VALIDATING ASSESSMENTS FOR STUDENTS WITH DISABILITIES

Validating assessments for students with disabilities can be challenging. If test scores derived from those assessments are used for state accountability purposes, threats to validity must be identified and actions taken to remove or reduce those threats. For example, a student with a vision impairment may have trouble reading a complex mathematics problem presented in small print on a single page. The ability of the test to accurately measure this student's performance on the task could be compromised. However, by changing the mode of presentation—for example, using a larger font—the effect of the vision impairment is removed, thus providing the student a fairer opportunity to perform. Appropriate accommodations in the design or administration of a test for students with disabilities may be necessary to improve the validity of the results.

The study of validity is greatly aided by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Experts in testing developed the *Standards* to help sponsors provide the highest quality testing programs possible.

Validity

States and school districts have a responsibility to educate students with disabilities. Part of this responsibility is validating the interpretations and uses of the students' test scores. This validation process needs to be comprehensive and include all components of the assessment system: the development of items and tasks, the training of teachers on implementation, the scoring and reporting procedures, and any number of statistical analyses that need to be conducted on the outcomes. In the end, the validation process focuses on the interpretation of the assessment results for making accountability decisions. The validation process is described here from an ideal perspective, with the understanding of the reality that all states and school districts have limited resources for validation.

What Does Validity Address and How Is It Operationalized?

Validity offers a way to reason about a desired interpretation of test scores and their subsequent use. The process followed to make valid interpretations and uses of test scores, *validation*, is used to answer two important questions:

(a) How valid is an interpretation of a set of test scores?

(b) How valid is the use of this set of test scores in an accountability system?

The questions are the same for all students, although the focus of this paper is on validation of assessments designed to be used in a state or school district with students who have disabilities.

The process of validation involves four stages: (a) defining what students learn, (b) stating the validity argument, (c) making the claim for validity, and (d) gathering evidence to support the argument and claim. In the final stage of the validation process, a qualified evaluator considers the logic of the argument and its plausibility, the claim, and the evidence in support of the claim. A summative judgment is made and recommendations are usually made for improving the testing program to strengthen supporting evidence and eliminate the threats to validity uncovered through the gathering of evidence.

Defining What Students Learn

The most important and fundamental step in any testing program to ensure a valid assessment of student learning is to define the academic content (American Educational Research Association et al., 1999, chap. 1). Understanding what students learn (or are supposed to learn) in school is fundamental to designing tests that help assess student learning. A basic education typically includes acquiring both knowledge and skills. A convenient way to think about achievement in this context is to imagine a *domain* of objectives that students learn to accomplish. In elementary mathematics, for example, this domain might include adding, subtracting, multiplying, and dividing whole numbers, fractions, and decimals. Other parts of this domain are linked to student learning objectives based on state content standards. A test score represents a level of achievement in that domain, for example, how many tasks students can perform or how well they can perform them. Levels of learning, for example beginning, intermediate, and advanced, can be set, based on the assumption that students go through natural stages of learning. For purposes of accountability, we tend to use terms to describe achievement levels such as the ones shown in Table 1.

Table 1

Academic Achievement Standards

Levels[1]
Advanced—Well above the minimum acceptable level of mastery of the material in the state's academic content standards
Proficient—Mastery of the material in the state's academic content standards
Basic—Progress of lower achieving students toward mastering the proficient and advanced levels of achievement

These levels are defined by a panel of subject matter experts whose experience with students should enable them to make valid determinations of the cut scores that separate achievement into these levels. Typically, the process for establishing these achievement standards includes a number of steps that provide some of the procedural evidence for validation and eventually results in an impact analysis that provides the statistical evidence for validation. Both types of evidence are further discussed below.

Traditional criterion-referenced and domain-referenced testing, which were popular in the past, featured tests designed to measure a student's status with regard to a large domain of knowledge and skills for which every bit of knowledge and every skill had a reference to a student learning objective. We could isolate the knowledge and skills well, and we could teach them effectively. The results of each achievement test were intended to be a representative sample of student learning from that domain. A test score would signify the level of achievement in the domain. For example, a score of 75 percent might lead to the conclusion that the student could perform proficiently on 75 percent of all items in that domain even though not all items had been presented to the student.

Most state content standards contain objectives that identify knowledge and skills that fit into this view of achievement. Table 2 shows examples of student learning objectives for three subjects that reflect knowledge and skills needed by all students to succeed in school.

[1]These terms are used in U.S. Department of Education, Standards and assessment peer review guidance, p. 2.

Table 2

Objectives for Knowledge and Skills

Subject Matter	Example of a Student Learning Objective
Reading	▪ Identify main characters in a story. ▪ Identify a cause and effect relationship in a story. ▪ Identify root words. ▪ Distinguish fact from opinion.
Writing	▪ Proofread a text. ▪ Place commas correctly. ▪ Use active voice as appropriate to purpose. ▪ Spell correctly.
Mathematics	▪ Identify examples of mathematics terms (e.g., logic, manipulative, pi, integer, scatter plot). ▪ Construct a bar graph with data provided. ▪ Order numbers from low to high.

For many reasons, the most desirable format for measuring knowledge and skills is multiple choice or selected response (Haladyna, 2004). The main reason is that the multiple-choice format provides the best chance for very good sampling from a domain, which usually allows tests to be more reliable. Other reasons include logistics and costs. Multiple choice tests are usually substantially less expensive than other formats. Nevertheless, for some important skills (e.g., reading skills such as phonemic awareness and reading fluency), multiple choice tests are not suitable. Instead, brief constructed tasks are necessary.

The Validity Argument

The validity argument states that some tests will produce scores that can be interpreted validly as measures of student achievement and used validly in a manner that is stated publicly (Kane, 1992). Making this argument involves assumptions about the causal connections between student learning and out-of-school factors, such as family background characteristics (socioeconomic status, mobility, etc.) and in-school factors, such as quality and quantity of instruction, learning environment, opportunity to learn, and school leadership. We also determine whether the student's disabilities interfere in some way with the validity of any assessment of his or her learning.

Part of the validity argument is that interpretations of test results for students with disabilities are valid if certain conditions are met satisfactorily. A student's disability should not interfere with the assessment of his or her learning. For instance, a student with mild mental retardation and low

reading comprehension may have difficulty reading a mathematics test that features items from the domain of complex mathematics problem-solving tasks. Administering an unaltered mathematics problem-solving test prevents the student from performing even if he or she has the ability to solve the problems. Addressing mathematical problem-solving in a manner that is more suitable for a student with this type of cognitive functioning might require simplification of the problem or its cognitive demand. In other words, accommodations are developed to remove factors that obscure a valid assessment of each student's learning.

Abedi (2004) presents examples of linguistic modifications of math test items in which, although used with English language learners, simplified language would provide similar advantages to students with disabilities. In simplifying linguistic features, Abedi removes idioms and words that are long or unfamiliar in context. Complex sentences are simplified by removing the passive voice and subordinate, conditional, and adverbial clauses. As Abedi, Hofstetter, & Lord (2004, p. 17) note, these changes narrow the performance gap of English language learners and other students by "modifying the language of the test items to reduce the use of low-frequency vocabulary and complex language structures that are incidental to the content knowledge being assessed."

Validity Claim

An argument about the validity of an interpretation based on a test of student achievement is subject to analysis, evaluation, and approval by state policymakers. This may be a public encounter in which all constituencies take part. Participation is one contributing form of validity evidence (Kane, 2002). After all is said and done, the sponsor of the tests—the state—needs to make the claim that the use of the test scores for students with disabilities is sufficiently valid. At the end of this validation process, the evidence should support the claim. Even though the validation is intended to support the argument and claim, the state has a duty to seek out evidence that may not support either the argument or claim. By doing this, the state shows that its work was done with integrity. In addition, identifying evidence that threatens validity provides critical data for recommending changes in the testing program that can potentially eliminate or reduce these threats. Validity research is a key to uncovering these threats to validity (Haladyna, in press).

Validity Evidence

Evidence collected in the validation process takes two major forms: (a) procedural and (b) statistical or empirical. Each is discussed in this paper. Assembling a complete body of

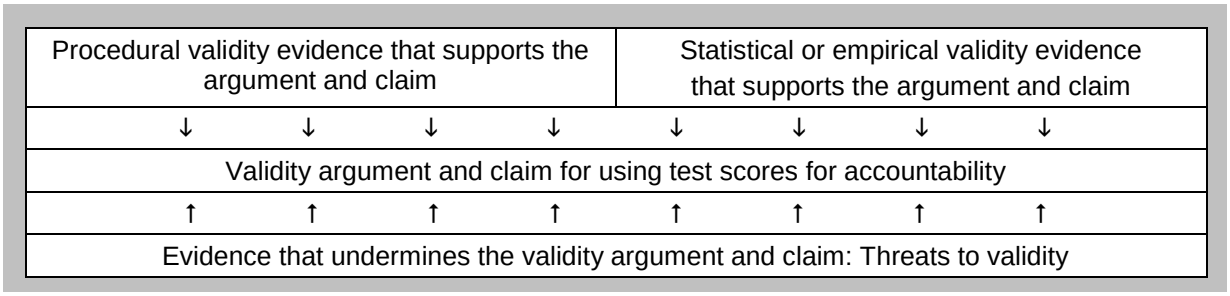
evidence can take many years. In fact, this process can be seen as unfolding in an evolutionary way over a long time. At the very least, collecting procedural evidence from item bias review panels takes a year to complete after pilot testing. Field-testing (in which differential item functioning analyses may be conducted) can easily extend this time beyond a year. And with alternate assessments, extensive professional development is needed to ensure that teachers follow proper procedures for gathering student work samples. Teachers must also learn to carefully score the portfolios or performance tasks to ensure the dependability and credibility of the work samples for making accountability decisions. All of this can take several years before any validity claims can be substantiated or refuted. Meanwhile, it is important to document the procedures used to train teachers and to develop items and tasks for use in the assessment, at the same time analyzing their performance on the items and tasks. Of course, if the validity claim focuses on attaining proficiency—as it is likely to do—then the process of setting standards also becomes part of the evidence. In this case, both procedural and statistical or empirical evidence will take at least two to three years to collect, analyze, and report. In summary, validity evidence is accumulative and needs to consider the process and outcomes, as they fit within the "construct definition" of the target behaviors, to provide both the theoretical support and the decisions for the construct definition.

The strength of the evidence supporting the claim for validity may increase each year unless threats to validity are uncovered and ignored. However, changes in the educational and political environments could actually decrease support. In any effective testing program, the process of validation includes reviewing the validity argument, considering the claim for validity, and evaluating the validity evidence to make a summative judgment about validity. After making this summative judgment, a formative process leads to recommendations for improving the testing program and, in so doing, increasing the validity evidence and the validity of the desired interpretation and use.

Figure 1 shows how the two kinds of validity evidence support the argument and claim and how some evidence also works in the opposite direction. Although evidence can support the argument claim, it also can threaten or detract from the validation argument and claim. The goal in validating assessments is to strengthen evidence that supports the argument and claim while minimizing or reducing evidence that disallows the argument and claim.

Figure 1

How Validity Evidence Works for and Against the Validity Argument and Claim



Some important observations about validity evidence can help states, school districts, and testing companies make better decisions about what kind of validity evidence to collect and why. Decisions about validity evidence are best made by first considering the purpose of the test—in this instance, to ensure state, district, and school accountability. We need to infer degrees of improvement in learning for groups of students.

Procedural validity evidence. Procedural validity evidence provides documentation that the assessment was developed and delivered in a way that is consistent with testing standards (American Educational Research Association et al., 1999). For example, technical specifications are important documents that help stakeholders understand how tasks and items were developed and from what domains they were sampled. Full discussions of how the tasks were reviewed and formatted or worded for universal design would help stakeholders understand the steps taken to develop actual administration and scoring procedures. Finally, documentation of the alignment and standard-setting processes would ensure that the achievement standards fully reflect appropriate content and levels of proficiency.

At the same time, the documentation provides evidence that the test content and administration are appropriate for each student with a disability. The procedural evidence arising from following the *Standards* (American Educational Research Association et al., 1999) is a valuable source of validity evidence. The *Handbook on Test Development* (Downing & Haladyna, in press) also offers advice about procedures that provide validity evidence.

It is important to document test scoring procedures. Post-test activities also are part of the evidence that supports the argument and claim for validity. The *Standards* (American Educational Research Association et al., 1999, chap. 6) are clear about the need for documentation of procedures and other actions as a type of validity evidence that supports a

specific test score interpretation or use. Haladyna (2002a) discusses the need for documentation with respect to developing a technical report, a major source of validity evidence.

Table 3 summarizes the bodies of evidence that can be used to characterize procedural evidence. Procedures and their documentation contribute greatly to the body of validity evidence supporting a state's policy on testing students with disabilities as well as the validity of interpreting and using these scores.

Table 3

Types of Procedural Validity Evidence

Type of Procedural Evidence	Description and Reference
Content-related evidence	The content of the test is well defined using a set of content standards. The structure of the content is known, unidimensional or multidimensional. The basis for identifying content is systematic (Messick, 1989, 1995a, 1995b; Kane, in press; Webb, in press). Implicit in this type of evidence is a precise definition of what is and is not intended for measurement.
Item quality	Items are developed following well-established principles and procedures, and documents provide evidence of this. Items can be field tested using nonquantitative techniques (e.g., think aloud) (Downing & Haladyna, 1997; Haladyna, 2004; Welch, in press).
Reliability	Because not all reliability coefficients are equally relevant for different tests, the appropriate type needs to be determined. Furthermore, the procedures for designing, implementing, and scoring the test must be documented to help interpret these coefficients. Reliability is needed before proceeding to validity claims, which then need to be independently established.
Scaling for comparability	When tests are modified, the resulting score should be on a scale that is validly interpretable. Procedures for achieving valid accommodations should be spelled out in documents.
Test design	Any test can be assembled using a variety of strategies that are based on differing statistical theories and principles. If tests are altered, the rationale for any alteration should be stated. These choices and actions should be well documented.
Test administration	▪ Test administrator's guide ▪ How accommodations are determined ▪ Procedures for test administration ▪ Reports of irregularities
Test scoring	▪ Scoring protocols ▪ Quality control
Standard setting	▪ Report of a standard-setting study
Reporting results	▪ Report of the development of scores reports (see Ryan, in press)
Security	▪ Security policies and procedures

Implicit in this list and the one below is that the population characteristics of the actual test takers match those of the intended target population in distributions of types of disabilities or other factors and whether an adequate range of format and altered testing conditions will be offered.

Procedural validity evidence typically appears in written form as a report, memorandum, letter, newsletter, brief, or file document. The document exists in an archive, and many states post

such documents on their Web pages as a means of documenting their validity evidence. Keeping a well-organized archive of these documents and a record of procedures is crucial to providing an adequate body of validity evidence for supporting the claim for validity.

Empirical or statistical validity evidence. This type of evidence complements procedural validity evidence. In addition to procedures that contribute to the body of validity evidence, we must provide data that support our inferences about what a test score means for a group of students with disabilities. Studies of reliability of scores, the structure of test data, the quality of items, and the equivalence or comparability of different tests (including those with accommodations and alternate assessments) are crucial to forming this body of empirical validity evidence. Haladyna (in press) discusses the types of validity studies that can be conducted and their bearing on the assembled body of validity evidence. Policymakers and others often identify problems that they believe threaten the validity of their testing program. Validity studies can be used to reveal the seriousness of the threat and recommend ways to reduce or eliminate each threat. Every state must engage in empirical studies to the greatest extent possible to assemble validity evidence that complements the procedural evidence.

Table 4 shows categories of empirical validity evidence. The entries in this table are suggestive, not exhaustive. Of the many types of empirical validity evidence that exist, most come naturally from standards described in the *Standards* (American Educational Research Association et al., 1999).

Table 4

Types of Empirical Validity Evidence

Types of Empirical Validity Evidence	Descriptions
Content-related evidence	▪ Analysis of item responses to identify dimensions ▪ Relationships with other (criterion) variables (e.g., formerly known as predictive validity and concurrent validity) ▪ Correlations to like and unlike variables (convergent and discriminant validity evidence)
Item quality	▪ Item analysis ▪ Ratings of test items based on subject-matter experts ▪ Differential item functioning
Reliability	▪ Estimates of reliability for individuals or groups ▪ Estimates of standard errors around cut scores
Scaling for comparability	▪ If new scales are produced for students with the most significant cognitive disabilities, the validity of the extended scale should be reported. ▪ Equating results
Test design	▪ Procedures for test design should produce tables of specifications with test information curves and differential item functioning as well as results from accommodations studies in which changes have been made.
Test scoring	▪ Descriptive statistics
Standard setting	▪ Results of a standard-setting study ▪ Impact study of a recommended cut score

As part of any validation, both procedural and empirical validity evidence should be assembled to support the argument and claim. No single source of evidence is sufficient; the mix of evidence is important to the evaluator called upon to make a summary judgment about validity. The next section calls for another type of validity evidence that is seldom considered in validation.

The Search for Negative Validity Evidence

As noted in the previous section, personnel for any testing program must identify procedural and empirical or statistical validity evidence that supports the validity argument and claim for validity, which is the main purpose of validation. However, as Cronbach (1987) observes, the claim for validity is stronger when challenges to validity have been investigated and dismissed. Kane, Crooks, and Cohen (1999) state that the chain of reasoning in validation is only as strong as its weakest link. Therefore, all testing programs should seek negative validity evidence with the objective either of not finding any or, when finding some, concluding that it is immaterial. When negative validity evidence is material, the threat to validity must be considered and resolved.

The alternative is to reject the intended interpretation. Negative validity evidence is a serious problem when testing students with disabilities, so added scrutiny is needed. In part, negative validity evidence results from the exclusion of students with disabilities in large-scale assessment programs prior to the *Individuals with Disabilities Education Act* amendments of 1997. Lack of attention to negative validity evidence also reflects relatively recent efforts to understand test accommodations and their interaction with student characteristics.

Negative validity evidence provides counterarguments that refute criticisms of any validity arguments. For example, in developing simplified items and tasks that would ensure access to content, researchers could collect evidence of items or tasks that are not simplified (and that might predict poorer performance on those items or tasks for students who are not skilled readers). Furthermore, researchers could collect negative validity evidence of the skills of students for whom such accommodations are not usually recommended. In general, such accommodations are evaluated only with students for whom they are recommended and for whom no other objective measure for documenting their reading skills is available.

Two major types of negative validity evidence exist: (a) construct misrepresentation or underrepresentation and (b) construct-irrelevant variance. When measuring more than is intended (due to construct-irrelevant factors), scores are typically inappropriately low. Conversely, if measuring less than intended (for example, through construct underrepresentation), inappropriately high scores are obtained. Of course, exceptions are possible.

***Construct misrepresentation or underrepresentation*.** Test accommodations are made to ensure that students with disabilities have appropriate access to the assessment. An accommodation is “a general term for any action taken in response to a determination that an individual’s disability or level of English language development requires a departure from established testing protocol” (Koenig & Bachman, 2004, p. 1). When a test accommodation is provided for a student with a disability, the most important question is whether the test result leads to a misrepresentation or underrepresentation of the intended content. If a student with a disability who requires a longer test administration time does not receive sufficient time to complete the test, his or her score will be underestimated and the test results will lead to an underrepresentation of the intended content. That is, the score will underestimate the student's proficiency (because the student's disability interfered with his or her ability to demonstrate knowledge of the content in the allotted time) and the test results will lead to an

underrepresentation of the intended content (because the student was not given the full sample of items that represent the domain). Similarly, a student with a hearing impairment may need an accommodation so that the verbal instructions typically given before or during the test are appropriately transmitted and received. If the accommodation is not given, or is inadequate, that student's test score might underestimate the student's proficiency (because the student's disability interfered with his or her ability to understand the directions for completing test items) and lead to an underrepresentation of the intended content (because the student did not understand the performance expectations for the intended content). Accommodating a test or altering the construct that the test measures can create a risk of misrepresenting or underrepresenting the construct. Of course, misrepresentation may include overrepresenting a construct by making the task or item more than it is designed to be. A test that is altered to better suit the needs of students with the most significant cognitive disabilities may alter the achievement domain being assessed.

No surefire procedure exists for determining the seriousness of this threat other than professional judgment by an appropriate content specialist—for example, a reading consultant. Both research and the sensitivity and understanding of the teacher and the rest of the IEP team are needed to guide us in the valid assessment of a student.

Construct-irrelevant variance. Any factor that is independent of the achievement domain and raises or lowers a test score unfairly is an instance of construct-irrelevant variance (CIV). Construct-irrelevant variance can reflect either construct-irrelevant easiness or construct-irrelevant difficulty. The term *irrelevant* indicates that a factor unrelated to the test content is distorting a student's score in an upward or downward direction. The many sources of CIV (see Haladyna & Downing, 2004) include (a) inappropriate test preparation; (b) student factors—mainly of an emotional nature—such as low motivation, negative attitude, or test anxiety; and (c) cheating. Many other factors can contribute to CIV. In one state, a scoring error led to lower scores for a group of students who, partially on the basis of the low scores, were then identified as at-risk. Only after the students were assigned to and completed a summer remedial program was the error discovered, to the embarrassment of the test company that had made the error (Haladyna, 2002b).

Random error can be large or small, positive or negative. Although it cannot be calculated accurately, reliability offers a way to estimate the error term. This estimate is called the *standard error of measurement*. By knowing the characteristics of the test, we can estimate the

approximate amount of random error in a test score with a certain degree of confidence that the score obtained is representative of a student's status in a domain. If cut scores are used to categorize students, this margin of error is crucial to understanding in which category a student might belong.

Construct-irrelevant variance also can be large or small, positive or negative. In instances of cheating, the error overestimates the actual score. With a scoring error, the source of CIV usually underestimates the actual score. The goal in designing and developing any testing program is to make the third expression (i.e., systematic error) in the equation attributed to CIV equal to zero. In other words, in the context of assessing the achievement of students with disabilities (or, for that matter, all students) systematic error should be as close to zero as possible. CIV should be equal to zero or be so small as to be judged immaterial (i.e., irrelevant to the intended construct rather than immaterial in magnitude—hence the term *construct-irrelevant*).

Table 5 lists some sources of CIV for the population of students with disabilities. This list is not intended to show the full range of threats to validity that CIV presents.

Table 5

Sources of Construct-Irrelevant Variance for Students With Disabilities

Source	Description
Sampling	Exclusion rates vary by states; lack of uniformity misrepresents students with disabilities.
Accommodations	Accommodations offered across states are not standardized, leading to differential treatment of students as a function of the state where each student resides (Haladyna & Downing, 2004).
Identification	Students with disabilities are identified in different ways, resulting in inconsistent identification of those requiring special education.
Data	Nonstandardized data make it difficult to compare the achievements of students with disabilities across different states.
Research on accommodation	Current research on accommodations is not exhaustive or conclusive (Koenig & Bachman, 2004, p. 103). Therefore, the usefulness of the research base in guiding IEP teams and policymakers in the formulation of policies and procedures for improving accommodations is limited.
Effects of disability	A student might inconsistently track visually from the test booklet to the scan sheet, marking the correct answer on the wrong line. Because another student's handwriting might be nearly indecipherable, he or she might print some words to accommodate a tendency towards reversals. Although this student's voice, composition, and conventions are excellent, most raters cannot read the student's writing product. These effects of disability are construct-irrelevant if the skills (e.g., visual tracking or handwriting) are

Source	Description
	irrelevant to the construct being assessed.

CIV is an underrepresented field of study. It is natural inclination not to look for evidence that undermines a claim for validity. However, not looking for CIV risks the exposure of frailties in the argument, claim, and validity evidence that undermine the entire effort to better assess students with disabilities. Because more instances of CIV are encountered among the population of students with disabilities than with other populations of students, greater attention to CIV in this population is recommended.

Consequences of Testing Programs for Students

In its July 2000 position statement on high-stakes testing, the American Educational Research Association states, "Where credible scientific evidence suggests that a given type of testing program is likely to have negative side effects, test developers and users should make a serious effort to explain these possible effects to policymakers." These consequences are subject to evaluation; negative consequences should be reported and action taken to remove them. For example, underestimating a student's reading, writing, or mathematical problem-solving ability may lead to an education program that fails to let the student achieve what is possible. Overestimating achievement may lead to unfair expectations and disappointment later in the student's instructional life.

According to Koenig and Bachman (2004), giving accommodations to students with disabilities increases participation in testing programs such as the National Assessment of Educational Progress. If this consequence is true, then states' and school districts' differential accommodation policies might have a noticeable effect on participation in testing programs and the results of these programs as they benefit students with disabilities.

Documenting the Validation Process

Several sources provide guidance on documenting the validation process (American Educational Research Association et al., 1999, chap. 6; Becker & Pomplun, in press; Haladyna, 2002a). All states, school districts, and other testing program sponsors are encouraged to document validity evidence and make this information available to the public and all concerned, interested parties. Technical reports, Web sites, conferences, newsletters, and press releases are effective means of documenting and disseminating evidence.

Establishing Assessment Validity: Summary and Recommendations

This paper provides an overview of the elements of the validation process (i.e., defining what students learn, stating a validity argument, making claims, and collecting evidence). Two important forms of construct representation and construct-irrelevant variance are considered as critical elements of this process. Validly assessing student learning for students with disabilities is a challenging task. In the domain of knowledge or skills representing state testing programs, validation of test score interpretations or uses is a continuing responsibility to help students learn. Toward that end, states must build a validity argument, make a claim, and collect evidence to support the claim. Validation is ongoing because this continuing process works to improve the testing program and the inferences that the testing program is purported to support.

The sponsor is responsible for engaging in validation procedures to support the use of scores in an accountability system for the state or school district. Although testing contractors may perform many test development and scoring services, it is the sponsor who must ensure its public that it is doing what is best with respect to assessing students' achievement.

In their extensive discussion of why a testing program's tests should be evaluated, Buckendahl and Plake (in press) write: "There are public and professional interests that are served through continuous evaluation and improvement. The independent auditing of an industry that is so critical to society offers needed protection to the public. Ensuring the credibility of tests and testing programs through external verification enhances the reputation of our profession."

Toward that end, the following recommendations are offered:

(a) Create a research agenda for validity studies that seek to solve problems or uncover threats to validity for students with disabilities.

(b) Document all validity evidence. The best tool for collecting and disseminating this documentation is a comprehensive technical report. Establishing and continually updating an archive also are important.

(c) Conduct an annual evaluation to determine the strength of the validity evidence supporting a claim. The evaluation should recommend actions needed to reduce threats to validity and should help strengthen the overall testing program via recommendations for improvement.

References

Abedi, J. (2004, March). Will you explain the question? *Principal Leadership*, 27–31.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*(1), 1–28.

American Educational Research Association. *AERA Position Statement on High-Stakes Testing in PreK-12 Education, Adopted 2000*. Retrieved May 14, 2014 from http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx.

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

Becker, D. F., & Pomplun, M. R. (in press). Technical reporting and documentation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Buckendahl, C., & Plake, B. (in press). Evaluating tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1987). Five perspectives of the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Mahwah, NJ: Lawrence Erlbaum Associates.

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.). (in press). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (2002a). *Essentials of standardized achievement testing: Validity and accountability.* Needham Heights, MA: Allyn & Bacon.

Haladyna, T. M. (2002b). Supporting documentation: Assuring more valid test score interpretations and uses. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment*

for all students: Validity, technical adequacy, and implementation (pp. 89–108). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (in press). Roles and importance of validity studies in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice 23*(1), 17–27.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kane, M. T. (in press). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook on test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. T., Crooks T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Koenig, J. A., & Bachman, L. F. (Eds.). (2004). Keeping score for all: The effects of inclusion and accommodations policies of large-scale educational assessments. Washington, DC: The National Academies Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan.

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

National Research Council. (1999). High-Stakes Testing. Washington, DC: National Academy Press.

Ryan, J. (in press). Practices, issues and trends in student test score reporting. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

U.S. Department of Education. (2004). Standards and assessment peer review guidance: Information and examples for meeting the requirements of the No Child Left Behind Act of 2001. Washington, DC: Author.

Webb, N. (in press). Identifying content for assessing student achievement. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Welch, C. (in press). Item/prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

The U.S. Department of Education is reviewing public comments received on the notice of proposed rulemaking regarding modified achievement standards. As this analysis is not completed, the content of this document may not necessarily reflect the final views or policies of the Department concerning modified achievement standards.

This document was produced under U.S. Department of Education Contract No. EDO4CO0025/0002 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.