VALIDITY EVIDENCE

In his extensive essay on test validity, Messick (1989) defined validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment” (p. 13). Four essential components of assessment systems need to be considered in making a validity argument: content coverage, response processes, internal structure, and relations to external variables. To put it simply, validity is a judgment about the degree to which each of these components is clearly defined and adequately implemented. Validity is a unitary concept with multifaceted processes of reasoning about a desired interpretation of test scores and subsequent uses of these test scores. In this process, we seek answers to two important questions. Whether the students tested have disabilities or not, the questions are identical: (1) How valid is our interpretation of a student's test score? and (2) How valid is it to use these scores in an accountability system to make judgments about students’ performance as it relates to a set of content standards?

Validity evidence may be documented at both the item and total test levels. This paper focuses only on documentation of validity evidence at the total test level. At this level, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) calls for evidence of the type noted above regarding content coverage, response processes, internal structure, and relations to other variables. Examples for each source of validity evidence are provided using illustrations from some large-scale assessment programs’ technical documentation. We subsequently provide a discussion of each of these sources.

Evidence of Content Coverage

In part, evidence of content coverage is based on judgments about “the adequacy with which the test content represents the content domain” (AERA et al., 1999, p. 11). As a whole, the test comprises sets of items that sample student performance on the intended domains. The expectation is that the items cover the full range of intended domains and that there are a sufficient number of items so that scores credibly represent student knowledge and skills in those areas. Without a sufficient number of items, a potential threat to the validity of the construct exists because the construct may be underrepresented (Messick, 1989).

Once the purpose of the test and intended constructs are determined, test blueprints and specifications serve as the foundation of validity evidence for determining the extent to which the test provides sufficient content coverage. Among other things, *Standards for Educational and Psychological Testing* (AERA et al., 1999) suggests that specifications should "define the content of the test, the number of items on the test, and the formats of those items" (Standard 3.3, p. 43). Test blueprints that include this information often are released to educators (and the public) via test manuals. Table 1 provides an example of a South Carolina test blueprint for fifth-grade mathematics on a fixed format (pencil and paper) given with or without accommodations for the regular assessment.

What might test blueprints look like for alternate assessments? Those based on standardized tasks or items may include information similar to the regular assessment. For example, administration manuals for Texas's *State-Developed Alternate Assessment II*, a pencil and paper test, include test blueprints for reading, writing and math by instructional level (grade or grade band). The blueprint for each subject indicates the number of items used to assess each objective in the *Texas Essential Knowledge and Skills* (Texas Education Agency, 2004). The manual briefly describes item formats elsewhere. The educator manual for Massachusetts' portfolio-based alternate assessment, an example with less standardized tasks or items, instructs teachers to include portfolio evidence that relates to specific grade-level standards within the state's curriculum framework (Massachusetts Department of Education, 2004). Teachers have some flexibility in determining the specific standards for which evidence is provided (e.g., selecting three of the five possible standards). Maximizing the specificity and clarity of instructions to teachers may help standardize the type of evidence for alternate assessments in general and portfolios in particular, thus allowing for more consistent interpretations of scores (as would be desired in large-scale assessment programs).

Table 1

Palmetto Achievement Challenge Tests (PACT) Blueprint for Grade Five Mathematics (South Carolina Department of Education, n.d.)

Distribution of Items

Type of Item	Constructed Response	Multiple Choice	Total on Test
Number of Items	4–5	32–34	**35–39 items**
Value of Each Item	2–4	1	
Total on Test	**11–13 points**	**32–34 points**	**45 points**

Distribution of Points by Strand

Strand	Constructed Response Points	Multiple Choice Points	Points per Strand
Numbers & Operations	At least one constructed response item in at least 4 of the 5 areas for a total of:	The balance of the points will be multiple choice and distributed through the strands according to the total points listed in the next column.	11–13
Algebra			7–9
Geometry			8–10
Measurement			8–10
Data Analysis & Probability			7–9
TOTALS	**11–13 points**	**32–34 points**	**45**

When performance assessments such as checklists, rating scales, or portfolios are used, student test scores are based on fewer items than for regular assessments (and therefore on a smaller sample of the intended domain). In these cases, it is especially important that the assessment items or pieces of portfolio evidence be representative of the intended domain (AERA et al., 1999, Standard 4.2) and also specified in the assessment administration instructions. Otherwise, construct underrepresentation is likely to be a threat to validity.

Evidence related to content coverage may become increasingly complex and nuanced, as state assessment systems use universal design principles to shift from a finite series of discrete

testing options to a broader continuum consistent with the cascade of options. As such, tables of test specifications may need to be tailored to groups of students who have multiple paths of participation in the cascade of assessment options. These test specifications may need to explicitly describe the populations of students for whom the test is intended as well as their selection criteria. In that case, high-quality items will serve as a foundation for content-related validity evidence at the assessment level. This topic represents an area in which considerable empirical evidence is needed.

Methods for Examining Content Coverage

The *Standards for Educational and Psychological Testing* (AERA et al., 1999) and other resources (see Downing & Haladyna, 2006) provide overviews of procedures for examining the match between test specifications and the items that comprise the test. The procedures that test developers follow in specifying and generating test items should be well documented (Standard 1.6). In large-scale assessment systems, where multiple forms of each test are needed, item writing and reviewing may continue on an ongoing basis. Trained item reviewers with content area expertise may be given a set of items and asked to indicate which standard the item matches. Comparisons would then be made between item reviewers' ratings and the item classifications initially made by the test developer. A critical consideration in collecting evidence related to content coverage is the standards upon which the assessment is based. The most tenable judgments of content match will be made if items are compared to specific objectives under a content standard (i.e., the smallest unit of measurement, or "grain size"), whether this match is for the grade-level, modified, or alternate achievement standards. The items in all forms of the alternate assessment need to be aligned with state grade-level content standards.

Evidence of content coverage also may come from methods for investigating the alignment of standards and assessments. For instance, Webb's (1999) alignment criteria consider not only alignment at the item level but also statistical indicators about the degree to which the test as a whole matches the standards. The **range of knowledge** correspondence criterion indicates the number of objectives within the content standard with at least one related assessment item; an acceptable range of knowledge exists when at least 50 percent of the objectives for a standard had one or more assessment items (which Webb reported as varying from 0 percent to 100 percent). Resnick, Rothman, Slattery, and Vranek's (2003) study of regular assessments in five states used the test blueprint as the basis for the intended content standards and calculated a proportional range statistic similar to that of Webb. They found that the regular assessments in

the five states that were examined tended to have lower-range scores than the criterion they established as acceptable. The research findings suggest that, as part of their continuous improvement efforts, states should pay particular attention to ensuring that the range of knowledge correspondence for regular assessments falls within acceptable ranges.

The range of knowledge indicator also has been applied to alternate assessments. In an analysis of two states' alternate assessments, Flowers, Browder, and Ahlgrim-Delzell (2006) found very low levels of range of knowledge match (0 percent to 37 percent of standards met the criterion). In contrast, Roach, Elliott, and Webb (2005) found that the majority of standards on one state's alternate assessments in reading, language arts, math, science, and social studies met or weakly met the criterion for acceptable range of knowledge. For a variety of reasons, meeting an acceptable range of knowledge correspondence is particularly challenging for alternate assessments. For example, unlike the discrete multiple-choice items on regular assessments, the embedded items that make up portfolio assessments do not readily lend themselves to discrete matching across content objectives. As such, more evidence is needed to establish criteria for acceptable range of knowledge correspondence for alternate assessments, as discussed below.

While Webb's (1999) range of knowledge criterion indicates the extent to which items match specific objectives within a standard, the **Balance of Representation** criterion indicates the degree of proportionality with which assessment items are distributed across objectives within a standard. In other words, the criterion indicates the extent to which items are **evenly distributed** across objectives. Balance of Representation is a calculated index, with 1 indicating a perfect balance and values closer to zero indicating most items were aligned with only a few objectives within the standard. (See Flowers, Browder, Ahlgrim-Delzell, & Spooner, 2006, or Tindal, 2005, for the Balance of Representation formula and calculation instructions.) In an analysis of four states' mathematics and science assessments, Webb used .7 as the criterion for an acceptable balance, and found a fairly high percentage of standards (71 percent to 100 percent across states and tests) met that criterion. Applying this criterion to one state's alternate assessment, Roach et al. (2005) found that all of the standards had acceptable results for Balance of Representation. In contrast, Flowers et al. (2006) found that only one of three states' language arts alternate assessments had any standards that met Webb's criterion for acceptable or weak balance of representation. The other assessments had 0 percent acceptable balance of representation across the standards. Resnick et al. (2003) also

investigated balance using qualitative judgments about "the relative importance of test items [given] to content and skills" (p. 20) in comparison with the importance stated in the standards. Based on responses to a set of open-ended questions, item reviewers rated the balance of sets of items on a scale (good, appropriate, fair or poor). Raters found that very few standards were rated appropriate or higher on language arts and math tests at the elementary, middle and high school levels. These findings may be partly attributed to limitations of providing an adequate sample of target skills from which to make inferences regarding the distribution of assessment items across content standards, particularly when the content standards are described at the granular level (i.e., the content standard includes multiple objectives that target discrete skills) needed to provide "entry points" for students with the most significant cognitive disabilities.

Criteria for Evidence Related to Content Coverage

Regardless of the type of assessment or the achievement standards upon which it is based, similar types of information should be included in test specifications in order to evaluate validity evidence related to content coverage. Standard 3.3 provides a comprehensive list of contents (AERA et al., 1999, p. 43):

> The test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information.

While the same kinds of information should be provided across assessments, the criteria established to judge content coverage as acceptable (e.g., sufficient number of items per standard) may vary. For example, the more partitioned a state's content standards are, the more difficult it may be for assessments to meet Webb's (1999) suggested Range of Knowledge criterion; using only that criterion as a target for "good" coverage may result in tests that are prohibitively long. The level of specificity of the content standards (e.g., global standards vs. specific grade-level objectives) against which items are aligned also may influence the statistics that are obtained on content coverage. Not until states develop evidence related to content coverage and consistently share that information with the psychometric and special education communities will a sufficient body of evidence exist to establish criteria for what is considered

"acceptable" content coverage across the range of assessment options in large-scale assessment systems.

Evidence of Response Processes

Considering again the range of assessment formats available for students with disabilities, response processes represent the cognitive behaviors required to respond to an item. In constructed response items, such as those seen on many regular assessments, the student's response process is the primary focus and intended to reflect a range of cognitive dimensions. However, these same assessments also require judging teachers (e.g., in assembling a portfolio or completing a rating scale), or raters (e.g., of extended response items, performance assessments or checklists), because an understanding of a person's response process also contributes to the validity evidence for test score inferences. When statements about examinees' or raters' cognitive processes are included in validity arguments, evidence about those processes should be described (Standard 1.8). The sections below will guide the collection of such evidence.

Item Specification and Review Procedures

While test blueprints described in the previous section are based on content and item type, tables of specifications also may be developed by forming a matrix that includes content and cognitive demand. Each cell in the table shows the number of items used to assess each topic or strand at each level representative of a cognitive demand. States use various frameworks for describing cognitive demand, each with a different number of levels and accompanying descriptors. For instance, the blueprint for South Carolina's PACT tests in science describes six levels of cognitive demand (South Carolina Department of Education, 2001). Webb's (1999) alignment criterion for depth of knowledge includes four levels. Haladyna, Downing and Rodriguez (2002) highlight the need for a taxonomy to classify items on the basis of cognitive demand that is empirically based and more universally used. While this suggestion was made in the context of classroom assessment, a well-founded taxonomy would also be useful for validity studies of large-scale assessments.

Response Processes of the Student

The processes of students' responses to most regular assessment items (selected response, short constructed response) may be considered through typical item review procedures. Test developers may start with a table of specifications, and external reviewers' judgments about

content and cognitive demand for each item may be compared with what was intended by the test developer. As with any other type of expert review process, reviewers should be well trained, and their selection, expertise, training and rating procedures should be thoroughly documented (Standard 1.7).

While expert ratings of cognitive demand provide some indication of students' use of lower- and higher-order thinking skills on the assessment, they do not rule out the possibility that the student's response is based on something other than what was intended. Haladyna (1999) describes a range of procedures designed to determine students' cognitive processes when responding to paper and pencil test items. One commonly used method is a think-aloud procedure, in which students orally describe everything they think about while working through a problem.

Cognitive processes may be more difficult to assess directly for alternate assessments aligned to alternate achievement standards. Think-aloud procedures are not feasible when students cannot communicate verbally, and interview methods — even with the benefit of assistive devices for students who communicate nonverbally — may not work for this population of students. In these cases, direct observational data collection methods may be needed to document possible cognitive processes. Administration of performance assessments may be videotaped, for example, and analyzed for behavioral evidence of particular processes. Videotaped accounts of instructional activities that yield products included in a portfolio also may provide supportive evidence of the student's response processes. Obviously, such resource-intensive data collection methods would be appropriate for empirical studies about the assessment rather than a component of the assessment given to all students.

One unique approach to standardizing the cognitive demand in alternate assessment is seen in Pennsylvania's Alternate System of Assessment (PASA Project, 2003), in which a series of otherwise standardized performance tasks are administered in one of three ways, depending upon the level of cognitive demand the teacher determines to be appropriate for a student. Level A is designed for the most simple response processes, usually providing a very simple discrimination context and requiring few steps; Level B takes the problem to a slightly more complex response by providing a more rich context (elements of a problem) and requiring more steps; finally, Level C is the most complex response with the most context provided and the most extended response. In order to encourage high expectations for student performance, the

level of cognitive demand is incorporated into the student's score.

Response Processes of Teachers and Raters

When someone other than the student is partially responsible for the responses to an assessment, that person's cognitive behavior also exerts some influence on the student's score. Studies on one state's portfolio-based alternate assessment support this assumption, as teacher training and understanding of the portfolio scoring system was associated with strong assessment scores (Browder, Karvonen, Davis, Fallin, & Courtade-Little, 2005; Karvonen, Flowers, Browder, Wakeman, & Algozzine, in press). When a teacher administers a performance assessment, completes a checklist, or assembles a portfolio, the teacher's exact responsibilities should be delineated in the assessment's specifications. The influence of judges and raters attenuates measurement reliability, which in turn impacts validity.

When assessment administration procedures are more standardized, there is less opportunity for the teacher's cognitive processes to influence the student's score. Clearly written, easy-to-follow scripts may help teachers adhere to a prescribed sequence of minimally intrusive prompting on performance tasks. Another basic method to maximize standardization is to have a neutral monitor present during the administration of performance tasks (South Carolina Department of Education, 2005).

As they are used with students, think-aloud procedures also may be used with teachers and raters. Teachers may be interviewed about the administration of a performance assessment, or asked to think aloud about completing a checklist or rating scale or about compiling a portfolio-based alternate assessment. Think-aloud procedures may be used to examine how raters follow the prescribed rating process, and possible places where failure to follow the process raises questions about validity of the score interpretation (Heller, Sheingold, & Myford, 1998). Post-hoc interrater agreement indices also may be used to assure consistency in rating a performance or scoring a product.

Criteria Related to Evidence Based on Response Processes

Considerations in evaluating criteria related to response processes are similar to those for content coverage. While well-established methods exist to collect this evidence, there is not yet a sufficient body of empirical evidence to provide clear standards for what is acceptable. Early studies on alignment (Flowers et al., 2006; Resnick et al., 2003; Roach et al., 2005; Webb,

1999) provide some possible benchmarks but should not be considered gold standards at this point. States should be aware of methodological problems in considering these sources of evidence that might attenuate the statistics obtained (e.g., judging the alignment on Webb's depth of knowledge criterion when comparing specific assessment items to vaguely worded content standards).

Evidence of the Internal Structure of the Test

The internal structure of assessment instruments reflects the dimensionality of the score or the degree to which the outcome can be explained by the format of the problem (e.g. selected response multiple choice versus constructed response short answers). The dimensionality of the problem has direct influence on the degree or clarity with which a construct is defined or can be inferred from the score. The *Standards for Educational and Psychological Testing* (AERA et al., 1999, pp. 13–15) call for a study on internal structure as part of test validation. For valid test score interpretations and validity generalization, it is expected that (1) the items show some level of internal consistency (Standard 1.11); (2) the internal structure of the test remains stable across major reporting groups (p.15); and (3) the internal structure of the test remains stable across alternate (and equivalent) forms of the same test (pp. 51–52). This section addresses these three topics.

Inter-Item Consistency

With an assessment consisting of items measuring the same construct or tapping the content standards in the same subject area (like reading or mathematics), it is expected that these items show some level of consistency among themselves. In other words, it is desirable that student responses to various items or parts of the test be logically related and not contradict each other in any substantial way.

Internal consistency can be checked by looking at the inter-item correlation matrix. It is desirable that all inter-item correlations be positive. Internal consistency is also manifested in overall test reliability indices such as KR20 or Cronbach's alpha. Phillips (2000) reported that a coefficient alpha of at least 0.85 was generally considered by most assessment experts as adequate for standardized tests such as Texas' high school graduation test. It should be pointed out, however, that the performance variability among students taking an alternate assessment is typically not as large as the performance variability among students taking the regular assessment. Therefore, the coefficient alpha for a modified assessment may be lower than the

0.85 threshold.

Inter-Strand Consistency

An assessment (based on grade-level, modified, or alternate achievement standards) may be designed to measure knowledge and skills in a number of separate content strands. In mathematics, for example, the strands may range from simple computations (operations) to more complex topics such as probability and statistics. In other domains like the assessment of writing, a strand may be one aspect of writing (such as expository or persuasive) and various strand scores are sometimes obtained by scoring the same student response using different rubrics. Because these separate strands are parts of a larger construct, it is desirable that they be somewhat related to each other (but not exceedingly so). Inter-strand correlations are expected to be positive and moderate. High inter-strand correlations are not desirable because the strands may essentially reflect very similar types of skills or abilities and may be overly redundant. Table 2 provides an illustration of data on between-strand correlations for the South Carolina PACT assessments in grade 8 mathematics in 2003 (South Carolina Department of Education, 2003).

Table 2

Inter-Strand Correlation Matrix for PACT Grade 8 Mathematics Assessments

MATHEMATICS					
Strand	**Number of Students**	**Mean Score per Strand**	**Standard Deviation**	**Minimum Points/Items per Strand**	**Maximum Points/Items per Strand**
Number & Operations	50,070	8.901	3.849	0	17
Algebra	50,070	10.832	4.113	0	18
Geometry	50,070	7.238	2.901	0	16
Measurement	50,070	2.909	2.168	0	8
Data Analysis & Probability	50,070	5.158	2.504	0	13
Between-Strand Pearson Correlation Coefficient Matrix					
	Number & Operations	Algebra	Geometry	Measurement	Data Analysis & Probability
Number & Operations	1.000	0.764	0.655	0.665	0.661
Algebra	0.764	1.000	0.654	0.644	0.661
Geometry	0.655	0.654	1.000	0.608	0.607
Measurement	0.665	0.644	0.608	1.000	0.606
Data Analysis & Probability	0.661	0.661	0.607	0.606	1.000

Unidimensionality

It is of considerable importance to check the number of dimensions (constructs) as operationally measured by the assessment and then reflected in the test data. Subject areas such as reading, mathematics and science (at the lower grade levels) are typically thought of as single constructs; however, student performance on the assessment items may be contingent on other unintended and irrelevant factors. Performance on constructed response items in mathematics, for example, may be dependent partially on reading level. Likewise, to solve a science problem,

reading and writing skills as well as some knowledge of mathematics may be essential.

Assessments based on alternate and modified achievement standards typically are given to students with a wide range of disabilities and academic training, so the internal structure of the assessment instrument may be more complex than it is for typical students assessed on grade-level achievement standards under standardized conditions. Performance tasks administered with scaffolded assistance may reveal both the ability of the student and the assessor's judgment regarding the level of prompting needed by the student in order to demonstrate his or her knowledge, which brings an extra dimension to be considered in any interpretation of the test data.

Interpretation of test scores is more straightforward when the assessment taps into one unique construct like reading or mathematics. Unidimensionality may be checked via a variety of statistical methods ranging from classical principal component analysis (PCA) or factor analysis (FA) techniques to more modern methods based on item response theory (IRT). The scree test, for example, in PCA and FA has been used quite often in this regard. Other IRT techniques include the DIMTEST and similar procedures developed by Stout and his associates (Stout, 1987; Nandakumar & Stout, 1993).

Similarity of Factor Structure and Differential Item Functioning Across Accommodated and Nonaccommodated Conditions

Many authors, such as Geisinger (1994), have noted major measurement issues when students are assessed by standardized tests that have been administered under nonstandard conditions. By adhering to standard procedures, errors in scores can more likely be attributed to random or individual errors, rather than systematic administrative errors, and scores can be interpreted similarly for all participants (Geisinger, 1994). Some students with disabilities require accommodations to allow an assessment to tap into a student's ability on the construct(s) being measured, curtailing or neutralizing the effect of a student's disability on his or her test result. A true accommodation should allow a student to be assessed in such a way that a disability does not misrepresent the student's actual level of proficiency. Students ideally should be placed on equal footing and not advantaged or disadvantaged because of a disability, and the interpretations of their scores should be valid for the purpose of the test.

In her landmark writing on the balance between the individual rights of the student with a

disability and the integrity of the testing program, Phillips (1994, p. 104) indicated the need to address a number of questions including the following two: (1) Does accommodation change the skill being measured? and (2) Does accommodation change the meaning of the resulting scores?

A host of psychometric issues need to be considered in addressing these two questions. Willingham (1989) pointed out that in order to achieve score comparability, various forms of the assessment need to display similar factor structures, and no differential item functioning (DIF) should exist across student groups. In a long review of psychometric, legal, and social policy issues about test accommodations for examinees with disabilities, Pitoniak and Royer (2001) concluded that further legal decisions would determine assessment accommodation policies and that more research is needed on test comparability.

The PCA, FA, or IRT techniques previously described can analyze factor structures. DIF analysis may be performed on the quality of test items. It should be mentioned that evidence regarding similarity in factor structure and DIF is more easily collected when exactly the same assessment is administered under varying conditions, as separate test forms. This is not usually the case where the accommodated form is the same as the regular form. More often than not there are significantly fewer numbers of accommodated versus nonaccommodated students, and still fewer when breaking out analyses by accommodation, limiting the appropriateness of conducting PCA, FA, IRT, or DIF analyses. In addition, most students receive multiple accommodations in a given administration.

However, when accommodated and nonaccommodated forms are developed and constructed to reflect exactly the same strands (content standards), PCA or FA may be carried out at the strand level rather than at the item level. This type of analysis was used in two studies based on the South Carolina High School Exit Examination (Huynh, Meyer, & Gallant-Taylor, 2004; Huynh & Barton, 2006). An analysis of the similarity of inter-item and inter-strand correlations may also provide evidence of the similarity across various groups of the construct assessed by the instruments.

Similarity of Factor Structure and Differential Item Functioning Across Major Reporting Groups

When data are available, it may be of interest to determine whether a test's factor structure is

similar across important reporting groups such as students with varying disabilities. An example of this type of analysis may be found in Huynh and Barton (2006), who compared the factor structure of the same test across groups of students with physical, learning and emotional disabilities. It is noted that this type of evidence may be hard to compile with small reporting groups. In this case, perhaps a consensus agreement via "juried" assessment, like the one used in Oregon, may be a good choice (Oregon Department of Education, 2004).

Special Considerations for Tests With Cross-Grade Items: Internal Structure and Differential Item Functioning

The U.S. Department of Education recently published a Notice of Proposed Rulemaking (NPRM) in the *Federal Register* that would allow states to develop modified achievement standards and use assessments aligned with those modified standards for a group of students with disabilities who can make progress toward, but may not reach, grade-level achievement standards in the same time frame as other students.[1] An assessment based on modified achievement standards may be narrower in content breadth and coverage than an assessment based on grade-level achievement standards. An assessment based on modified achievement standards may comprise a number of **cross-grade** items that were originally designed for students at other grade levels. Cross-grade items may be used under certain conditions to enlarge the item bank for the purpose of test form construction.

In some cases it may be possible to find a section of the assessment based on grade-level achievement standards (GLAS) that is similar to the assessment based on modified achievement standards (MAS) either at the item level (all items are identical) or at the strand level (all strands are the same). In these cases, it may be possible to check the similarity between the GLAS section and the MAS section in terms of factor structure. When enough data are available, it may be possible to check for DIF between students who took the GLAS section and those who took the test with MAS.

Similarity in Types of Errors Made by Students

Evidence of similarity of constructs across major reporting groups may also be collected by analyzing the major types of errors made by students on various assessments administered to these groups. A study along these lines for the South Carolina High School Exit Examination was reported in Barton and Huynh (2003). Overall, the question is, "Are the types of errors

[1] Retrieved from the World Wide Web on Feb. 8, 2006 at http://www.ed.gov/legislation/FedRegister/proprule/2005-4/121505a.html

made by students taking the regular assessment with accommodations or students taking the test based on modified achievement standards similar to those of general education test takers at a similar grade level, level of mastery, etc.?" Similarity in the errors is a good indicator that the assessment instruments tap similar constructs.

Evidence Based on Relations to External Variables

The *Standards for Educational and Psychological Testing* (AERA et al., 1999, pp. 13–15) also calls for validity evidence based on relations to other variables. External evidence for the construct being measured may be found in the relationship between the test and other similar or dissimilar measures. Evidence of this type is sometimes referred to as "convergent" and "divergent," respectively. Applying this type of validity evidence to large-scale assessments in reading and math should result in reading assessment scores that are more closely related to other reading scores than to math scores, and math assessment scores that are more closely related to other math scores than to reading scores.

For example, Table 3 reports correlational data on the relationship between Palmetto Achievement Challenge Tests (PACT) scores and scores on the TerraNova tests for grades three and six. The table is extracted from the 1999 PACT Technical Documentation written by Huynh, Meyer, and Barton (2000). In this table, PACT scores are scale scores and TerraNova subtest scores are normal curve equivalents. The data indicate that PACT English Language Arts (ELA) assessments relate more strongly to the TerraNova reading and language components than to the TerraNova mathematics component. Similarly, PACT mathematics assessments relate more strongly to the TerraNova mathematics component than to the other two TerraNova components. However, the divergent evidence for validity is stronger for PACT mathematics assessments than for ELA assessments.

Table 3

Correlation Between the PACT and the TerraNova for Grades Three and Six

Grade/Content	TerraNova		
	Reading	Language	Math
Grade 3 ELA	.776	.768	.729
Grade 3 math	.689	.687	.803
Grade 6 ELA	.755	.754	.741
Grade 6 math	.710	.705	.863

Conclusions and Recommendations

Together with the information about validity evidence at the item level, this paper provides the reader with a set of methods for collecting evidence to evaluate the validity of score interpretations in large-scale assessment systems. The good news is that many of the techniques for examining evidence related to content coverage, response processes, internal structure, and relationships between test scores and other variables, are well established in the psychometric literature. The challenge will come in determining what constitutes “good” validity evidence using each of the techniques described in this paper. At this time, there is not a sufficient body of theoretical and empirical evidence to recommend minimally acceptable values of statistical indicators for all of the sources of validity evidence. Especially for some assessments that are yet to be designed (e.g., assessments based on modified achievement standards) and for assessments that may be taken by small groups of students, new small-sample techniques for gathering and assessing validity evidence are needed. Following are some recommendations for collecting validity evidence:

1. States need to collect and document validity evidence for all four general areas described in this paper: content coverage, response processes, internal structure, and relations to other variables. Even if strong validity evidence is collected in one or two areas, the failure to collect evidence across all four areas will weaken arguments that may be made for the validity of score interpretations.
2. Validity evidence for content coverage and response processes should be collected

with the greatest precision possible. Blueprints for assessments should clearly state the strands or sub-strands within each content area, the types of items used to assess those strands, and the levels of cognitive demand required to respond to the items. Evidence of content coverage will be stronger if links can be made between specific achievement standards (grade level, modified, or alternate) and assessment items. Evidence collected through alignment methods that consider breadth and depth of content coverage will provide a richer understanding of content match than item-level judgments alone. States will need to carefully consider methods to assess response processes to rule out possible problems with construct-irrelevant variance. The response processes of teachers, raters of performance assessments, and students with the most significant cognitive disabilities who complete tasks for performance-based alternate assessments should be considered as well.

3. Evidence related to the internal structure of the assessment may include inter-item correlations, inter-strand correlations, summary data from test dimensionality, similarity of the test's internal structure across major reporting groups, and the nature of the errors major reporting groups made on the test.
4. Validity evidence collected via relationships between scores on the target assessment and external measures should include relationships with both similar measures (convergent evidence) and dissimilar measures (divergent evidence). Correlational analyses of scores on these measures may be used to make judgments about the quality of evidence from relationships with external measures.
5. It will continue to be of great importance to thoroughly document the backgrounds of students who are involved in the field-testing process (Standard 1.5). Disability labels alone may not be good proxies on which to base assumptions about group homogeneity. Careful documentation of testing accommodations will also be needed (Standard 1.13) for subsequent analysis of the potential unintended impact of accommodations (e.g., construct irrelevant variance).

In general, well-established data collection guidelines should be followed (Downing & Haladyna, 1997) and validity evidence should be clearly documented. The challenges associated with collecting new types of evidence should not discourage the continued study and collection of item- and test-level validity evidence for all types of assessments.

References

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Barton, K., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement, 63*, 602–614.

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). Taxonomy of educational objectives: The classification of educational goals. *Handbook I: Cognitive domain*. New York, Toronto: Longmans, Green.

Browder, D. M., Karvonen, M., Davis, S., Fallin, K., & Courtade-Little, G. (2005). The impact of teacher training on state alternate assessment scores. *Exceptional Children, 71,* 267–282.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61–82.

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook on test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fueuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Flowers, C., Browder, D., & Ahlgrim-Delzell, L. (2006). The alignment of three states' alternate assessments to state standards. *Exceptional Children, 72*, 201-215.

Flowers, C., Browder, D., Ahlgrim-Delzell, L., & Spooner, F. (2006). Promoting the alignment of curriculum, assessment, and instruction. In D. M. Browder & F. Spooner (Eds.), *Teaching reading, math, and science to students with significant cognitive disabilities*. Baltimore: Brookes.

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7,* 121–140.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309–334.

Heller, J. I., Sheingold, K., & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*, 5–40.

Huynh, H., & Barton, K. (2006). Performance of students with disabilities under regular and oral administrations for a high-stakes reading examination. *Applied Measurement in Education*, 19(1), 21–39.

Huynh, H., Meyer, P., & Barton, K. (2000). *Technical documentation for the 1999 Palmetto Achievement Challenge Tests of English Language Arts and Mathematics, Grades Three through Eight*. Columbia, SC: South Carolina Department of Education, Office of Assessment. Retrieved July 13, 2005, from http://www.myscschools.com/offices/assessment/Publications/Index_of_Technical_Reports.htm.

Huynh, H., Meyer, P., & Gallant-Taylor, D. (2004). Comparability of student performance between regular and oral administration for a mathematics test. *Applied Measurement in Education, 17*, 39–57.

Karvonen, M., Flowers, C. P., Browder, D. M., Wakeman, S., & Algozzine, B. (In press). A case study of the influences on alternate assessment outcomes for students with disabilities. *Education and Training in Developmental Disabilities*.

Massachusetts Department of Education. (2004). *2005 educator's manual for MCAS-Alt*. Malden, MA: Author. Retrieved June 25, 2005, from http://www.doe.mass.edu/mcas/alt/05edmanual.pdf.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.

Minnesota Department of Education. (2005, March). *Minnesota Comprehensive Assessments Series II: Test specifications for reading and mathematics*. Roseville, MN: Author. Retrieved July 5, 2005, from http://education.state.mn.us/content/087562.pdf.

Nandakumar, R., & Stout, W. (1993). Refinement of Stout's procedures for assessing unidimensionality. *Journal of Educational Statistics, 18*, 41–68.

Oregon Department of Education. (2004). *Juried assessment 2004–2005: Guidelines for using the juried assessment process and guidelines for jurying a modification*. Salem, OR: Author. Retrieved June 25, 2005, from http://www.ode.state.or.us/teachlearn/testing/admin/juried/juriedassmtmanual0405.pdf.

PASA Project. (2003). *Pennsylvania alternate system of assessment administrator's manual*. Pittsburgh, PA: Author. Retrieved June 25, 2005, from http://www.pattan.net/files/instruction/adminmanual.pdf.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7,* 93–120.

Phillips, S. E. (2000, April). Legal corner: GI Forum vs. TEA. *NCME Newsletter, 8*(2), n.p.

Pitoniak, M., & Royer, J. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*, 53–104.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003). Benchmarking and alignment of standards and testing. *Educational Assessment, 9*, 1–27.

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate assessment. *The Journal of Special Education, 38*, 218–231.

South Carolina Department of Education. (n.d.). *Blueprint construction of PACT in mathematics.* Columbia, SC: Author. Retrieved June 25, 2005, from http://www.myscschools.com/offices/assessment/Publications/PACTblueprints.htm.

South Carolina Department of Education. (2001). *PACT science assessment: A blueprint for success.* Columbia, SC: Author. Retrieved June 25, 2005, from http://www.myscschools.com/offices/assessment/Publications/PACTblueprints.htm.

South Carolina Department of Education. (2003). *Technical documentation for the 2003 Palmetto Achievement Challenge Tests of English Language Arts, Mathematics, Science, and Social Studies*. Retrieved July 13, 2005, from http://www.myscschools.com/offices/assessment/Publications/PACT-Tdoc03.doc.

South Carolina Department of Education. (2005). *High school assessment program alternate assessment (HSAP-Alt) test administration manual*. Columbia, SC: Author. Retrieved June 25, 2005, from http://www.myscschools.com/offices/assessment/Publications/HSAPAltTAM030805.pdf.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Texas Education Agency (2004, August). *State-Developed Alternative Assessment II information booklets*. Austin, TX: Author. Retrieved June 25, 2005, from http://www.tea.state.tx.us/student.assessment/resources/guides/sdaa/index.html.

Tindal, G. (2005). *Alignment of alternate assessments using the Webb system: An abbreviated version*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. (NISE Research Monograph No. 18). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education. (ERIC Document Reproduction Service No. ED440852).

Willingham, W. W. (1989). Standard testing conditions and standard score meaning for handicapped examinees*. Applied Measurement in Education, 2,* 97–103*.*

This document was produced in December 2005 under U.S. Department of Education Contract No. ED4CO0025/0002 with the American Institutes for Research. Renee Bradley served as the contracting officer's representative. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this report or on Web sites referred to in this report is intended or should be inferred.