Assessment of English Language Learners Under Title I and Title III: How One Testing Program Can Inform the Other

Stanley Rabinowitz

Assessment and Accountability Content Center

WestEd

Full inclusion of English Language Learners (ELLs) in assessment and accountability systems represents a landmark shift in U.S. federal education policy. Under the No Child Left Behind Act (NCLB), states are required to not only assess ELLs, but to demonstrate, with accountability consequences, that ELLs are becoming English proficient and achieving proficiency in the same core academic content as all other students. Both Titles I and III of NCLB include assessment and accountability requirements for ELLs. While the main focus under Title III is assessment of and accountability for ELLs' English language proficiency, under Title I it is on proficiency for all students—including ELLs—in reading/language arts, mathematics, and science.

In principle, holding schools, districts, and states accountable for the achievement of ELLs is widely, though not universally, perceived as a positive development. What better way to increase awareness of the academic needs and progress of ELLs? In practice, however, states are grappling with just how to meet NCLB's requirements for ELLs. In

large part, this is because state-of-the-art of assessment practices and accountability policies for ELLs lag behind those for non-ELLs.

In a recent report, the Government Accountability Office (GAO) (March 2007) noted that about one-third of the 33 states they contacted said that they want more guidance from the U. S. Department of Education (ED) concerning ELL assessment. Many states have expressed similar wishes directly to ED. In response to this need for additional guidance, ED will provide support in a variety of ways. One way is through written guidance and frameworks that detail research-based practices. Another is through publishing papers related to specific ELL assessment issues, such as linguistic accommodations and native language assessments to ensure valid measurement of ELLs' proficiency in academic content and different approaches states might take to improve the assessment of ELLs apart from statutory and regulatory requirements. This paper is an example of the latter. Unlike the other papers that focus on a specific assessment issue, this paper addresses a broad assessment strategy. It explores and explicates a potential relationship between Title I and Title III assessments for ELLs.

The inclusion of ELLs under NCLB in large-scale assessment and accountability systems presents unique challenges to states, the most important being the need to demonstrate the validity of all assessments and the interpretability and usability of resultant test scores. Valid assessments for ELLs result from the "minimalization or removal of sources of construct-irrelevant variance in order to facilitate students' ability to demonstrate their construct-relevant knowledge and skills" (Sato, 2007). While strategies to remove

sources of construct-irrelevant variance may differ somewhat across various at-risk student populations (ELLs, students with disabilities, low performing general education students), refinement of standard test development and validation procedures can create efficiencies that will support the assessment goals and requirements of both Titles I and III.

Several states have explored the possibility of developing one test for both Title I and III purposes. Such attempts have been unsuccessful to date, given the differences between the two titles in their goals and the constructs to be assessed (i.e., language vs. content proficiency). Adding to the difficulty are the technical and logistical challenges related to developing and administering assessments for diverse student populations and multiple purposes. Despite these challenges, there are several ways that Title I and Title III assessment programs could be made more coherent and provide information that is valuable across testing programs and for a state's overall accountability system.

Specifically, this paper explores ways in which Title I and Title III assessment programs for ELLs can inform one another so as to avoid redundant or incoherent testing of ELLs. The following questions will be addressed:

- What are the similarities and differences in Title I and Title III assessment requirements, practices, and technical considerations for ELLs?
- How can information be shared across Title I and Title III assessment programs in order to maximize data utility and create efficiencies?

- What specialized validity studies should states plan and implement to support inferences across Title I and Title III assessments?

Title I and Title III Assessment Requirements

The major assessment-related provisions for Titles I and III are briefly summarized in Table 1. As shown, Title I deals mainly with requirements around state academic assessment (in reading, mathematics, and science). Its intent is to ensure that all students, including particular subgroups of students, are included in the annual state academic assessments. As such, it requires that students with limited English proficiency[1] take part in state academic assessments. To allow students to demonstrate their academic knowledge, states can offer assessments in a student's native language or offer accommodations, such as allowing use of a bilingual dictionary or providing additional time to take a test. While all students with limited English proficiency must be tested after their first year of attendance in U.S. schools, their assessment scores are included as part of the accountability determinations only after they have attended U.S. schools for three years. Title I also requires that states annually assess the English language proficiency of all students with limited English proficiency, measuring students' oral language, reading, and writing skills in English.

Whereas Title I deals with all students and specific subgroups of students, Title III focuses specifically on students with limited English proficiency. It requires states to

[1] Consistent with the language of NCLB, this paper will refer at times to students with "limited English proficiency" (LEPs). We recognize that many researchers and practitioners prefer the term English-Language Learners (ELLs) or English Learners (ELs). We will use these terms interchangeably.

establish English language proficiency (ELP) standards and track the progress of students with limited English proficiency in order to help ensure that these students make progress in learning English and attain English proficiency.

Table 1: Selected Requirements for ELLs from Title 1 and Title III[2]

Selected Requirements from Title I

Academic Assessment Requirements	States must provide for participation of all students, including those with limited English proficiency in their academic assessment program. Assessments are given in reading and mathematics in grades 3–8 and at least once in high school. Science assessments are given once in each of elementary, middle, and high school (starting in 2007–08). Assessments must be aligned with challenging state academic standards. Students with limited English proficiency must be assessed in a valid and reliable manner, and provided with reasonable accommodations. To the extent practicable, they should be assessed "in the language and form most likely to yield accurate data" on their academic knowledge. Students who have been in U.S. schools for 3 years or more generally must be assessed in English.
Academic Proficiency Accountability Requirements	States must set annual goals that lead to all students achieving proficiency in reading and mathematics by 2014. To be deemed as having made adequate yearly progress (AYP) for a given year, each district and school must show that the requisite percentage of each designated student group, as well as the student population as a whole, met the state proficiency goal. Students with limited English proficiency must be assessed, starting the first school year after they have arrived in the U.S., but are not included for determining AYP until they have been in U.S. schools for 3 years or have reached English proficiency (whichever is sooner).
English Language Proficiency Assessment Requirements	States must annually assess the English language proficiency of all students with limited English proficiency, measuring students' oral language, reading, and writing skills in English.

[2] This summary is adapted from a recent report of the GAO (July 2006) on NCLB.

Selected Requirements from Title III

English Language Proficiency Standards	States must establish English language proficiency standards that are aligned with the state's challenging academic content standards.
Tracking Student Progress in Learning English	States must establish objectives for improving students' English proficiency in speaking, listening, reading, and writing. States receiving grants under Title III must establish annual goals for increasing and measuring the progress of students with limited English proficiency in (1) learning English, (2) attaining English proficiency, and (3) meeting adequate yearly progress goals in attaining academic proficiency outlined in Title I.

An Expanded View of Assessment Development Practices in Support of Coherent, Efficient Assessment of ELL Students

Development or selection of large scale assessments—whether for academic content or English language proficiency—entails considerations that are universal for assessments for all populations, and others that are more relevant for the ELL population particularly when testing for high stakes purposes, such as school and district accountability under NCLB. Thus, it is important to identify for each major assessment development step, the specific actions that must be taken to ensure valid and fair testing for all participating students, *including ELLs*. These major development steps include the following:

1. Determine the purpose of the test
2. Identify and prioritize the standards on which to base test content
3. Develop test specifications
4. Draft items consistent with general item-writing guidelines and the particular test specifications
5. Conduct content and bias reviews of items

6. Pilot test items on small groups of students and revise items using information from pilot test results
7. Field test items with a representative sample of students
8. Conduct analyses on the field test data
9. Assemble items into operational test forms that are consistent with the test specifications
10. Conduct technical analyses (reliability, equating)
11. Conduct validity studies

Below, we discuss these major steps in the context of Title I and Title III testing.[3] In so doing we pinpoint where there is potential for one testing program to inform, support, or enhance the other, thereby promoting efficiency and coherence across testing programs.

Ideally, the steps described below might most efficiently be implemented at the beginning of the development of a new or revised state testing program. We realize, however, that all states already have Title I and Title III programs in place, with varying degrees of technical evidence and buy-in among state constituencies. States may use the procedures described below to make incremental improvements to their current testing programs. They may also consider fuller implementation of the recommended steps below as part of the periodic update and redesign of their testing programs.

[3] While the focus of this paper is on ELLs, many of the same considerations apply to other subpopulations such as students with disabilities.

1. Determine the purpose of the test. As previously indicated, the primary purpose of Title I academic testing of ELLs is to ensure that ELLs make adequate progress towards academic proficiency, while the primary purpose of Title III English proficiency assessment is to ensure that they are making progress in reaching mastery in English academic language required for success in school. That said, both Title I and Title III assessments of ELLs can serve other purposes, consistent with their primary mission. For example, the findings of both Title I and Title III testing can inform professional development efforts and areas for improvement in teaching practices at the classroom, school, and state levels. Teachers need to learn how to serve the instructional needs of their ELL students, increasingly in English-only, academic classes. District and state officials need to determine the effect of their assessment and accountability policies and provide appropriate support at the classroom level. Together, results from both assessment programs can paint a complete picture of each child's needs and can help educators at the state and local levels revise policy and practice consistent with student achievement data.

2. Identify and prioritize the standards on which to base test content. State standards should drive the content for both Title I and Title III testing: academic standards for the former and ELP standards for the latter. Each set of standards should be developed or revised with the other in mind—mastery of ELP standards should ensure that the student is able to participate fully in the instructional experience necessary to achieve proficiency in challenging content across all subject areas. Content standards should be developed to signal the language requirements needed to access the content itself.

To the extent that there is overlap in the constructs embodied in academic content standards and ELP standards, there may be overlap in the content of Title I and Title III testing. Therefore, states should examine the standards identified for Title I and Title III instruction and assessment for overlap in order to coordinate and reinforce coverage of common content and skills across the two testing programs.

Clearly, the content area with the most potential for content overlap with English Proficiency testing is reading/English language arts (ELA). States should examine their ELP standards to ensure that they fully support achievement of reading/ELA and vice versa. They should ensure that (1) ELP standards do not simply duplicate their ELA counterparts, especially at the earlier grade levels (K–3) and (2) mastering the ELP standards positions students to be successful readers across different genres . Examination of reading/ELA standards alongside ELP standards may lead some states to consider expanding their reading/ELA standards to include writing, speaking, and listening, resulting in a more comprehensive construct of reading/ELA proficiency.

Similarly, ELP standards should be examined to ensure that they facilitate and reinforce achievement of other academic content standards (i.e., math, science). Specifically, ELP standards could cover the academic language necessary to support or scaffold student learning and achievement of math and science standards. Content assessments typically include knowledge of key vocabulary/terminology germane to the domain.

3. Develop test specifications. The next major step in the development process is to produce test specifications and blueprints. Test specifications typically stipulate the breadth and depth of assessed content, provide more specific information about the prioritization of the targeted content and intended difficulty of the items, dictate the format of the items and response modes, and delineate administration and scoring procedures. Test blueprints indicate the relative weighting of each content strand and item type for a specific test form.

In order to ensure accessibility of test items to ELLs, test specifications for both Title I and Title III testing should proactively incorporate principles of Universal Design. The central idea of Universal Design is that assessments should be built from the onset to be accessible to the widest range of students and with the needs and characteristics of all student groups in mind (Thompson, Johnstone, and Thurlow, 2002). The developers of Title I assessments should be cognizant that the language demands of all assessed content must be consistent with only what is required to demonstrate mastery of that content. Additional language complexity can become a major source of construct irrelevant variance. Ensuring that teachers and other content experts who draft, edit, or review potential test items receive training on the similarities and differences between ELP and content standards, as well as the characteristics and needs of ELLs (and the conditions that affect access). Sato, Rabinowitz, & Gallagher (2008) have identified several features of language that create difficulties for non-native speakers (e.g., context, vocabulary, syntax, graphics, language load, grammar, sentence structure). Good training for item writers can mitigate the effect of these pernicious factors on assessments for ELLs.

With respect to item formats, Title I testing typically consists of multiple-choice items, sometimes exclusively, but often in combination with constructed-response items. In contrast, Title III testing must venture beyond the traditional paper-and-pencil, multiple-choice assessment format because it requires assessment of ELP skills not amenable or easily assessed through that approach. Specifically, Title III calls for assessing oral language, which requires that students produce language rather than select a response. Likewise, it requires assessment of listening skills, which means that the item stimulus must be oral rather than written. Finally, Title III requires assessment of writing skills, which typically involves constructed-response (e.g., writing prompts) as well as multiple-choice item formats.

Because Title I testing does not explicitly call for assessment of listening, speaking, or writing, there is less inherent in the content that requires item formats beyond multiple-choice. Nevertheless, numerous states include constructed-response items for Title I testing in order to ensure more comprehensive and valid measurement of student achievement.

4. Draft items. Once the item specifications are established, item writers draft items to be consistent with these specifications. States should ensure that item writers are fully cognizant of the requirements of both content and ELA standards and blueprints, regardless of which test they are developing. Training for item writers should include models of accessible items for ELL (and other at-risk student populations) and common

pitfalls to avoid that might hinder access (e.g., complex sentence structures in item stems), and thus increase bias and limit test validity.

States might also consider expanding their Title I content standards and testing programs to include item types used in Title III testing. Doing so would not only promote consistency and coherence, but also reinforce a more comprehensive view of *content proficient* that requires students to construct rather than just identify their content knowledge. Where there are similarities in item specifications and content between Title I and Title III testing, the same items could be used for both programs, thus creating efficiencies in item development. (The advantages of appropriate joint content are described in various sections below.)

5. Conduct content and bias review of items. A fundamental step to ensure the validity of test items under development is review by external committees. Draft items are reviewed by knowledgeable, trained individuals for overall quality, content appropriateness, and absence of bias. Content and bias review committees are typically set up to conduct such reviews.

Inclusion of ELLs in Title I and Title III testing has implications for the make up of the content and bias review committees. Specifically, linguists should be included on the content review committees and the bias review committees to minimize language barriers that could limit access of the test to ELLs. Also, individuals with relevant content (e.g., reading/English language arts) or linguistic expertise might be used for both Title I and

Title III content review committees. In fact, the bias review committees for Title I and Title III testing could be merged into one committee, since the ideal composition for both bias review committees would be similar (e.g., linguists, educators, community members). Having one bias review committee would cut down on the resources and costs associated with recruitment, training, and review and would ensure that the needs of ELLs are at the forefront of both types of review[4].

6. Pilot test items and revise items using information from results. Prior to more formal field testing, assessment items can benefit from small scale, less formal tryouts, particularly those item types that are difficult and expensive to assess (e.g., writing prompts, constructed response, performance tasks). Including ELLs in Title I and Title III testing and reporting their results means that more attention should be paid at the pilot test stage on how these students perform on various item formats across the range (breadth and depth) of assessed content. *Cognitive labs* and interviews with ELL students on how they perceive and work through specific items would provide useful information during the pilot test phase. A cognitive lab "is a method of studying the mental processes one uses when completing a task such as solving a mathematics problem or interpreting a passage of text" (Zucker, Sassman, and Case, 2004, p.2). Cognitive labs yield valuable qualitative data for refining as well as validating assessment items. Specifically, information on which content strands, item types, and specific language features of items create the greatest difficulty for ELLs can inform item development (step 4, above) and be used in the set of validity studies described in detail below (step 11).

[4] As indicated in a previous footnote, similar considerations should be made with respect to other student subpopulations such as students with disabilities.

While cognitive labs and interviews are not yet a standard part of the test development process, their addition should be considered given the inclusion of ELLs and other subpopulations in Title I testing, whose interactions with items can be more complex, requiring additional research and attention. States may find this approach especially helpful as new standards and their related assessments are being developed and phased into a comprehensive system. Periodic check-ins with samples of students after full implementation of the testing program are also desirable to see if accessibility is increased over time due to increased content familiarity and changes in support services and instructional practices.

During the pilot stage of Title III testing, results of distinct subgroups of language learners (e.g., students from different language groups; students with different levels of language proficiency across the domains of language) should be examined, if the sample sizes are sufficient. This will help ensure that items are appropriate for the major ELL subpopulations.

Finally, proposed accommodations should also be examined during the pilot test stage so that necessary adjustments to accommodation policies could be identified and incorporated prior to a full-scale field test. Accommodations should be research-based and specific to the needs of ELL students across the full continuum of language

acquisition and other relevant cultural considerations. They should not simply be adapted from state policies for students with disabilities (Abedi, 2001; Rivera & Collum, 2006)[5].

Once items are pilot tested, the results are examined and revisions (or deletions) are made to items based on these results. Items that survive the process are ready for formal field testing.

7 & 8. Field test items with a representative sample of students and conduct data analyses. The field test is considered a "dress rehearsal" for the official, operational assessment. Items are included in final form and assessed under conditions as close to a live administration as possible (e.g., using the same administration time parameters and taking items in the same order and location on the test booklets). Ideally, field testing should be embedded within an actual live administration to ensure student motivation is not a factor in interpreting the results of the item tryouts. However, for ELL students, this potential advantage needs to be weighed against the need for a longer testing administration time, since both live and field test items would need to be included on each form. Furthermore, the possibility exists that some field test items will not perform well for ELLs; hopefully, such relatively poor (non-counting) items will not have differential negative effect on overall ELL performance on the Title I test.

Regardless of approach, it is essential that field testing involve a sufficiently large, representative sample of students. This means that for Title I testing, sufficient numbers

[5] An updated look on the effectiveness of accommodations for ELL students is under development (Rivera et al, 2008) as part of the set of LEP Partnership developed papers.

of ELLs must be included in the field test sample (across all field test forms), including as feasible, students across the range of language groups and degree of English proficiency. Similarly, for Title III testing, sufficient numbers of *major subgroups of ELLs* must be included, allowing analyses across languages and proficiency level. As compared with pilot testing, it is possible to conduct more formal analyses following field testing. Differential item functional (DIF) analyses should be conducted to ensure that items are performing appropriately for ELLs (i.e., bias free and accessible). For Title I testing, DIF analyses should be conducted at a minimum for ELLs as a group, whereas for Title III testing, DIF analyses should be conducted on major ELL subgroups, as available and appropriate. Results from both sets of analyses should be reviewed in conjunction to determine if certain item types and language features consistently disadvantage ELL subpopulations or ELLs as a group.

9. Assemble items into operational test forms. Based on field test results, operational test forms are built consistent with the test specifications and blueprints. Depending on what conclusions have been drawn at previous development stages, the final test forms could create the opportunity to *directly* transfer information across the Title I and III programs.

Such opportunities could occur at several levels. At the most ambitious, a state might build in identical items or content on both assessments if they determine from a joint review of their ELP and content standards that sufficient overlap exists to justify this strategy. Use of identical items would allow for less testing time, whereas use of overlapping content would result in broader content coverage. Overlap across content

and specific items could also provide joint scaling opportunities across the two programs (see section *Conduct technical analyses* below for a fuller description of potential benefits of a common scale).

If building in content overlap is not deemed appropriate or desirable, content should be coordinated across the programs, ensuring that the ELP (Title III) assessments contain language specifications and items that scaffold (i.e., support) the academic language needed to be successful in each of the various content areas. States should conduct alignment studies of ELP assessments to each content area the state assesses, beginning with the NCLB-required areas of reading/English language arts, mathematics, and science, and potential other areas (social sciences and beyond), to ensure that English proficiency translates into academic readiness (from a language perspective).

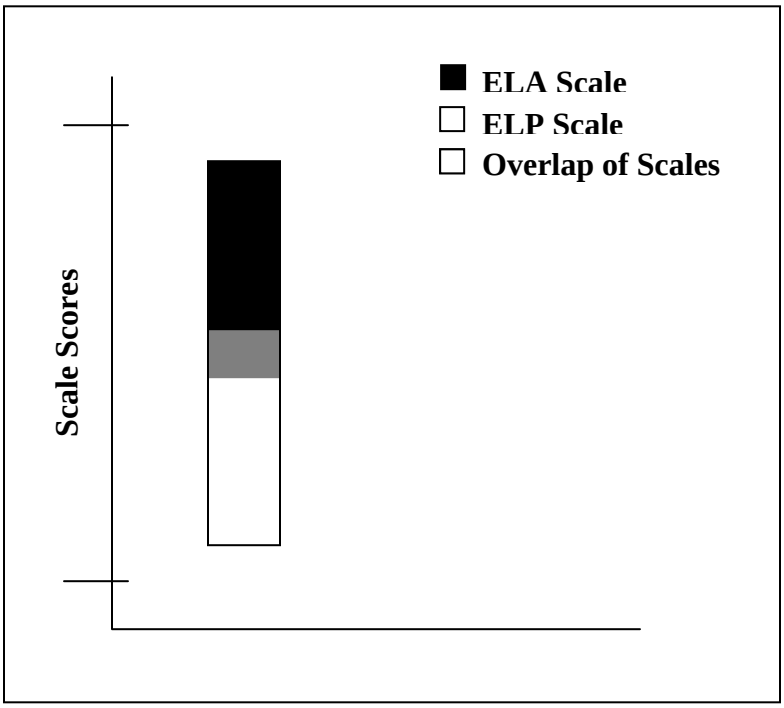
10. Conduct technical analyses (reliability, equating). Increasingly, states are examining the technical adequacy of their high-stakes assessments using more sophisticated approaches. For example, rather than just computing overall reliability indices, many programs report the reliability at key points of the score scale, such as the cut score which divides proficient from below proficient performance. States may also expand their analyses to include differential reliability estimates for key subgroups, such as ELL students. Since each subgroup has school and district accountability decisions linked to their test performance, differential reliability rates could provide states with evidence of the confidence they can have in these judgments. Lower reliability for some subgroups will help identify areas of potential invalidity that states will need to address.

As indicated above, benefits could also accrue from formally or informally placing Title I and Title III assessments on the same score scale. If this proves to be possible, states would have a common way to describe performance across the two assessment programs. The stronger the relationship, the more the state will be able to track the interaction between the attainment of language proficiency and mastery of grade level content standards.

Using a common score scale for both assessments requires satisfying certain conditions regarding content and dimensionality. With respect to content, the standards on which both assessments are based would need to overlap sufficiently. To determine whether or not there is sufficient content overlap, states could apply the criteria they use for other aspects of their assessment program when determining if a vertical scale is supportable (e.g., Is the content continuous? Is the relationship between standards primarily linear?). To meet dimensionality conditions, various forms of comparability would also be required. Take, for example, the potential relationship between performance on a state's ELP assessment and ELA test. If the standards suggest a linear relationship between the two and student performance supports this supposition, then the top of the ELP scale (equivalent to full English-Language Proficiency) could overlap with the bottom of the ELA content scale. This would set English proficiency as a necessary but not sufficient requisite for performance on the ELA test. (See Figure 1 for an idealized depiction of this relationship.) Other relationships among ELP and academic assessment content scales could be postulated and investigated, such as research and language-based relationships

built around specific strands or subscales of the various content assessments with the four ELP domains, or overlap of the dimensionality of the ELP and content scales. For example, because of similarity in content, format and mode of test administration, the reading and writing domain scores for the ELP assessment may correlate more highly with the content assessments than with the speaking and listening subscores. States may want to limit their scaling studies to the former domains only, excluding information from either the speaking or listening sections of the ELP assessment.

Figure 1: Idealized Scale Overlap, ELA and ELP Assessments



In order to perform this scaling, states need to have (1) a sufficient number of common items across both forms (as dictated by the overlap in the content standards and test specifications) or (2) common students across both tests (a very strong likelihood given

Title I and III requirements[6]). Satisfying the first condition would create the most valid scenario because sufficient content overlap is an important pre-condition to a strong interpretation of scale overlap. Success at scaling using the common item approach can also be used as evidence of overall validity of the assessments and the common standards that may underpin them. (More detail on joint validation efforts is presented below.) However, even meeting the second condition (i.e., use of common student scaling alone), may be beneficial for states because it allows them to talk about how performance across the two testing programs is similar and different for students at various levels of English proficiency and academic achievement.

11. Conduct validity studies. States have an explicit obligation to demonstrate the validity of their assessments for all stated purposes, especially those accountability decisions that affect students and schools (AERA/APA/NCME, 1999). As such, planning joint validity studies between ELP and content assessments could have great payoff for both programs, providing an efficient and authentic means to judge their technical quality and effectiveness.

The focus of validity studies differs somewhat, depending on the stage of development and maturity of the assessment program(s). For example, when assessments are newly developed, validity studies should focus on defining the targeted constructs and ensuring that items align to the standards that operationalize those constructs. As tests become

[6] In some cases, states may indeed have common students (as required) but due to limitations in data tracking systems or lack of a unique student identifier, they may not be able to match student performance across assessments.

operational, the emphasis of validity studies should shift to accessibility (lack of bias), reliability, and generalizability.

Specialized Validity Studies for Title I and Title III ELL Assessments

Whereas all testing programs require validity studies, more large-scale inclusion of ELLs in state testing programs requires new specialized types of these studies. The validity challenge is even greater given that the tests themselves have changed over the past decade, from a focus on social/conversational to academic language. Finally, the NCLB requirement that the results of this population be included as parts of both Title I and Title III school accountability means that states must be sure that the results of ELL students, from both ELP assessment and content assessments, are truly valid.

This section presents a series of specialized validity studies that will allow information across Title I and Title III assessment to inform and improve both programs. Most of the studies included will not require states to change administration practices or collect any additional information than they are already receiving from these programs (standard practice). We also recommend other studies (non-standard practice) that require additional effort, the value of which should outweigh the extra burden.

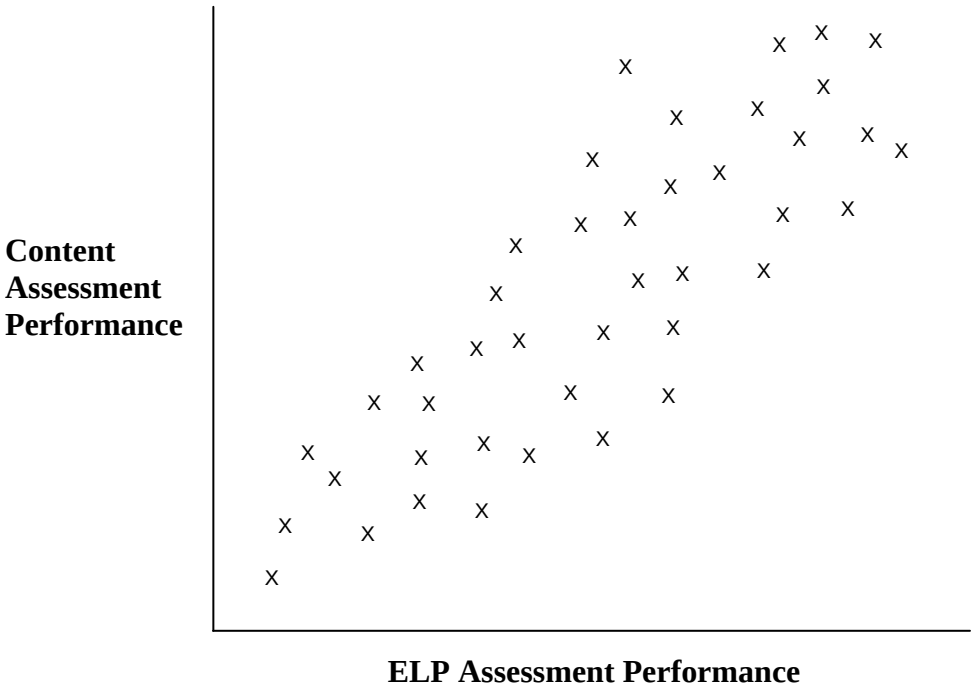
Standard Practice Specialized Validity Studies

As described in Table 1, states under NCLB must administer both content assessments and ELP assessments to their ELL population and content assessments to their non-ELL students. This provides a great opportunity to formally analyze the relationship among

these various assessments *without creating any extra burden on either students or schools.* However, while states routinely meet these administration requirements, very few systematically analyze the relationship of student performances across these various assessments, nor link such performances to other indicators of success (e.g., local assessments).

Category I Studies: ELL performance on ELP assessments vs. content assessments. The validity of assessments is directly related to whether they fulfill their primary purpose(s). Specifically, states should be able to predict (based first on linguistic theory and then supported by empirical data) how students of various ELP levels should perform on content assessments. The typical expected relationship can be seen in Figure 2 below, where increased levels of English proficiency are associated with higher content achievement scores.

Figure 2: Expected Relationship between ELP and Content Assessment



Francis and Rivera (2007) begin the discussion as to what extent ELP assessments can and should predict success on content assessments. They examine two state data sets, focusing on the direct relationship between the two types of tests as well as whether some intervening variable (e.g., years in the U.S.) affects that direct relationship. This section expands on that work and includes other relevant model studies.

States need to be able to address the following questions in order to more fully validate both their ELP and content assessments:

- *How strong should the relationship be between ELP level and content mastery?* Increased levels of English proficiency should result in higher content achievement since facility with English should certainly improve students' ability to access classroom information and other support materials (e.g., textbooks). However, if the relationship is too high (e.g., .8–.9), then the tests are providing redundant information; too low a correlation (< .3) might suggest bias or other sources of invalidity. States need to determine this relationship and whether there are certain student groups who fall outside the typical pattern.

- *Should the relationship between ELP levels and content mastery differ by content area?* While all assessments administered in English have a language load, the degree of language dependency differs across content areas. For example, ELA assessments should require a greater degree of English proficiency than a mathematics test. States should verify whether this is indeed the case, through correlational analysis and/or expert judgment. If so, the validity of all three instruments—ELP assessment, ELA assessment, mathematics assessment—is supported. If not, states need to review the language expectations of their non-ELA assessments (mathematics and others) to ensure that measurement of true content achievement in not being suppressed by construct-irrelevant variance.

- *Should the relationship between ELP levels and content mastery differ by language group (or other demographic indicators)?* While politically sensitive, evidence across NAEP and many state assessment programs indicates that

performance differs across ethnic groups on content assessments. States need to be aware if differences in the relationship (e.g., correlation) between ELP levels and content mastery also exist across ELL subpopulations. Such evidence might suggest differential validity and potential bias against lower performing subgroups. The purpose of these analyses is to develop strategies to move beyond such findings by developing more valid assessments or more effective instructional practices geared to the needs of the lower performing students.

One outcome of the series of validity studies described above is that states can develop prediction tables that indicate the likelihood that a student at a given English proficiency level will score at a given content achievement level. These tables can be computed separately for each content area, showing at a glance how performance across the various tests differs by proficiency level, achievement level, and content area. Table 2 presents a hypothetical prediction table. The table supports the expectation that increased levels of English proficiency results in greater likelihood of content proficiency. It also supports the notion that the mathematics test has a lower language load than the ELA test. Taken together, these two findings would support the validity of the ELP assessment and both content assessments.

Table 2: Hypothetical Prediction Table (ELP to ELA and Mathematics)

ELP Proficiency Level	Probability of Basic on ELA / Math Assessment	Probability of Proficient on ELA / Math Assessment
4	20% / 10%	60% / 70%
3	30% / 35%	40% / 50%
2	20% / 30%	20% / 25%
1	10% / 10%	5% / 5%

Category II Studies: ELL performance vs. Non-ELL performance on content assessments. States routinely report significant performance differences between ELL and non-ELL students on their content assessments. This is not surprising given the language load of state assessments and other factors more common for ELL students (e.g., poverty, years in school) that are associated with academic achievement. States should conduct analyses that delve more deeply into performance differences in order to further validate their assessments and pinpoint areas for instructional focus. The following questions can help both validate the content assessments and inform strategies to improve performance of both ELL and non-ELL students.

- *Are there content strands that show different patterns of performance than overall test score?* Content strands differ in the language load found in items developed to assess them. For example, items testing a mathematics strand focusing primarily on computation should be less affected by English proficiency level than content requiring word problems measuring more complex mathematical concepts. Predicting and verifying where such differences are to be expected and found (and not found) provide evidence that both the ELP and content assessments are performing in a valid fashion.

  For example, just as items are often classified based on depth of knowledge (DOK), they could also be classified using a committee of content experts and applied linguists on expected language load. The resultant language load score for each item can then be compared with the items' p value or IRT difficulty level

(b parameter). A pattern of higher language load/difficulty correlations for ELLs vs. non-ELLs would be positive evidence of the validity of both the state's content assessment and the ELP assessment.

While performance at the student or classroom level is typically not reliable enough to support these types of analyses, school, district, and statewide performance should be sufficiently reliable to allow meaningful comparisons and draw generalizable conclusions.

- *Does performance differ across item types (i.e., multiple choice vs. open response/writing samples)?* Most states incorporate multiple item types in their statewide content assessments. Typically, this involves the use of multiple-choice items supplemented with open response and writing samples. The latter two methods require a level of language proficiency that generally exceeds what is needed to respond to multiple-choice items. States should examine any differential patterns that exist (or don't exist) between ELL and non-ELL students across these item types. To the degree that constructed response items influence total scores (in some states, the weighting approaches 50%), ELLs might be differentially disadvantaged. States might consider the use of alternate assessments or accommodation in these instances (see below).

- *Are there strategies that might mitigate the differences across content strands and/or item types?* If content assessments are believed or found to underestimate

the performance of ELLs, states may choose from a range of strategies to "even the playing field." Such approaches range from providing appropriate and targeted accommodations (Rivera, 2006), on through more time and cost intensive ones such as providing translations and modified English versions of the content assessments (Abedi, 2007; Sato, 2007; Stansfield, 2007). As states consider options, they must calculate the time, effort, and cost required to perform comparability studies between the original and modified versions of the content assessments, as well as whether they have the in-house content, language, and psychometric expertise to implement such a challenging process. As appropriate and necessary, states may need to turn to their contractors, consultants, and technical advisory committees to successfully move in this direction.

- *Are there state or local practices that mitigate the differences?* Finally, states need to identify sites where larger than expected numbers of ELLs perform well on content assessments. Researchers such as Popham (2005) have questioned the instructional sensitivity of large-scale, high-stakes assessments. States have an obligation to show that good instructional practice leads to higher academic achievement (as measured by state assessments). Such pockets of excellence can also serve as model programs for statewide dissemination.

Non-Standard Practice Specialized Validity Studies

The various studies described above are relatively less difficult for a state to implement than the ones to be described in this section since most data elements needed to perform

them already exist. In this section, we describe additional studies that will further the knowledge of the validity of language and content assessments on the states ELL population.

Category III Studies: Non-ELL performance on ELP assessments. States rarely administer their ELP assessments to non-ELL (e.g., native English speakers) student populations. On the surface, the suggestion might seem superfluous or unproductive: Why administer a proficiency test to a population already presumed to be proficient? However, certain assumptions underlying that question need further examination. Most important is the implicit belief that the ELP assessment is indeed a valid measure of English proficiency. If such were the case, then native English speakers should routinely achieve at least at or near the proficient level. Rather than assume this to be the case, states should demonstrate that presumed proficient students meet the standard of proficiency; and all subsections of the ELP assessment support that decision.

For example, how might a state interpret the likely finding that a larger than expected number of native English speakers do not meet the highest achievement level of their ELP assessment? To answer that question states need to carefully examine:

- The content that is included on the ELP assessment, looking especially at the breadth, depth, and range of complexity of the expected domain (modalities) of language to support academic content achievement;

- The placement of the proficiency level (mastery cut score) for the ELP: is the standard reasonable and related to the purpose of the test?

- The differences between academic language and socio-functional language skills for native speakers, especially those who are not achieving well on content assessments. Many non ELLs are not able to achieve proficiency on the states ELA and other content assessments. This lack of achievement may be related to facility with the demands of academic English.

The validation of ELP assessments routinely focuses only on ELL students. Expanding such studies to include both non-ELL students and formerly ELL students will provide a level of increased confidence that the content and cut points are appropriate and valid, and that all students possess the requisite academic English skills to master challenging academic content across a range of subject areas.

Category IV Studies: ELL and Non-ELL performance on ELP and content assessments vs. other indicators of academic success**.** States should plan concurrent, predictive, and consequential validity studies to investigate the following:

- *What measures/outcomes should correlate with performance on Title I and Title III assessments?* A common practice to validate an assessment is to determine what other indicators are theoretically or practically related to it and then determine if performance across these measures do indeed correlate as expected (Cronbach and Meehl, 1955). The studies described above that encourage states to

determine the expected and empirical relationship between their ELP and content assessments are examples of concurrent validity evidence. States should consider what other types of indicators might be used to demonstrate the validity and usability of their ELP and content assessments for ELLs. Do higher levels of English proficiency or content mastery correlate positively with class grades, teacher observations, and other assessments (e.g., SAT/ACT, locally administered norm-referenced tests)? And do they correlate negatively with unfavorable indicators such as suspensions and other behavioral measures and dropout/graduation status? Validity is often a cumulative process—no one measure will allow states to say this "proves" the test is valid. However, the more evidence that points in that direction, the more states can feel confident that their ELP and content assessments are valid for the full range of their ELL students and all intended purposes.

- *What is the effect at the classroom level of Title I and Title III policies, including how the standards and assessments are used to support instruction?* While assessments play a key accountability function, they are also expected to affect classroom behavior. This can happen in two ways. First, the results of assessments (either the proficiency decision or relative mastery of content) should change teacher instructional practice, focusing on students' areas of deficit. Second, the content and format of the test itself should change classroom practice. This is a major reason states include constructed response items on state assessments—to signal to educators that the type of reasoning required to do well

on such items should be a part of regular classroom activities. States cannot assume that either benefit of their assessment programs is occurring just because the test is in place. Carefully designed studies—consisting of surveys, observations, and examination of local artifacts (e.g., lesson plans before and after the implementation of the testing program)—are an essential step to determine whether ELLs are receiving the type of instructional support required to both achieve English proficiency and master the expected challenging academic content contained in the state content standard and measured by the state's content assessments.

In examining the consequential validity of these assessment programs, states might address the following key question: Do the results of Titles I and III assessments allow teachers to work in a coordinated, coherent way to guide students through the dual challenge of language and content achievement or do strategies typically employed at the district, school, and classroom levels work at counter purposes? Studies of successful (and ineffective) practices could both (1) inform needed revisions to standards and assessments for both testing programs and (2) serve as guides for improving professional development on good local practice.

Concluding Remarks

In their efforts to satisfy the various Title I and Title III requirements, states should not lose sight of the need to build a comprehensive approach to the assessment of their ELLs. While recognizing that the assessment goals of Title I and Title III are not synonymous,

this paper has argued that there is sufficient overlap to (1) create efficiencies in the development and administration of the respective assessments and (2) use the findings of either assessments to supplement and reinforce the other.

All tests must follow the same basic development protocols. However, states may have unnecessarily erected "walls" between the processes for selection or development of academic content and English proficiency testing. By breaking down the walls between the two testing programs—whether it is in regards to conceptualizing ELP and content standards, processes to review draft items, or the procedures to evaluate the technical adequacy of items and score inferences—both Title I and Title III tests can achieve greater construct and instructional validity.

Perhaps most importantly, a comprehensive approach to ELL assessment requires a systematic examination of the theory and practice-based relationships between language acquisition and achievement of academic content. States must endeavor to better understand the interactions between language acquisition and academic content achievement because these are crucial and mutually-supporting goals of schooling for their ELL populations. These interactions have important implications, particularly for the development of:

- ELP and content standards,
- Title I and Title III assessments, and
- instructional strategies.

States can and must take into account the complexities in the relationship between language acquisition and achievement of academic content in order to achieve greater accessibility, validity, and support in assessment of ELLs.

References

Abedi, J. (2001). *Assessment and accommodations for English language learners: Issues and recommendations.* Policy Brief 4. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. http://www.cse.ucla.edu/CRESST/Newsletters/polbrf4web.pdf

Abedi, J. (2007). *Language factors in the assessment of English language learners: The theory and principles underlying the linguistic modification approach.* Paper developed for the US Department of Education LEP Partnership.

AERA/APA/NCME (1999). *The Standards for Educational and Psychological Testing.*

Chronbach, L.J. & Meehl, P. E. (1955). *Construct validity in psychological tests.* Psychological Bulletin, 52, 281–302.

Francis, D. J. & Rivera, M. O. (2007). *Principles underlying English Language Proficiency Tests and Academic Accountability for ELLs.* In Abedi, J.: English Language Proficiency in the Nation: Current Status and Future Practice, University of California, Davis, School of Education.

No Child Left Behind Act: *Assistance from Education Could Help States Better Measure Progress of Students with Limited English Proficiency.* GAO-06-815. Washington, D.C.: July 2006.

No Child Left Behind Act: *Education Assistance Could Help States Better Measure Progress of Students with Limited English Proficiency.* GAO-07-646T. Washington, D.C.: March 2007.

Popham, W. J. (2005). *'Failing' Schools or Insensitive Tests?* The School Administrator.

Rivera, C., & Collum, E. (2006). *State assessment policy and practice: A national perspective*. Hillsdale, NJ: Earlbaum.

Rivera, C. et al (2008). *A Handbook of best practices in test accommodations and state assessment policies for English Language Learners.* Paper developed for the U.S. Department of Education LEP Partnership.

Sato, E. (2007). *A Guide to Linguistic Modification: Strategies for Increasing English Language Learner Access to Academic Content.* Paper developed for the U.S. Department of Education LEP Partnership.

Sato, E., Rabinowitz, S., & Gallagher, C. (2008). *Access and Special Student Populations—the Similarities/Differences in the Needs of English Language Learners and Students with Disabilities: Implications for Standards, Assessment, and Instruction* [working title].

Stansfield, C. W. (2007). *Sight Translation of Assessments: Answers to Frequently Asked Questions*. Paper developed for the U.S. Department of Education LEP Partnership.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Zucker, S., Sassman, C., & Case, B. J. (2004). *Cognitive labs.* San Antonio, TX: Harcourt Assessment. Retrieved 10/11/2006 from https://harcourtassessment.com/hai/Images/resource/library/pdf/CognitiveLabs_Final.pdf.