda, 2005).

Furthermore, ELL students’ level of English proficiency is an important criterion in determining their readiness for participating in the state content-based assessments such as reading/language arts, math, and science. Because state content-based assessments that are used for No Child Left Behind Act (NCLB; 2002) Title I accountability purposes are mainly constructed and field tested for students who are fluent in English, they may be subject to linguistic factors that could seriously undermine their validity for ELLs (Abedi, 2006a). It seems reasonable that ELL students may need particular levels of proficiency in English to meaningfully participate in the state content-based assessments written in English.

Because of the importance of adequately assessing a student’s level of ELP, the NCLB legislation requires that schools receiving Title I funding assess ELL students using reliable and valid measures (NCLB, 2002). With this mandate, the legislation plays an important role in bringing the need for English language assessment to the forefront of education accountability.

The main objectives of this article are (a) to provide an overview of the national efforts in the assessment of ELP for ELL students, (b) to review the content and psychometric characteristics of these assessments, and (c) to examine their power in predicting ELL student performance in content-based areas such as math and science, the subject of Title I assessments.

**NCLB Title III Mandates for Assessment**

In addition to emphasizing the need for ELP assessment, the NCLB Title III legislation provides a set of guidelines for constructing ELP assessments that are intended to render reliable and valid estimates of a student’s level of English proficiency (NCLB, 2002). These guidelines provide specific tools to help the measurement community be more vigilant of ELL assessment needs and to be better prepared for assessing ELLs. Literature has documented shortcomings of ELP assessments developed prior to the implementation of NCLB (Abedi, 2007; Zehler, Hopstock, Fleischman, & Greniuk, 1994). Many of the pre-NCLB assessments were not based on an operationally defined concept of English proficiency, had limited academic content coverage, were not consistent with states’ content standards, and had psychometric flaws (Del Vecchio & Guerrero, 1995).

NCLB Title III contributes to improving the quality of ELP assessments in many different ways. By making such assessments a requirement for ELL students, NCLB encourages the education community to pay greater attention
to this area. The NCLB description of ELP assessment requirements also helps define ELP assessments more operationally. For example, NCLB requires ELP assessments to include four domains (reading, writing, speaking, and listening), measure student’s academic English proficiency, and be aligned with the states’ ELP standards—as well as content standards—across three academic topic areas and one nonacademic topic area related to school environment (Fast, Ferrara, & Conrad, 2004). By introducing the concept of academic English and academic content into ELP assessment, NCLB requires states to focus on the aspects of language acquisition that are more directly linked to students’ academic success.

These are significant milestones in the history of ELP assessment. By assessing academic language proficiency, states more thoroughly address language needs related to students’ academic success. Alignment of ELP assessment content with the states’ ELP content standards provides assessments that are relevant to students’ academic needs. Other conditions, such as requiring states to provide evidence for the reliability and validity of their assessments and to assess students’ proficiency in Grades K through 12, also contribute to improved assessment of ELP.

The Newly Created ELP Assessments

The work of four consortia of states that carried out the challenging task of developing post-NCLB assessments illustrates the advances in ELP assessment that have been achieved (for a detailed discussion of the four consortia of states and their test development efforts, see Abedi, 2007). Each consortium produced a test battery that included items in the four domains of reading, writing, listening, and speaking and provided outcome measures for each as well as scores for comprehension (listening and reading) and for overall performance. The consortia also conducted standard-setting procedures to set achievement levels in several categories including below basic, basic, proficient, and above proficient for all four domains. ELP tests were developed for several grade clusters (e.g., kindergarten through Grade 2, Grades 3 through 5, 6 through 8, and 9 through 12). These newly developed assessments underwent extensive pilot testing and field testing on large and representative samples of students. The content and psychometric properties of the individual items, as well as the total tests, were carefully examined and improvements were made.

The newly developed ELP assessments are based on theoretical frameworks of second language acquisition and other principles in the field of linguistics (Bauman, Boals, Cranley, Gottlieb, & Kenyon, 2007; see also Cummins, 1981). The assessments were also informed “by second-language development theory of communicative competence which posits that ELP tests should measure communicative and participatory language in the context of the classroom and that they should be age/grade appropriate” (Lara et al., 2007, p. 48).
In addition to basing target constructs, and grounding the ELP test development process in theories of language learning, each consortium created sets of standards to guide test development that are common across the participating states in the consortium. Formal studies of the alignment of the assessments to these standards also were conducted (see, e.g., Bauman et al., 2007; Lara et al., 2007; Rebarber et al., 2007).

The processes used to develop the consortia assessments provide some important examples of good practice. Rather than focusing on a particular grade or age level, these assessments cover kindergarten through Grade 12, using a grade-cluster approach that grouped test instruments into several clusters. There were substantial numbers of common items built into the assessments to facilitate vertical scaling across the clusters (Bauman et al., 2007; Lara et al., 2007; Mathews, 2007; Rebarber et al., 2007). In addition, some consortia used test development methodologies to facilitate understanding the developmental nature of the language proficiency within the assessments. For example, a validation study of the English Language Development Assessment was based on the latent-class methodology in which items were studied in terms of ELP development (Lara et al., 2007). Similarly, the World-Class Instructional Design and Assessment (WIDA) consortium conducted developmental level ratings of items in which educators experienced in teaching ELLs and/or language testing were asked to assign items with performance-level designations that best identified the language proficiency level necessary to answer each item correctly (Bauman et al., 2007). Bias reviews, an important step in any assessment development, were performed by all the consortia.

Although these efforts have established a solid foundation for the newly developed (post-NCLB) ELP assessments, we believe there are still issues to be resolved in order to bring the new assessments to the level of providing reliable and valid ELP outcome measures. In the next section, we elaborate on some of these issues.

ISSUES TO CONSIDER IN THE NEW GENERATION OF ELP ASSESSMENTS

In this section, we discuss areas that need additional attention in developing valid and reliable ELP assessments. Although addressed somewhat in the literature (Abedi, 2007), we need to examine issues concerning (a) ELP standards, (b) standard-setting for ELP assessments, (c) creation of composite scores, (d) pilot and/or field-test findings, (e) the baseline scores used in reporting progress, and (f) the concept of academic language. In this article, we briefly discuss these issues and provide recommendations based on both our review of the
literature on existing assessments and our knowledge of the newly developed ELP assessments.

Alignment of State ELP Standards and Consortium ELP Assessment

NCLB requires states to develop ELP standards suitable for ELLs. Then, based on these standards, states are to implement a single, reliable, and valid ELP assessment that annually measures listening, speaking, reading, writing, and comprehension. However, in the context of consortia efforts, this requirement raises the question of whose ELP standards and from which of the participating states should the standards be drawn? Many states participating in the four consortia did not have a set of well-defined ELP content standards at the beginning phase of the development process (Fast et al., 2004). Even for consortia where all participating states may have had well-established ELP content standards, the decision as to which standards from which states should be used to develop the common assessment was fraught with difficulty. Is it reasonable to establish a common set of ELP standards across the participating states in the consortium? If states have these standards in common, does each of the standards have the same level of importance for all the participating states? Is it reasonable to expect the same degree of alignment between state standards and a consortium-developed test as with a custom-designed state test? If not, how should the alignment of state standards with ELP assessments be addressed?

Standard-Setting Challenges for New ELP Tests

Literature reports major inconsistencies between the achievement-level classifications produced by different standard-setting techniques (e.g., Impara & Plake, 1997; Jaeger, 1989; Kiplinger, 1996; Loomis, 2001; Musick, 2000; Texas Education Agency, 2002). For example, the use of the bookmark procedure (Lewis, Mitzel, & Green, 1996) may result in a different cut score for classifying students as “intermediate” in proficiency than would the results of the Angoff method (Angoff, 1971) for the same test; when tests differ in terms of specific standards to which they are aligned, test formats, and content, there can be no expectation that their results will yield consistent classifications. Thus the standards set by each method may be quite different, which may lead to different

---

1ELP standards, English language development standards, and English as a second language standards are terms that have been used, often interchangeably, to describe state adopted standards that guide the instruction and assessment of English learners toward the achievement of English language proficiency. The NCELA glossary (2007) indicates that ELP is “often used in conjunction with NCLB and Title III guidelines.” We therefore use the ELP term in this article.
interpretations of the students’ level of ELP. Even when the same standard-setting technique is applied to the same test, the results may vary greatly depending on factors such as the educational background and training of judges involved in the standard-setting process (e.g., Impara & Plake, 1997; Jaeger, 1989; Kiplinger, 1996).

In addition to the inconsistencies in classification that arise from the use of different standard-setting approaches and different assessments, other factors may introduce bias into the process. For example, in deciding the number of performance levels their assessments should address, the consortia were faced with a dilemma. Fewer cut-points may require fewer items and thus shorter tests, which may be a favorable feature for many states. A greater number of performance levels provides the opportunity for more subtle distinctions between students’ performance but requires a greater number of items and longer tests (Fast et al., 2004). Typically, five performance levels were used in the newly developed ELP assessments. The performance level descriptors were slightly different across the different tests developed by the four consortia, but Level 1 usually refers to no or very low proficiency in English and Level 5 represents high proficiency. Similarly, whereas different states set their criteria for reclassification of ELL students from limited English proficient (LEP) to fluent English proficient (FEP) at different proficiency levels, ELL students are typically reclassified from LEP to FEP at ELP performance Level 4 or above.

There are other issues concerning standard-setting for the ELP assessments as well. Among these issues are inconsistencies between performance levels set for the different domains of reading, writing, speaking, and listening. When performance levels are set separately for each of the domains, then discrepancies between such levels across the domains can make interpretation of the results difficult. For example, many students can be classified as proficient or above in one domain but may be classified as below proficient in other domains. How can such issues be resolved (see, e.g., Bunch, 2006)? Should achievement levels be established for the test as a whole? If so, then how should the total test score be obtained? This raises a whole new set of issues concerning dimensionality of ELP tests with the four different subscales, the topic of our next section.

Creation of Composite Scores: Dimensionality Issue

NCLB Title III requires states to measure the annual growth of students’ English language development in reading, writing, listening, and speaking—and in comprehension. In addition to the scores from each of these four domains, states use composite scores from a number of the domains. Although an overall composite of the four domains is commonly used by states, other composite scores based on only some domains are also used by some members of the
consortia. For example, in the Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs®) assessment, “reports include four weighted composite proficiency scores: an Overall composite score reflecting all domains, an Oral Language composite score (listening and speaking), a Literacy composite score (reading and writing), and a Comprehension composite score (listening and reading)” (Bauman et al., 2007, p. 90). Sometimes these composites are based on unequally weighted subscale scores. For example, the states in the WIDA consortium (ACCESS for ELLs) decided on and computed the overall composite as 15% listening, 15% speaking, 35% reading, and 35% writing and the comprehension composite as 30% listening and 70% reading (Bauman et al., 2007). Similar policies have been adopted by other consortia of states.

However, to create these composite scores, it is important to know how these different subscales are correlated and whether they measure a single construct (i.e., ELP) or whether they measure four different constructs, namely, reading, writing, listening, and speaking. If the subscales are highly correlated, the decision to combine them as well as the weightings would be more understandable than if the subscales are not highly correlated. Therefore, the issue of dimensionality needs to be addressed prior to any decisions about creating composite scores. It also is important to provide evidence to justify the particular weights used to create composite scores. There can be a big difference between decisions based on the views of state policy makers and those based on solid empirical evidence.

To begin with, researchers should ask, “Should the four domains be considered as four separate subscales/dimensions or should they be considered as a single latent trait that encompasses all four domains?” The number of constructs being measured seriously affects reporting and interpretation of scores. If the four domains are measuring a single construct (i.e., an overall ELP latent variable), then scores from the four domains can be combined and a single score can be used for reporting annual measurable achievement objectives and for classification purposes. On the other hand, if each domain has a unique contribution to the ELP construct, then it is difficult to justify aggregating the domain scores. Different models for creating composite scores are suggested in the literature (see, e.g., Abedi, 2004; Sawaki, Stricker, & Oranje, 2007). In a factor analytic model, when a single ELP construct is postulated, the common variance shared across the four domains is used and a single latent variable of ELP is computed. This requires high or near perfect correlations between scores of the four domains. If individual domains contain large specific variance in addition to the overall ELP variance, then among the two most commonly used options for aggregating domain scores (compensatory vs. conjunctive models), which one would be preferable? In the compensatory model,
a low score in one domain may be compensated by high scores on another domain. As Abedi (2004) elaborated, the preferred model for NCLB is the conjunctive model in which students should score at the proficient level in each of the four domains to pass annual measurable achievement objectives requirements. If this model is used, then creating aggregated scores becomes even more complex.

Available Indicators for Establishing Concurrent Validity

A major strength of the ELP assessments recently developed by the consortia relative to the prior generation of ELP assessments is the number of psychometric and validation studies incorporated into the development process. Many different approaches in validation of the ELP assessments were utilized. These approaches include latent-class analyses, a criterion-related validity approach using both concurrent and predictive approaches, as well as a Multi-Trait/Multi-Method approach within a structural equation modeling framework. The validation phase also included content validation through alignment to ELP content standards and construct validations using the confirmatory factor analytic approach (e.g., multiple group confirmatory factor analyses).

A major issue in estimating the validity of ELP assessments through the criterion-related or construct approach is the validity of the criteria used for validation of ELP measures. As indicated earlier, reviewers of the pre-NCLB assessments expressed concern over the soundness and validity (Del Vecchio & Guerrero, 1995; Zehler et al., 1994) on some of the ELP tests. Thus, although these tests are nearly the only readily available criterion for examining concurrent validity, they have many flaws for this purpose. A low correlation between the newly developed ELP assessments and the older assessments might be due to psychometric problems in the older assessments and/or the differences between the content and structure of the new and old generations of ELP assessments.

Bauman et al. (2007) found correlations between the four domains of a new ELP assessment (listening, speaking, reading, and writing) and four existing ELP assessments (IDEA Proficiency Test [IPT], Language Assessment Scales [LAS], Language Proficiency Test Series [LPTS], & Maculaitis Assessment of Competencies [MAC] II) to range between 0.468 and 0.765 with an average of 0.604. Although this correlation is considered relatively high, it explains only 36% of the variance between the ACCESS for ELLs test and the existing ELP tests used as the criterion variables. Once again, there are many factors that could explain the lack of a strong correlation between the newly developed ELP assessments and the existing ELP assessments. Among the most important sources contributing to a low correlation is the low content and psychometric comparability between the two sets of assessments.
Baseline Scores for the NCLB Title III Assessment Reporting

As the implementation phase of NCLB Title III began around 2002, efforts to develop new ELP assessments based on the NCLB requirements began as well. It took more than 3 years for most of the consortia’s assessments to become fully developed and field tested. In 2002, there were many existing ELP assessments on the market—a majority of which, however, did not meet the NCLB Title III assessment requirements. Because the newly developed ELP assessments were not available at the start of NCLB implementation, states had no other choice but to use whatever existing ELP assessment they found relevant. This situation obviously introduced flaws into the reporting of ELP progress, one of which was that subsequent tests might not be comparable with the tests they replaced (George, Linquanti, & Mayer, 2004; Gottlieb & Boals, 2006; Linquanti, 2004).

Now that many states have access to the newly developed ELL assessments that meet the NCLB requirements, they are faced with the quandary of linking baseline results based on then available off-the-shelf ELP assessment tests with the results from their new ELP assessments. Incomparability of content is a first issue, in that many of the baseline assessments did not include all four domains of ELP (reading, writing, speaking, and listening) and baseline and current assessments reflect different theoretical emphases in the definition of measured constructs. In addition, baseline tests are not usually aligned with state’s ELP content standards and do not reflect the importance of academic language development. Therefore, even a high statistical correlation between ELP assessments used as the baseline and the new ELP assessment would not be enough to establish a strong link between the two assessments.

The WIDA consortium conducted a study that may shed light on the use of pre-NCLB assessment measures as the baseline. This bridge study was conducted with a sample of 4,985 students enrolled in Grades K through 12 from selected districts in Illinois and Rhode Island. Students in this study took both the ACCESS for ELLs and one of four older ELP tests: IPT, LAS, LPTS, and MAC II. The purpose of this study was to predict performances on ACCESS for ELLs from performances on the older tests using a linear regression procedure.

One can argue that if ELP assessments claim to measure students’ levels of English proficiency, there must be high correlations between those assessments—whether they are pre- or post-NCLB. However, as Bauman et al. (2007) elaborated, one may not expect high-level relationships between the pre- and post-NCLB assessments because they are very different in many different aspects, including content, construct, and psychometric characteristics. The findings of the bridge study by Bauman et al. suggest a moderate level of relationship between the four domains of ELP assessment (listening, speaking, reading, and writing) and four existing ELP assessments (IPT, LAS, LPTS, MAC II).
The Concept of Academic Language

Among the NCLB instructions provided to states for developing reliable and valid ELP assessments is the requirement to incorporate the concept of academic English into the process. The major goals of NCLB Title III are

to help ensure that limited English proficient (LEP) children attain English proficiency, develop high levels of academic competence in English, and meet the same challenging state academic content and student academic achievement standards that all children are expected to meet. (Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students, 2003, p. 5)

Similarly, in response to the question, “B-5. Why must English language proficiency standards be linked to academic standards?” the U.S. Department of Education indicated that:

The statute requires English language proficiency standards to be linked to state academic content and achievement standards in reading or language arts and in mathematics beginning in the school year 2002–2003. This is required in order to ensure that LEP students can attain proficiency in both English language and in reading/language arts, math and science. English language proficiency standards should also be linked to the state academic standards in science beginning in the school year 2005–2006. (Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students, 2003, p. 10)

Clearly, the focus of the ELP assessment mandate is academic English. Many of the newly developed measures of ELP are therefore based on the need to test academic English, which in turn facilitates learning content knowledge across the following three academic topic areas: English/Language Arts; Math, Science and Technology; and Social Studies, as well as one nonacademic topic area related to the school environment (Fast et al., 2004).

However, there have been controversies over what is covered under the concept of academic English. Should ELP assessments include the language of the aforementioned content areas or include the language that facilitates learning that content? Fast et al. (2004) clarifies this issue by indicating that ELP assessments “are not tests of academic content, in other words, no external or prior content-related knowledge is required to respond to test questions” (p. 2) and that the concept of academic English refers to the language that facilitates learning of the academic content not the mastery of academic concepts (Fast et al., 2004). That is, eventually, ELL students should be able to demonstrate proficiency in the academic English that is required to successfully function within their grade level. We believe this is an area that needs attention from experts in a variety of
disciplines. Experts in the field of linguistics with knowledge and experience in academic language, along with content and measurement experts, should join in the efforts to operationally define academic language and provide guidelines for test item writers who are assigned to ELP test development. It is also important to include teachers, bilingual coordinators, and state personnel working with ELP assessment experts. Meanwhile, it might benefit states to review their current ELP tests and evaluate the test items in terms of academic English content.

ELP Assessments in Determining ELL Students’ Readiness for Title I Assessment

Results from ELP assessment can inform decisions on when ELL students are ready to participate in the NCLB Title I assessments in English. State assessments that are used for the NCLB Title I accountability purposes are mainly constructed and field tested for students who are fluent in English. Therefore, there might be linguistic factors that could seriously undermine the validity of content-based assessments for ELLs. Hence, it is imperative to determine at what level of proficiency in English ELL students can meaningfully participate in the state content-based assessments written in English.

We used data from three different sites: Site 1 provided data from pre-NCLB assessment, and Site 2 and Site 3 provided post-NCLB data. The pre-NCLB site (Site 1) was a district with a large number of ELL students. Data from a group of 916 ELL/non-ELL third-grade students were used (Staley, 2005). For each student the data included six measures of ELP, three standardized achievement English language measures, and six standardized achievement mathematics measures. Of the 916 students, 602 had complete data on all the measures used in the analyses. The post-NCLB data were obtained from two locations which we refer to as Site 2 and Site 3. Site 2 (post-NCLB) is a state with a large number of ELL students. This site provided data from the entire state for Grade 4 students. Similarly, Site 3 (post-NCLB) provided data from an entire state for Grade 4 students. Data included reading scores for ELL students in Grade 4 and their reading scores in Grade 3. In addition, the data set included ELP scores in reading, writing, listening, and speaking. The data from the three sites were analyzed using a canonical correlation approach.

It must be indicated at this point that the structure of the data from the pre-NCLB and the post-NCLB sites were quite different. Different ELP tests based on different theoretical foundations with different content and psychometric characteristics were used at the pre- and post-NCLB sites. Such differences may limit our ability to draw valid comparisons between the pre- and post-NCLB data. However, we considered ELP as a construct. This construct was defined based on a more limited coverage of ELP in the pre-NCLB era and was defined in a more comprehensive manner in the post-NCLB era. Therefore, we
are comparing two constructs, the construct of reading/language arts and the construct of English proficiency in the pre- and post-NCLB sites rather than specific sets of measures. We understand that the variables used for creating these constructs may not be highly correlated; therefore, dimensionality might be an issue here.

A canonical correlation approach was deemed relevant for this case because (a) we were interested in predicting ELL student performance in content-based assessments (Title I) from her or his performance in ELP assessment (Tile III), and (b) there were multiple dependent and multiple independent variables.

For the pre-NCLB canonical model, reading language arts scores were used as the first set of variables and scores of LAS Reading and LAS Writing were used as the second set. Table 1 summarizes the results of canonical correlation for this site. As data in Table 1 show, the canonical correlation model yielded two canonical variates. The canonical correlation between the reading language art scores (in the first variable set) and ELP scores (second variable set) for the first canonical variate was 0.74, explaining about 55% of the variance in the canonical variate, and for the second variate the canonical correlation was 0.09 explaining less than 1% of the variance of the canonical variate. Because the second canonical variate explained a negligible amount of variance of the canonical variates (less than 1% of the variance), we decided to drop this canonical variate and focus on the first variate.

The first pair of variables in this canonical variate is reading score for the first set (with a canonical correlation of 0.99) and Reading subscale of the English proficiency, LAS from the second set (0.96). The second pair of variables is language score (0.89) with LAS writing (0.83). These results clearly suggest that the LAS Reading subscale score (Tile III) has the power of predicting

<table>
<thead>
<tr>
<th>Assessment</th>
<th>First Canonical Variate</th>
<th>Second Canonical Variate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Reading &amp; Language Arts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>.99</td>
<td>.77</td>
</tr>
<tr>
<td>Language</td>
<td>.89</td>
<td>.27</td>
</tr>
<tr>
<td>English proficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAS Reading</td>
<td>.96</td>
<td>.74</td>
</tr>
<tr>
<td>LAS Writing</td>
<td>.83</td>
<td>.36</td>
</tr>
<tr>
<td>Canonical correlation</td>
<td>.74</td>
<td>.09</td>
</tr>
</tbody>
</table>

Note. NCLB = No Child Left Behind Act; LAS = Language Assessment Scales.
scores on reading language arts (Title I) at the medium/high level of prediction (with a canonical correlation of 0.76).

Table 2 presents a summary of canonical analyses for the post-NCLB data obtained from Site 2. In this post-NCLB model, data from 6,749 ELL students in Grade 4 were analyzed. The first set of variables included students’ reading scores in Grade 4 as well as their Grade 3 reading scores. The second set of variables included scores from the four ELP domains (reading, writing, listening, and speaking). As the data in Table 2 suggest, this model yielded two canonical variates. The canonical correlation for the first variate was 0.83 and for the second variate the canonical correlation was 0.06. Similar to the pre-NCLB model in Site 1, the second canonical variate did not explain much of the variance of the two sets of variables (less than 1% of the variance of the variables was explained by the second canonical variate); therefore, the second variate is not discussed further.

As the data in Table 2 show, the variable with the highest correlations, with the first canonical variate in the first set of variables, is students’ reading scores in Grade 4 (0.97), and in the second set of data the variable with the highest correlation is scores of reading subscale (0.95). Therefore, these two variables create a pair. The second pair is students’ reading scores in Grade 3 (0.87) with their scores in the writing subscale of the language proficiency test (0.87). These results suggest a strong relationship between reading/writing content areas and students ELP scores.

Table 3 presents the summary results of the analyses for the post-NCLB Site 3. The variables included in the canonical correlation model for this site are very similar to those that were included in the model for Site 2 post NCLB. The first

<table>
<thead>
<tr>
<th>TABLE 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical Correlation Summary Results, Post-NCLB, for Site 2</td>
</tr>
<tr>
<td>Assessment</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Reading &amp; Language Arts</td>
</tr>
<tr>
<td>Read4 06 Tot Scal</td>
</tr>
<tr>
<td>Read3 05 RSS05</td>
</tr>
<tr>
<td>English proficiency</td>
</tr>
<tr>
<td>Speak</td>
</tr>
<tr>
<td>Listen</td>
</tr>
<tr>
<td>Read</td>
</tr>
<tr>
<td>Write</td>
</tr>
<tr>
<td>Canonical correlation</td>
</tr>
</tbody>
</table>

*Note. NCLB = No Child Left Behind Act.*
### TABLE 3
Canonical Correlation Summary Results, Post-NCLB, for Site 3

<table>
<thead>
<tr>
<th>Assessment</th>
<th>First Canonical Variate</th>
<th></th>
<th>Second Canonical Variate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation</td>
<td>Coefficient</td>
<td>Correlation</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Reading &amp; Language Arts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read4 06 ScalScor</td>
<td>.95</td>
<td>.63</td>
<td>-.32</td>
<td>-1.24</td>
</tr>
<tr>
<td>Read3 05 ScalScor</td>
<td>.89</td>
<td>.45</td>
<td>.46</td>
<td>1.32</td>
</tr>
<tr>
<td>English proficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speak</td>
<td>.46</td>
<td>.11</td>
<td>.88</td>
<td>1.05</td>
</tr>
<tr>
<td>Listen</td>
<td>.65</td>
<td>.23</td>
<td>.10</td>
<td>.03</td>
</tr>
<tr>
<td>Read</td>
<td>.90</td>
<td>.55</td>
<td>-.13</td>
<td>-.37</td>
</tr>
<tr>
<td>Write</td>
<td>.80</td>
<td>.38</td>
<td>-.11</td>
<td>-.21</td>
</tr>
</tbody>
</table>

Note. NCLB = No Child Left Behind Act.

set of variables include students’ Grades 3 and 4 reading scores and the second set of variables include scores of the four domains of ELP (reading, writing, speaking, and listening). Similar to the model for Site 2, this model yielded two canonical variates. The canonical correlation for the first variate was 0.77 explaining about 60% of the variance of the variables in this model, and the canonical correlation for the second variate was 0.04 explaining less than half a percent of the variance of the model. Therefore, it was decided to drop this variate from the analyses.

In the first canonical variate, the variable with the highest correlation in the first set was Grade 4 reading (0.95) and the variable with the highest correlations in the second set of variables was reading (0.90). Therefore, the reading language arts and reading domain of the ELP assessment create the first pair. The next variable with the highest correlation in the first set is student’s Grade 3 reading (0.89), which pairs with writing domain in the second set (0.80).

These results are almost identical with the results from Site 2 which provides high level of cross validation evidence for this study. Also, results of canonical correlation analyses from the pre- and post-NCLB are very similar in many aspects. In both cases, the first canonical variate explained most of the shared variance of the two sets. The size of correlation between the variables in both sets and the first canonical variate was quite similar across the two sites. The only major difference between the results from pre- and post-NCLB was that there was a slightly higher level of association between the two sets of variables in the post-NCLB (0.83 for Site 2 and 0.77 for Site 3) as compared with the pre-NCLB (0.74 for Site 1). One possible explanation for this slightly higher relationship between Title I and Title III assessments in the post-NCLB may
be the improvement of the ELP assessments. Another possible explanation for this difference is the fact that the data are from two different sites with different measures and different variables.

To present a clearer picture of the relationship between ELP and ELL student performance in reading/language arts, we compared ELL students’ level of English proficiency with their performance in reading/language arts assessments. Data for these analyses came from the 2005–2006 academic year from two states with large numbers of ELL students. Because of confidentiality considerations, we call these two sites Site X and Site Y. Table 4 presents results of analyses for Site X and Table 5 presents the results for Site Y.

As data in both Tables 4 and 5 suggest, there is a strong relationship between students’ level of English proficiency as judged by ELP assessments and their performance in reading/language arts. For example, the data in Site X show that

<table>
<thead>
<tr>
<th>Language Proficiency</th>
<th>Frequency</th>
<th>N</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEP</td>
<td>1,238</td>
<td>859</td>
<td>475.59</td>
<td>90.46</td>
</tr>
<tr>
<td>LEP</td>
<td>4,868</td>
<td>4,788</td>
<td>531.93</td>
<td>61.49</td>
</tr>
<tr>
<td>FEP</td>
<td>2,222</td>
<td>2,216</td>
<td>588.88</td>
<td>41.36</td>
</tr>
<tr>
<td>Non-ELL</td>
<td>47,392</td>
<td>47,151</td>
<td>597.21</td>
<td>56.69</td>
</tr>
<tr>
<td>Missing</td>
<td>3,635</td>
<td>3</td>
<td>588.67</td>
<td>91.31</td>
</tr>
<tr>
<td>Total</td>
<td>59,355</td>
<td>55,017</td>
<td>589.29</td>
<td>61.82</td>
</tr>
</tbody>
</table>

*Note.* ELL = English language learner; NEP = non-English proficient; LEP = limited English proficient; FEP = fluent English proficient.

<table>
<thead>
<tr>
<th>EOG_LEPSTATUS</th>
<th>N</th>
<th>No. of Participants</th>
<th>M Reading</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELL (new arrivals)</td>
<td>440</td>
<td>325</td>
<td>237.77</td>
<td>5.45</td>
</tr>
<tr>
<td>ELL</td>
<td>4,935</td>
<td>4,825</td>
<td>246.16</td>
<td>7.65</td>
</tr>
<tr>
<td>Reclassified ELL</td>
<td>1,641</td>
<td>1,633</td>
<td>256.61</td>
<td>6.50</td>
</tr>
<tr>
<td>Non-ELL</td>
<td>95,084</td>
<td>94,209</td>
<td>253.42</td>
<td>8.59</td>
</tr>
<tr>
<td>Total</td>
<td>100,992</td>
<td>100,042</td>
<td>253.08</td>
<td>8.66</td>
</tr>
</tbody>
</table>

*Note.* ELL = English language learner.
ELL students at the lower level of English proficiency (non-English proficient [NEP]) have the lowest reading mean score ($M = 475.59$, $SD = 90.60$) as compared with the native English speakers ($M = 597.21$, $SD = 56.92$), a difference of more than 120 score points (about 2 $SD$s). However, the data show that the ELL group at the lowest level of English proficiency is not as homogeneous in their test scores as other groups with higher levels of English proficiency. For example, the standard deviation for NEP students is 90.46 as compared with the highest standard deviation of 61.49 for the other three groups based on their language proficiency.

It is also clear from the data in both tables that ELL students at the lower level of English proficiency have lower rates of participation in reading/language arts assessments. The number of ELL students who are participating in reading/language arts assessments is smaller than the number for non-ELL students. For example, of 1,238 NEP students, 859 or 69.4% of them participated in reading/language arts assessments, whereas of the 47,392 non-ELL, more than 99% (47,151) participated in the assessment. These data suggest that ELL students at the lower level of English proficiency may not be ready to meaningfully participate in content-based assessments.

Although, because of differences in the scaled scores, the results from the two sites may not be directly comparable, trends of performance differences at categories of students’ level of proficiency are quite similar across the two sites (Sites X and Y). As data in Table 5 show, ELL students at the lower level of English proficiency (newcomers) have a mean reading score of 237.77 ($SD = 5.45$) as compared with the mean score of 253.42 ($SD = 8.59$) for non-ELLs, with a difference of more than 15 score points (about 2 $SD$s) between the two groups.

These analyses clearly suggest that ELP assessments can be used as valid criteria for decisions regarding participation of ELL students in Title I assessments. This finding is an extremely important milestone in ELL students’ academic career. If these students proceed with taking Title I assessments without having the sufficient linguistic knowledge to understand and respond to assessment questions, then the results of Title I assessments for these students may be questionable.

**SUMMARY AND DISCUSSION**

Measures of English language proficiency are of immense value to ELL students as they have the greatest impact on their academic life. Such measures inform decisions on classification, curriculum planning, and participation in content-based assessments in English. ELP measures with unknown quality may cause grave academic consequences for ELL students. Invalid measures
MEASURING STUDENTS’ ENGLISH PROFICIENCY LEVEL

of ELP may result in misclassified and inappropriate instruction. Reviews of ELP assessments indicated that most of the ELP tests that existed prior to the implementation of NCLB had poor quality, both in terms of content and psychometric properties. They measured different content and were based on different theoretical emphases prevalent at the time of test development; they also were not necessarily connected to the state’s ELP content standards and did not address the importance of academic language development. In fact, many states did not even have a set of clearly defined ELP content standards.

The NCLB legislation made the assessment of ELL students’ level of proficiency in English mandatory once a year and provided useful guidelines for developing ELP assessments. This mandate along with its useful guidelines helped improve the quality of ELP assessment significantly and led to the development of several batteries of ELP assessments, either through consortia of states or by test publishers. These improvements were reflected in several areas: (a) The newly developed assessments were based on states’ ELP standards; (b) they included a more comprehensive set of measures of different aspects of English proficiency (i.e., reading, writing, listening, and speaking); (c) they incorporated the concept of academic language, which is an essential requirement for ELL students’ performance in the academic content areas; and (d) they were tested in extensive pilot and field studies. However, as elaborated earlier in this article, there are still issues with the newly developed ELP tests to be resolved.

Among the most important functions of ELP assessment is to inform decisions regarding participation of ELL students in the content-based assessments such as reading/language arts, math, and science (NCLB Title I assessment). Literature has clearly demonstrated that participation of ELL students in content-based assessment in English when they are not able to understand the assessment questions may not be productive. For example, literature suggests that ELL students do poorly in content-based assessments in English as compared with non-ELL students, but the performance gap between ELL and non-ELL students increases as the level of language demand of assessment increases (see Abedi, 2006a; Abedi & Gandara, 2006; Maihoff, 2002; Solano-Flores & Trumbull, 2003). Unfortunately, however, issues concerning the impact of language factors are not considered in the development process of many standardized achievement tests. Therefore, many of these assessments at the state levels may not present a comprehensive picture of what ELL students know and are able to do. It is therefore imperative for ELL students to participate in the state content-based assessments when their ELP level matches that of the content-based assessment language.

However, if ELP assessments ought to play such an important role in the assessment and accountability system for ELL students, they have to be sound in terms of content and psychometric characteristics. A comparison between the content and psychometric properties of the pre- and post-NCLB measures of
ELP indicated that the ELP assessments that were based on the NCLB guidelines present a better quality measure when compared with the ELP assessments prior to the implementation of NCLB. However, there is not enough research evidence to fully support this notion. The main question here is whether improvements in the quality of the newly developed ELP assessments help to make a more informed decision about participation of ELL students in the state content-based assessments. To answer this question, we compared the power of ELP assessments that existed prior to NCLB with those developed along the guidelines provided by NCLB in predicting ELL performance in content-based assessment.

We used ELP measures to predict ELL students’ performance in reading/language arts using a canonical correlation model. The results from the canonical correlation analyses indicated that ELP measures, particularly reading scores, were strongly associated with students’ performance in reading/language arts in both pre- and post-NCLB assessments, with a slight improvement in the post-NCLB assessments.

This finding has quite important implications for ELL assessment and accountability. Participation of ELL students in the state content-based assessment in English without enough proficiency would produce undesirable results and would seriously impact their academic career. Therefore, the decision must be based on ELP outcome measures. However, before using ELP outcome measures for such an important decision, one must make sure that these assessments are capable of being used for such purposes.

**RECOMMENDATIONS**

Next is a set of recommendations based on our knowledge of the field and the information that was presented here. Once again, we hope our presentations along with our recommendations will start a national dialog for improving the quality of ELP assessments that are essential for better instruction, assessment, and accountability systems for ELL students.

- Use multiple methods for setting standards and defining cut scores for achievement levels.
- Examine the comparability of the assessment used to establish the baseline with the newly adopted ELP assessment. If there is not a strong link between the two assessments, both in terms of content and psychometric characteristics, then use caution in making firm judgments about the growth of students’ English proficiency.
- Examine the content of your state-adopted ELP assessment and align the content with the state ELP content standards. This is important, as the ELP
MEASURING STUDENTS’ ENGLISH PROFICIENCY LEVEL

consortia’s assessments are not completely based on the ELP standards of any one state.

- Examine the pattern of possible differential performance of ELL students on the ELP assessments to make sure that the ELP assessment items do not differentially or unfairly perform across the subgroups within the ELL population.
- Use multiple criteria for assessing ELL students’ level of English proficiency, particularly with high-stakes decisions such as classification or reclassification of students.
- Use ELP assessment results along with other sources to make informed decisions about ELL student participation in Title I assessment, as the literature clearly suggests that assessments that are constructed for native speakers of English may not provide valid outcomes to ELL students at the lower levels of English proficiency.
- Train staff with high levels of knowledge and experience in measurement so that you can constantly review and monitor assessment issues, particularly in the area of English proficiency. As a part of their contract, test publishers may provide states with the needed technical information. In addition, states should have an independent evaluation capacity to examine the quality of state assessments.
- Incorporate a major measurement research component into your programs that can be supervised and run with your professionally trained staff. Once again, states must always reserve the right to examine the validity of their assessments and conduct analyses independent of what the test publishers/developers provide, to bring another layer of confidence into their high-stakes assessments.

Furthermore, it is imperative that states actively pursue research and development in maintaining the quality of ELP assessments. Conducting validity studies and examining test items for any indication of threats to their authenticity over time will help assure the quality assessment of students’ level of English proficiency. To reach this important goal, the following recommendations are provided (S. Ferrara, personal communication, September 2007):

- States’ ELP assessments are ongoing operational assessment programs, just like grade-level content area assessment programs. States should manage their ELP assessments accordingly.
- They should implement field testing procedures and replenish their item banks and operational test forms on a regular basis so that they do not overexpose current items, tasks, and test forms.
They should conduct ongoing reviews of the alignment of items and assessment tasks with ELP standards, and the psychometric characteristics of the item banks and test forms.

States should plan and implement validity studies on an ongoing basis to examine current issues in assessing ELLs and ELP assessments that were discussed in this article.

Our main objective in this article was to present information on the status of assessment of ELP in the nation. We acknowledge our limitations in both content and scope of this overview. This is an extremely complex area in assessment and deserves more attention. We hope this presentation opens dialog on the importance and quality of ELP assessments in the nation, as there is much to learn from our past experience.

REFERENCES


Texas Education Agency. (2002). *State accountability data tables, base indicators.* Texas Department of Education.

Copyright of Educational Assessment is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.
Copyright of Educational Assessment is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.