CHAPTER 4

Getting Started: Issues of Participation, Alignment, and Validating Access with Test Specifications

Building a large-scale achievement assessment system that properly recognizes the diverse population that will be taking the tests requires a complex, comprehensive approach. Issues such as for whom the test is intended, what groups will participate during the development of the assessment, and how they will be included, should be an integral part of the test plan. Further, it will be important to provide evidence that reasonable levels of alignment have been achieved among the assessment system, the content standards that drive the assessment, the performance levels that codify the levels of achievement associated with various scores on the assessment, and the enacted curriculum with which the students engage. This requires early identification of presentation, administration, and response options suitable for this population. It also requires the identification of alignment procedures and alignment targets, followed by the development of test specifications or blueprints outlining what will be tested and how. Test specifications provide the construct framework for operationalizing the test design through subsequent item development. Considerations relevant to English language learners (ELLs) need to be addressed at each of these points in order to construct a system that is appropriate for this population. This chapter extends the discussion of how to include ELLs in tests given to the general population, and the importance of including ELL experts and students in the substantive phases of test development. It then illustrates how alignment might mean something different for English language learners than for their native-speaking counterparts. Finally, a discussion of the creation of test specifications that will allow for more accurate testing of ELLs will be summarized.

Participation in Test Development

The field's measurement standards (AERA/APA/NCME, 1999) summarize how the testing discipline interprets its responsibilities to appropriately measure the targeted content in K-12 test takers. As Chapter 2 outlined, federal legislation and many state and district policies have driven the field to develop testing mechanisms which assess academic learning of a broader range of their student population than was true in the past. Further, the previous chapter explained that, in order to hold all students to the same standards of performance and be able to place them confidently on a common performance scale, it is advisable to build these large scale assessment systems that yield comparable scores across virtually all English language learners (Haertel and Wiley, 2003; Kopriva, 1999b).

This book will outline in some detail the procedures that should be able to improve mainstream testing systems for the range of ELLs. To further increase the likelihood that a coordinated assessment system represents that achievement of all children with comparable accuracy, English language learners should be represented during all stages of test development. ELL experts need to be included during test specification and item development, as well as during reviews. A full range of ELL students needs to be involved in item pilots and field tests.

Expert Participation in Test Development

Typically, experts with substantive knowledge of English language learners have been used primarily in bias reviews, where the charge has been very narrow and incomplete (see Chapter 5 for an explanation). They have not been included in the planning, item development, and decision-making processes to the same extent that mainstream teachers and content experts have been in recent years. This includes involvement throughout the design, construction, and technical phases of development. The tasks in which it would be appropriate for them to actively participate include the following:

- designing the comprehensive testing system;
- developing test specifications;
- writing and reviewing content items and rubrics that are appropriate for ELL population;
- providing training to other item writers and developers;
- trying out items in classes;

- evaluating forms for coverage and accessibility;
- making decisions about inclusion or exclusion of items, all testing materials, and administration/response options based on data from pilots, field tests, and other technical data collections; and
- scoring, reporting, and making decisions about test use for accountability and program evaluation.

English language learner experts who bring the most to the test development process have a deep understanding of content standards, experience with adapting academic teaching environments for these students, and knowledge of their students' strengths and challenges. Test developers should ask for evidence of these characteristics. Because few experts will have experience with the full range of English language learners, it will be necessary to include a variety of educators who bring diverse capabilities to the entire test development process, from designing the system to scoring, and making accountability decisions. Future chapters will, hopefully, make clear how this range of expertise can be used to make development decisions. Relevant expertise encompasses the following:

- educators from classrooms in which students are learning English as well as grade-level academic content;
- educators from mainstream academic classrooms in which ELLs are placed after they have reached a certain level of English proficiency;
- educators working with students who are newly arrived to the United States;
- educators working in classrooms in which the students' primary language (also known as their first language or L1) is the language of instruction or in bilingual (L1 and English) classrooms;
- educators with urban experience and educators with rural experience;
- educators working with migrant students; and
- educators who come from the primary language and cultural backgrounds of the students they teach.

Student Participation in Small Sample Pilots and Large Sample Field Tests

Just as the diverse perspectives of multiple ELL experts should be included during test development, so should a full range of English language learners be involved in all item and test data collections. Several researchers suggest that students respond differently based on their proficiency levels and adequate accommodations (e.g. Rivera and Collum, 2006, Abedi *et al.*, 2007; Emick, and Kopriva, 2007), and so participating students should range from new arrivals through former English language learners that have successfully transitioned. Further, Laitusis *et al*., (2004) found that ELL students appear to differ in how they respond to items based on their racial/ethnic background and these distinctions should be examined. As such, it is recommended that enough students from pre-identified language proficiency and racial/ethnic strata be included to be able to analyze the data by these subgroups.

In assessments that include constructed response items, ELLs should be purposefully sampled so that their work can be inspected at all item score points. Some researchers have noted that item considerations based on language or culture will sometimes come to light for those with different levels of targeted ability (Abedi *et al*., 2001; Kopriva and Lowrey, 1994; Kopriva, 2005b).

Test Specifications and Content Validity

Test specifications, sometimes called test blueprints or frameworks, outline the content and constructs each assessment in a testing system will cover. Created for each content area and grade level tested, test specifications prioritize and weight the areas to be covered and generally include information about test length and format, item types, and reporting requirements. Each assessment consists of one or more forms that are intended to be parallel, and each of these may or may not have separate specifications associated with them. The test specifications form the bridge from the content standards to the assessment and provide a framework that identifies what the test will measure and what items will be included in the test. This section discusses the various components of traditional test specifications and discusses the addition of a new component—access—to provide information about the appropriateness of the test for ELL students.

Components of Test Specifications

Test specifications represent a translation from the curricular priorities established in the content standards to a design for how assessment developers will evaluate what students have learned. As discussed above, assessments may cover a large or slim portion of the content standards. Further, they may either match or understate the range of academic complexity present in the standards, including the sophistication and depth of knowledge and skills the students are expected to learn within content topics.

Developing test specifications from content standards is one of the first procedures in building an assessment. Test specifications are used consistently as items are developed and content and complexity coverage is evaluated. Further, once the test is completed, test specifications are one of the best indicators of content validity, or the determination of what is being covered on the test and how. While specifications vary in "grain size" or specificity within and across topics, they should be precise enough to retain the integrity of the standards on which they are based without turning into highly constrictive objectives.

Test developers use specifications to summarize the content upon which all test takers will be tested. Test specifications form the basis for content area inferences about what students know and don't know. These inferences are valid to the degree that items properly reflect the standards they are intended to reflect and that test takers are able to understand the nature of the items, solve the problems to the extent they know the content, and explain their solutions. The inferences would be incorrect to the degree that test items distort what is being measured and that other variables influence student scores. Linn (1993) encouraged the assessment community to develop methods for tightening test specifications because they are the linchpin to ensure that important aspects of the standards are properly measured in the assessments and reported to stakeholders. Part of his emphasis was on ensuring that the blueprints specify the levels of complexity that tests include. He also indicated that differential item functioning and other information about performance by subgroups should be taken into account in test development.

Today, topic coverage continues to be one of the major components of test specifications, often including specifications of item types (for example, multiple choice, brief constructed response, essays) and item difficulties around anticipated points such as achievement-level cut points. Complexity coverage is frequently a part of test specifications as well. While item difficulty and/or item type are sometimes used as poor surrogates for complexity, many publishers and agencies understand that complexity refers to a range of skills, from basic to more sophisticated, as they are associated with content (see NAEP mathematics test specifications, 2001, for example (Solomon *et al*., 2001)). In the mid-1990s, the Delaware Department of Education developed a mechanism that defined complexity in items, which it called depth, as a function of four variables:

- approximate time required to complete an item;
- item scaffolding (ranging through three levels from step-by-step task guidance to no hints or guiding questions);
- level of generalization (in three levels, from highly specific items to items that require generalization); and
- complexity of process (in three levels, ranging from managing only a limited amount of information to processing or considering multiple pieces of information and/or procedures simultaneously).

Each item is coded on all four variables. In assembling forms and tests, depth balance across items is guided by the content standards. Delaware's approach was developed from the work of Rigney and Pettit (1995), who suggested criteria for assessing the depth and quality of student work in portfolios that were assembled as large-scale evaluation tools.

While coverage and complexity are routinely covered in test specifications, issues of accessibility are not typically made explicit, which means that test developers and test users do not gain a clear understanding of the extent to which test takers' scores reflect variables irrelevant to the targeted construct. Test development procedures such as analyzing differential functioning of some items in some subgroups and conducting bias reviews, along with the use of *post hoc* accommodations, are seen as ways to minimize problems. However, to date there does not seem to be any systematic way of dictating standards of acceptable access, specifying points where access should be displayed, and summarizing that adequate work has been done. For test score inferences to be used with confidence, it seems necessary that proper access to intended item content be documented for test takers, particularly those with identified challenges or limitations that run counter to the methods typically used in large-scale testing.

Documenting Content Validity for English Language Learners: The Access Specifications Package

English language learners are among the population of students whose challenges often affect the accuracy and validity of their test scores. This book explains various ways to improve the viability of their scores; here, it is argued that summary documents—referred to here as Access Specifications—also need to be put into place to explicitly insist on and monitor proper access to test coverage over the entire breadth of topics and the depth of cognitive complexity. The documents should target access at both the item and test levels, providing a planning and evaluation framework for access-driven item writing and test-level accommodation decisions, as well as evidence to demonstrate the validity of test scores. Test specification documents governing accessibility should also set an acceptable standard for matching procedures so that students with identified needs and challenges are matched to appropriate accommodations that address their barriers appropriately. Such access specifications would be used for the following purposes:

- to regulate planning and development in order to ensure acceptable levels of validity in score inferences across diverse students, including ELL and native speakers;
- to serve as a formative evaluation tool from which to make mid-course corrections; and

- to provide evidence to document the level and quality of access across the test and test development and implementation procedures.

At the core of the access specifications, as with other test specifications, would be the same content standards and targeted constructs, clearly specified. The intent of developing and using the access specifications is that the test score inferences for English language learners will be defensible and comparable to the inferences of test scores for other students taking the assessments.

As envisioned here, test developers should complete an Access Specifications Package. This package would comprise four documents: an item matrix, an accommodations summary, a description of the matching procedure for assigning accommodations, and a scoring, analysis, and reporting survey. Each of the four documents would summarize the expectations, criteria, and objectives to be used to guide test development, and then allow for the presentation of summary evidence that the work was completed. The four documents that make up the Access Specifications Package are described below.

Access Specifications Package

Access-Based Item Matrix The Access-Based Item Matrix allows each item in a test form to be rated for accessibility according to the following criteria: language simplification, reduced reading load, graphics, format simplification, and other appropriate item modifications (e.g., problem context); use of tools and resources; administration options; response options; and rubric accessibility (for constructed response items only). The matrices are completed by filling out the appropriate information by item, for each form of the test. Evaluators are encouraged to be detailed and specific about their findings and consider the composite effect of various criteria on the accessibility of each item. While it is unrealistic to expect that all items will have graphics, for example, it is reasonable to expect that attention is paid to broadening accessibility in some ways for all items.

Item Matrix documentation would include a description of criteria that would need to be addressed for each item (e.g., what constitutes language simplification or use of tools/resources) with additional criteria for rubrics for scoring constructed response items (for example, anchor papers and scoring notes that illustrate how students might use a diagram). Agencies would prepare an evidence document by reviewing and analyzing each item and filling in the grid as appropriate.

Accommodations Summary The Accommodations Summary identifies and describes all accommodations available for the test and explains the characteristics of the students for whom each accommodation would be appropriate. The accommodations could be grouped under broad categories as follows in Table 4.1. Options such as those listed below are described in several of the subsequent chapters.

The accommodations available and used should be checked off on the Accommodations Summary. If possible, the number of students who received each accommodation should be noted. Where substitutions are made, reasons for them should be noted and an attempt should be made to estimate their efficacy.

Description of the Matching Procedure for Assigning Accommodations This document specifies the procedures for determining how individual students are assigned one or more presentation, administration, and response accommodations. It summarizes the research base for the decision-making algorithm employed and *indicates* what information was gathered for each student and how accurately the decisions were made and how fully they were implemented. Some examples of how developers or agencies might address how to assign accommodations to individual students are summarized in Chapter 10.

TABLE 4.1 Selected Forms, Tools, Administration, and Response Options

Forms

- Standard form
- ELL or low language English form
- L1 or Side-by-side form in the following languages:________________________

Tools

- Bilingual word list/electronic translator in the following languages: ____________
- Picture dictionary
- Manipulatives or other problem solving tools

Administration

- Small group
- Oral L1 in the following languages: _______________________________
- Language liaison in the following languages: _______________________

- Other?

Response

- Written L1 or code-switching; scorers available for the following languages: __________________________
- Oral English
- Oral L1 or code-switching; scribes available for the following languages: __________________________
- Modeled/demonstrated response

Scoring and Analyses Documentation These documents ensure that issues related to ELLs are considered as scoring materials are developed and scorers are trained. They also should impact how scores for forced-choice and constructed-response items are evaluated, analyzed, and interpreted. Examples of descriptions for scoring criteria such as those listed below in Table 4.2 are presented, and agencies check off the items that apply. This is explained more fully in Chapter 11. A. discussion of empirical documentation which is particularly relevant for ELLs occurs in Chapter 12. This includes a validation design, and explicit evidence linking appropriate inferences to the weakening of alternate arguments which could wrongly infer the knowledge and skills of ELL students.

Access specifications are not meant to substitute for additional documentation required throughout the test development and analytic phases, but rather to summarize some of the work that has been done.

Alignment Considerations

Overview

The movement toward standards-based education has prompted educators to pay more attention to linkages among content and performance standards, curriculum, instruction, and assessment. State and local educational agencies, with input from subject experts and subject-area associations such as the National Council of Teachers of Mathematics, specify the knowledge and skills students are expected to gain at each grade level and the performance level considered acceptable. These content standards provide the framework for what schools should teach and guide the selection of appropriate classroom learning materials and activities.

TABLE 4.2 Examples of Scoring Documentation

- Prompts (including descriptions of acceptable response modes) are explicit and clear.
- Scorers understand what is being measured, and what is not, in each constructed response.
- ELL experts are included in item and scoring rubric development.
- Bilingual scorers are trained and available.
- Anchor papers representing ELL samples at all score points are provided during training.
- Anchor papers presenting ELL samples at all score points are provided during calibration exercises once the scoring gets under way.
- ELL issues are covered in training and in item notes.

Learners are assessed, both informally and formally, to evaluate their progress in relation to the specified content standards. The state or district defines what levels of mastery are adequate and, in today's parlance, these are referred to as achievement standards. They also distinguish levels of performance that approach the "adequate" level as well as one or more levels that surpass it. These levels might be indicated through designations such as Basic, Proficient, and Advanced or Does Not Meet Standards, Partially Meets Standards, Meets Standards, Exceeds Standards. Large-scale assessment results may be used diagnostically, within the classroom, to modify instruction, as well as for program evaluation and accountability purposes.

In the narrowest terms, alignment may be viewed as how well curriculum standards and tests match. When off-the-shelf or proposed custom-built assessments are examined in relation to content standards, the great majority of test items or tasks—usually 80 to 90 percent and sometimes even 100 percent—match up with what is specified in a state or district's content standards. When the extent of standards coverage is evaluated in relation to assessment coverage, however, it is not uncommon for 50 percent of fewer of the content standards to be measured on an assessment (Hansche, 1998). Another aspect of alignment involves the appropriate emphasis or weighting of content standards. Is the emphasis placed on various topics in the test the same as the intended emphasis in the content standards? Additionally, the measured skills should be consistent with the desired rigor described at each achievement level. Appropriate alignment of content standards to assessments means that the tests and test forms match the emphasis, depth (complexity of the skills being tapped within topics), and breadth (content coverage) of the content standards which are being referred to in the interpretation of the test results.

The issue of how many content standards to assess in any large-scale assessment system is an ongoing debate. On one hand, assessing more standards means that there is more sampling across the testing domain to which the inferences are referring. If the interpretation of test scores is to remain broad (for instance the inference may be a description of how well a particular student or school is doing in mathematics, broadly conceived), sampling is necessary as not all standards, at all levels of complexity, can be assessed on an exam of reasonable length. One consequence of this approach, however, is that breadth tends to outweigh depth in most assessments with the result that the tests are a more superficial evaluation of student knowledge and skills. Popham *et al.*, (2006) have encouraged the education community to test fewer standards but do so in more depth. They argue that this approach would provide more useful information to educators who have to take these results and subsequently develop learning priorities and relevant pedagogy to implement academic improvements, as necessary, in the schools. While more detail about this debate is outside the scope of this book, it is important to remember that, which ever method is used needs to be accompanied by test score interpretation explanations that are consistent with the approach. This is an essential step. Therefore, alignment needs to be not only the alignment of content standards and the assessment system, but needs to include an alignment of the explanations in the achievement standards and other score interpretations which are included in the testing materials.

Several agencies have developed methodologies to evaluate the alignment of tests to standards and achievement levels, ranging from the cursory to the more thorough (for instance, see http://www.ccsso.org/Projects/surveys_of_enacted_curriculum). Virtually all of these focus on the content standards, assessment systems, and on at least some of the interpretations of test scores (for instance, on the published state achievement levels). To date, it doesn't appear that any of these procedures directly address the validity of the alignment matches in terms of different populations of test takers. The closest integrated review system which includes an evaluation of the alignment of test forms and procedures for different populations (as well as other evaluations of other validity indicators) is the U.S. peer review process (USED, 2004, April, Standards and Assessments Peer Review Guidance). This process was first developed and utilized under the previous authorization of the Elementary and Secondary Education Act (ESEA). It has been adapted and used as part of the current authorization as a review of each state's academic assessment systems for purposes of school accountability (No Child Left Behind Act of 2001, 2002).

Finally, while grade-level standards are intended to drive curriculum and instruction, it cannot automatically be assumed that all children at a certain grade level are exposed to the same curriculum. Some teachers may not cover the expected material at the expected level of depth. If students are working below grade level, teachers may need to spend significant time providing remediation before starting in on the grade-level material. Students may miss instruction when they are "pulled out" for various auxiliary services such as reading or language instruction, speech therapy, or gifted or special education programming. They may also miss a significant amount of school due to illness, family travel, babysitting, or other jobs. They may attend schools in which there are disruptions to learning due to lack of materials, discipline problems, inexperienced or frequently absent teachers, etc. Researchers therefore draw a distinction between the intended curriculum and the enacted curriculum and some work has been done to attempt to identify and quantify what types of enacted curriculum students have received (e.g., Schmidt *et al.*, 2001.) In the mid-1990s California attempted to collect information on enacted curriculum when it was conducting its statewide examinations. The intent was to be able to interpret the achievement results within the context of differing opportunity (Burstein *et al*., 1994). However, this innovation was cut short when the state re-vamped their testing program. To date, it does not appear that any other state or assessment system have attempted to integrate the evaluation of opportunity into the measurement of student content. This information, along with other information about the students would, however, appear to be helpful in explaining test results and it is recommended that future work be completed to identify how this might be accomplished.

Alignment and English Language Learners

If curriculum standards, performance standards, curriculum, and assessment are aligned properly, it appears that the results obtained from the assessment system *should* be able to be used with confidence to evaluate whether programs deliver necessary services effectively, whether educational agencies are educating students properly, and whether all students learn the content specified as important in the state or district's content standards. Since the test items are assumed to be representative samples of the knowledge and skills identified in the standards, and all students are assumed to have been exposed to grade-level curricula, it follows that the assessment results would be considered a valid and reliable measure of this content. In other words, the results are taken to be generalizable evaluations of student performance on the content standards (or a tested subset of content standards) for all test takers. However, as noted above, rarely do alignment evaluations attempt to determine whether alignment results actually extend to specific subpopulations of students because results tend to reflect the experiences of the majority group or an artificial average across minority subgroups.

The relationship between systematic error and alignment is related to a careful evaluation of the conditions and testing materials for all populations, including ELLs. Specifically, there can be a substantial effect of systematic error over items and over persons for many English language learners, or any population where systematic errors have been documented (Haladyna and Downing, 2004). Systematic errors are errors that do not occur randomly or by chance, but which are sustained over cases and caused by non-targeted elements in the testing system which consistently impacts an individual or subgroup in a particular way. Unlike random error which balances itself out over items or students, small errors of this type will compound when items are aggregated to form test scores or when test scores are aggregated for students who share the same sets of qualities that produce the systematic error. This can lead to gross misrepresentation about the ability of these students, which, in turn, has implications for school accountability and for placement or other cases where educators use individual student scores.

Several chapters deal with ways to minimize this type of error but there needs to be evaluations and empirical evidence at several levels to determine whether the conditions and testing materials appear to be sufficient and suitable for the range of students who are taking the test. Specifically, in order to determine whether alignment is adequate for English language learners, evaluators need to determine whether adaptations made to the assessment system in terms of flexible tools, response options, administration options, and forms are appropriate to address the students needs. Further, the alignment between these forms, tools, and options and the standard assessment system needs to be evaluated as well as their relationship to the content standards and achievement levels.

Finally, it is important that the achievement levels fairly and comparably relate to the content standards and the assessment options that these students will be receiving and that there is no language in the achievement levels that might inadvertently bias attainment of any level. For instance, developers or agencies should be careful to allow for responses that have a minimal language load, as appropriate, for some ELLs, and attainment of any level should not assume or necessitate language except when this is the target of the composite score inference. To ensure the use of appropriate descriptors, evaluators looking at alignment should find evidence that the achievement level descriptors do not unduly bar English language learners with language or other challenges from being classified correctly. Levels in earlier systems sometimes required students to demonstrate performance through a restricted set of communication methods not related to the content areas (see Kopriva, 2000, for examples). While test development and implementation interventions should tend to minimize this problem because they allow multiple avenues of access, sometimes the use of multiple pieces in the assessment or accountability system reintroduces the problem. Procedures for establishing the levels themselves should also be evaluated as well as evidence of due process for non-mainstream students. Specifically, evaluators should have documentation of the technical rigor associated with the classification procedures themselves. What evidence is there to document that the cutoff scores are appropriate for ELLs or that multiple sources have been combined correctly? In situations where test forms, testing tools, and testing procedures appropriate for ELLs, as well as sufficiently rigorous matching procedures or appropriate validity documentation for ELLs are still being put in place, how are agencies determining whether the test scores or other pieces of information used to classify student achievement are representing the content-area mastery of ELL students properly? When incomplete evidence related to these sources is found, in many situations, other sources of achievement evidence should be used to support classification, either initially or upon appeal.

Concerns related to the alignment of assessments and content standards are raised and addressed throughout this book. Evaluators looking at alignment need to focus on whether appropriate evidence exists to determine that items and forms are accessible. Is there adequate evidence that the items and forms, test implementation, scoring, and reporting allow English language learners to demonstrate their achievement properly? The implementation of a reasonable number of the points outlined here can be used to provide evidence of alignment between the assessment system and content standards for this population. Such evidence should include the following:

- documentation of procedures related to item and forms development;
- documentation of matching, or how agencies determined which students got which forms and accommodations and why;
- rubrics, rubric notes, and scoring training that demonstrate that the content of students' constructed responses is not inappropriately confounded with language mastery where English proficiency is not the targeted construct;
- evidence that achievement levels allow for students with different levels of language to attain proficiency as warranted in the test score inference;
- results of analyses that show that force-choice responses, correct and incorrect, are not functioning differentially across ELL subgroups and the mainstream population; and
- results of other procedures and empirical analyses that determine that this population is understanding, using, and demonstrating test content properly.

In addition to these, it is particularly important for this population that there be some sort of evidence that ELL students are receiving academic instruction, in addition to the instruction in English. While opportunity to learn is not commonplace evidence across the country in the mainstream testing systems, there is precedence that, historically, certain programs for English learners have been focused entirely or more heavily on the learning of English to the exclusion of keeping up with grade-appropriate content. As such, it is important that academic opportunity be documented to ensure that this source of systematic error is minimized. Several authors (for instance Wong-Fillmore, 2006) emphasize how English can and should be taught within the context of academic instruction and these may be used as a model for evaluators to follow. However the documentation is accomplished, it is recommended here that information about their programs of study be included in any alignment evaluation process whenever ELLs are part of the student groups being tested.