SOURCES OF VALIDITY EVIDENCE

Table 2.2 summarizes five types of evidence that can be used to evaluate the legitimacy of a validity claim. Different sources of evidence can, and often should, be used to support the same inference. The key is matching the right types of evidence with the intended inferences and uses.

Table 2.2 Summary of Sources of Evidence

Evidence Based On	*Description*
Test content or construct	Extent to which the assessment items represent a larger domain of interest or construct
Relations to other variables	High correlations with other measures of the same variable or criterion measures and low correlations with measures of related but different variables
Internal structure	Extent to which items measuring the same thing are correlated
Response processes and results	Consistency between hypothesized processes used and expected results with actual processes used and results
Consequences	Extent to which intended and unintended consequences of the assessment are appropriate and desired

Evidence Based on Test Content

The most important type of evidence in our current standards-based climate is based on the content of what is assessed. The idea is that the test items, when systematically reviewed, adequately measure the learning standard or objective. This evidence is used in two circumstances: (1) when there is a specific learning target or standard; and (2) when the test items represent a larger domain of knowledge, understanding, or skill.

It is difficult, if not impossible, to test students on everything they are taught or have learned. Typically, an identified *domain* represents the nature of what it is that we want to make an inference about. The domain, or universe, consists of all the knowledge, skills, or constructs of interest. What we do is assess a *sample* from the larger domain. Evidence based on test content (also referred to as content-related evidence or content validity) includes logical and empirical analyses of how well the sample in the assessment that is administered is representative of the larger domain.

For example, suppose a fifth-grade teacher is giving a unit test on insects, and the teacher intends to use the scores to show how much each student knows about everything that has been taught during the 6-week unit. Can you

imagine how long the test would need to be to cover every fact, concept, and principle that students have been taught? The teacher must make some decisions to sample content from the entire domain and then use the scores on the sample items to make inferences about how much each student knows as defined in the larger domain. For example, if a student scores 75 percent correct on the test, the teacher infers that the student knows 75 percent of the content in the entire unit. How do you know if the teacher's decisions about the content to include in the test are such that the inference about the entire domain, which is made on the basis of the sample test items, is accurate? Here validity becomes a matter of professional judgment. In classroom assessment, the teacher usually makes a judgment about whether the sample is representative of the larger domain. This judgment process can be superficial or systematic. In a superficial review, the teacher makes the judgment in haste on the basis of appearance only. This is sometimes referred to as *face* validity. Face validity means that on a superficial review of the test, the content appears to be representative of the larger domain. While we clearly want to avoid poor face validity, more structured and systematic evidence is desirable.

In a similar way, test users make judgments about the nature of a construct that is being assessed by examining the items to determine if all aspects or components of the construct are represented in the appropriate degree. With constructs, we begin with a theoretical definition and rationale, then build the assessment to be consistent with that definition and rationale. This is important because of the abstract nature of construct. That is, there are different ways of conceptualizing a construct, none of which is necessarily better than others. Consider the construct "critical thinking." To examine critical thinking in education, you need a good match between what you want to emphasize in your school and the definition and theory represented in the particular measure you would like to use. Once the theory is consistent, an examination of the items is needed to make a judgment about how representative the items are with respect to the theoretical rationale.

Suppose a school decided to use a new student self-report instrument that was purportedly designed to identify students who are most at risk to fail and drop out of school. The instrument could be based on a theoretical model of resilience in which various factors contributing to resilience, such as having a hobby and a good relationship with an adult, were assessed. For use in a particular school, teachers would need to review the theoretical rationale and agree that it seemed reasonable for their students, then review the items to determine if the items were consistent with the theoretical rationale and weighted appropriately in the scoring.

Large-Scale Testing

In large-scale educational achievement testing, evidence based on test content begins with a detailed description of the content domain. Once the content domain is defined, items are developed and included in the test to represent the domain. These specifications, called *test blueprints* or *tables of specification*, will show the user and interpreter of the test the extent to which different content

areas have been covered. An example of a test blueprint for the Virginia state testing program is illustrated in Table 2.3. In this example, state "Standards of Learning" (SOL) were used to indicate the content and skills to be covered on the tests. To establish strong evidence based on content, experts in the subject areas reviewed the tests and made systematic judgments about whether the items represented the content. These experts also made judgments about whether the percentage of items in different areas was appropriate and whether some areas that would be important were not on the test. With several individuals making such judgments, the review process is fairly systematic.

In the development of large-scale national standardized tests, the test developers will invest significant resources to be sure that appropriate knowledge and skills are assessed. For commercial test companies, who want their tests to be used in as many schools as possible, this process begins with suggestions from nationally recognized subject matter experts and, more recently, with content standards identified by national associations. Leading textbooks would also be examined to determine the domain of content and skills. Teachers and

Table 2.3 Example of Large-Scale Third-Grade Science Test Blueprint

Reporting Categories	*No. of Items*	*Kindergarten SOLs*	*Grade 1 SOLs*	*Grade 2 SOLs*	*Grade 3 SOLs*
Scientific investigation, reasoning, and logic	10	K.1a-j K.2a, b	1.1a-h	2.1a-h	3.1a-k
Force, motion, energy, and matter	10	K.3a, b K.4a-e K.5a-c	1.2a-d 1.3a-c	2.2a, b 2.3a, b	3.2a-c 3.3a-c
Life processes and living systems*	10	K.6a-c	1.4a-c 1.5a-c	2.4a, b 2.5a, b 2.7a 2.8a-c	3.4a, b 3.5a-c 3.6a-c 3.10a
Earth/space systems and cycles*	10	K.7a, b K.8a-d K.9a, b K.10a-c	1.6a, b 1.7a-c 1.8a-d	2.6a, b 2.7b	3.7a-d 3.8 a, b 3.9a-c 3.10b-d 3.11a-e
SOLs excluded from this test: No SOLs are excluded.					
Total number of operational items:				40	
**Field-test items:				10	
Total number of items:				50	

*Standards from these resource strands are incorporated as these reporting categories.

**These field-test items will *not* be used to compute students' scores on the test.

Note: SOLs stands for Standards of Learning. Numbers and letters in the table refer to specific standards. Reporting categories are test subscales. This test includes SOL for four grade levels.

college professors might be used to indicate the nature of key concepts, ideas, and skills. Following item generation, teachers may be used to examine each item and classify it according to categories of the subject domain and type of cognitive skill being assessed (e.g., recall knowledge or understanding).

Local Classroom Assessment

For classroom assessment, test blueprints are sometimes used to indicate what will be assessed as well as the nature of the learning that will be represented in the assessment. An example of such a test blueprint is shown in Table 2.4. It is a two-way grid in which items are classified by content area and by the cognitive level of the learning. Although making such a blueprint provides a systematic approach to evidence based on content, many teachers will conclude that, in practice, the time it takes to do this outweighs the benefits derived. An alternative is to build a complete set of the learning objectives or targets, showing the number of items and/or percentage of test devoted to each.

Table 2.4 Example of Classroom Assessment Test Blueprint

	Cognitive Level of Learning			
Topic Area	*Knowledge*	*Understanding*	*Application*	*Total*
Types of clouds	5	3	4	12
Types of fronts	5	2	4	11
High and low pressure	6	6	5	17
Wind	7	3	6	16
Total	23	14	19	56

Note: The number of items is shown in this blueprint. Percentages of items can also be used to provide an overview of what is emphasized in different areas.

To make judgments about their assessments, teachers need to have a clear understanding of the nature and structure of the discipline that is taught. They need to know what constitutes true understanding and what is most essential to developing appropriate breadth and depth of the discipline. To do this, it is helpful for teachers to discuss with others what constitutes essential understandings and principles, and to review assessments to make judgments about whether an assessment, when considered as a whole, reflects these understandings and principles. This process is enabled by making sure that the language used in describing cognitive complexity is accurate. Table 2.5 shows how this can be accomplished to show differences between knowledge, understanding, and application.

Finally, with performance assessments, teachers need to extend the essential meaning of validity to how performance is scored. That is, the nature of the scoring criteria needs to reflect important learning objectives. For example, if

Table 2.5 Cognitive Levels of Learning

Cognitive Level	*Definition*	*Types*	*Key Verbs*
Knowledge	Remembering something	Declarative Procedural Recognition Recall Facts Claims Elements Comprehension	Identifies Retrieves Knows Selects Names Defines Reproduces Classifies Recognizes Define
Understanding	Use of knowledge to ascribe meaning	Simple Deep Explanation Interpretation	Understands Converts Translates Discriminates Explains Interprets Infers Distinguishes Predicts Compares Justifies Illustrate
Application	Use of knowledge and understanding to reason and solve problems	Analysis Synthesis Transferability Critical thinking Problem solving Judgment Designing Constructing Testing Perspective	Analyzes Synthesizes Transfers Reasons Generalizes Contrasts Infers Creates Hypothesizes

students are to learn a science skill in which a series of steps needs to be performed, a task that asks students to show their work would help establish a valid inference about whether students have the skill. In addition, to help provide a more valid inference, the scoring of the answers would take into account which steps the students demonstrated and which steps the students did not, giving partial credit where appropriate. If the teacher simply marks each item as correct or incorrect, the total score may not indicate very well what degree of skill the student actually possesses. That is, if items are scored solely as right or wrong, it would be an invalid inference to conclude that the student who missed all the items possesses none of the skills.

Evidence Based on Instruction

Whether the concern is with content or construct evidence to establish validity, it is important in both large-scale and local classroom assessment to have evidence based on instruction (what could be called *instructional* validity). Instructional validity is concerned with the alignment between what is taught or what students have the opportunity to learn, and what is assessed. What is the match between what was taught and what was assessed? Have students had an appropriate opportunity to learn what was assessed? These questions are important because they relate directly to many of the inferences to be made.

Large-Scale Assessment

In large-scale assessment, the alignment between instruction and assessment is critical to high-stakes judgments about students, teachers, and schools.

Suppose a national norm-referenced achievement test is used to determine mathematics achievement. If the mathematics content in the test is not matched with what students have been taught, it would be unreasonable to conclude that the low scores mean that the school is not doing a good job. Similarly, it would not be valid to conclude that a school with low scores on a state competency exam is deficient or poor if the instruction provided does not match well with what is on the exam.

Local Classroom Assessment

For a classroom teacher, this essentially means asking the question, "Were the concepts actually taught, and taught well enough, so that students can perform well and demonstrate their understanding?" Often, this type of judgment is made just before an assessment is written in final form and administered because the answer can be known only after instruction has occurred. Although teachers may begin with an instructional plan, and even have an assessment instrument that is already prepared, only when most of the instruction is completed can the teacher determine for sure the match between what has been taught and emphasized and what is on the test. In making this determination, the teacher, one hopes, will also be able to conclude that student performance is due to learning and not to other factors such as the format of the assessment (e.g., some students are better with multiple-choice), gender, social desirability (e.g., pleasing the teacher when completing an attitude survey), and other influences that would lessen the validity.

Evidence Based on Relations to Other Variables

A second way to ensure appropriate inferences from assessment results is to have evidence that the scores are related to other variables in significant and predictable ways. There are two types of such evidence, one based on how a measure is related to other, external measures (test-criterion relationships) and

one based on obtaining a pattern of relationships (convergent and discriminant evidence). Most of my emphasis will be on the first type.

Test-Criterion Relationships

One type of relationship occurs when a set of scores is correlated to another measure of the same content or construct or to some behavior or performance. This other measure, behavior, or performance is usually called a *criterion* measure. A correlation coefficient is calculated as a measure of the relationship and may be called the *validity coefficient.* Traditionally, there are two types of test-criterion relationships, *concurrent criterion-related* and *predictive criterion-related.* A concurrent coefficient indicates a relationship between two measures that are given at about the same time. A predictive criterion-related coefficient indicates how accurately test data can predict scores from a criterion measure that are gathered at a later time.

In establishing predictive evidence, there are many influences on the criterion measure, making a correlation difficult to establish. Consider the use of grades obtained in high school as a criterion measure for the SAT. Think for a moment about all the factors that affect student grades. Motivation, study skills, peer groups, work, family, effort, goals, and interpersonal skills are all important in determining grades, in addition to academic aptitude as measured by the SAT. Actually, student grades in high school are better predictors of college grades than any aptitude test because they tend to account for more factors. On the other hand, predictors that are almost the same as criterion measures (e.g., aptitude tests given at different ages) will result in a high correlation.

Given these influences, it is not surprising that the correlations typically reported as evidence based on predictive criterion-related relationships are moderate (e.g., .50 to .60). This means that the predictor measure can provide some degree of prediction, but it will be far from perfect. This is one reason why important placement decisions should never be made solely on the basis of a single test score. For instance, requiring students to take remedial summer school if they fail to achieve a designated "cut score" on a measure that purportedly predicts how well students will perform the following year is problematic because many factors contribute to student performance. The result achieved on the predictor test can represent only a part of the prediction.

Test-criterion evidence is used extensively by effective teachers by obtaining two or more measures of the same trait and looking for discrepancies between the scores. This is typically done informally though systematically. That is, the teacher knows what evidence is needed to corroborate the results of other assessments. Thus, teachers may see if homework results are consistent with in-class quizzes, or if observations of students working individually suggest the same level of understanding as evaluated by small group performance.

Teachers constantly make informal predictions of student learning, based on their observations of student work or answers to questions. With experience, effective teachers learn that certain types of behavior and student response will predict how well students will perform on tests. Once this relationship is

established, it can be used to help future students obtain the assistance they need to succeed.

Convergent and Discriminant Evidence

Strong evidence for validity is demonstrated when certain patterns of correlations are reported for two or more measures or instruments. *Convergent* evidence is obtained when scores from one instrument correlate highly with scores from another measure of the same trait or performance (similar to concurrent criterion-related evidence). *Discriminant* evidence exists when scores from one instrument correlate poorly with scores from another measure of something different. When convergent correlations are high and the discriminant correlations are low, the pattern suggests strong evidence. For example, scores from a measure of self-concept would be expected to correlate highly with scores from a different but similar measure of self-concept but to show low correlation with related but different traits such as anxiety and motivation. The discriminant pattern can also be found by examining the subscales within a single instrument. For self-concept, for example, there are often different subscales (academic, social, and physical). Strong evidence for validity would exist if the subscales are not too highly correlated with each other.

Convergent and discriminant evidence is used extensively in developing psychological instruments that assess constructs such as self-concept, personality, attitudes, values, interests, and beliefs. An array of correlation coefficients is presented to show the predicted pattern of relationships.

In the classroom, teachers can apply the logic of this type of evidence by taking note of the consistency of student performance on different measures of the same knowledge or skill (convergent) and by seeing if there is less correlation with assessments of different knowledge or skills (discriminant). This is useful in identifying specific areas that need attention. For example, a teacher might use the following logic as evidence that an inference about a student's comprehension skills is valid: "Sam is able to read all types of different reading passages well in class and on standardized tests" (convergent evidence) "but doesn't always demonstrate a clear understanding of what he reads" (discriminant evidence). The inference is that Sam needs further instruction in comprehension. This conclusion would be less valid if based only on comprehension scores because there would be no evidence that the measure of comprehension was different from reading ability.

Evidence Based on Internal Structure

Large-Scale Assessment

Large-scale and standardized psychological assessments are usually designed so that several items are used to measure each separate trait or important reporting category. The item clusters are identified by how a construct is defined or by identified categories. For example, a measure of classroom climate would have several similar items that indicate a given theoretical

dimension of climate, such as friendship or cohesiveness. If the items focusing on friendship are strongly related to each other and, at the same time, related less to items measuring other components, then there is good evidence based on internal structure. On the other hand, if the friendship items correlate highly with cohesion or goal orientation, or some other dimensions, then the evidence is weak, meaning that it may not be appropriate to report friendship as a separate aspect of classroom climate. Thus, evidence based on internal structure is provided when the relationships among items and parts of the instrument are empirically consistent with the construct, theory, or intended use of the scores.

Local Classroom Assessment

In the classroom, this type of evidence is typically used in criterion-referenced testing, albeit informally. Teachers use the combined results of several items that cover the same skill, concept, principle, or application. It is recommended that students need to answer a minimum of six to eight selected-response items to obtain sufficient consistency to conclude from the results that the students do or do not understand.

To illustrate, suppose a math teacher is constructing a unit geometry test. One of the skills to be assessed is the ability to determine the area of circles and cylinders. To obtain good evidence based on internal structure, the test needs to have several items that assess the ability of the students to determine area of circles and several items focused on cylinders. It wouldn't make much sense to give a two-item test—one item on circles and one on cylinders. Rather, several items for each are needed, and consistency in responses would provide good evidence for the validity of the inference that students do or do not know how to find the area of circles and cylinders.

Evidence Based on the Consequences of Assessments

In recent years, there has been considerable discussion among testing experts about whether the overall judgment of validity of uses and interpretations should include a consideration of the possible *consequences* of using the assessments (Messick, 1989, 1995; Popham, 1997; Shepard, 1997). Consequences could be both planned and unintended. For example, a desirable consequence of using essay questions may be that students learn the content with more depth of understanding than if they prepare for a multiple-choice test.

Large-Scale Assessment

For many large-scale tests, the implicit purpose is to have predetermined consequences, such as a placement test to determine which level of a foreign language the student should take, a high school graduation test to determine if a student is eligible to obtain a graduation certificate, and end-of-year course tests to screen students for summer school. Clearly, in all these cases, the use of the assessment involves important consequences.

The issue of evidence based on consequences takes on a different perspective when considering broader effects, social consequences, and usually negative, unintended effects. The current trend toward more and more "high-stakes" testing, in which large-scale assessments are used to deny grade promotion, high school graduation, and school accreditation, may result in several negative outcomes in the school, such as a narrowing of the curriculum so that it focuses on only the knowledge and skills measured. Important subjects that are not tested may be ignored. Drill-and-practice instructional activities may be used in excess to ensure that students score well, especially if the test emphasizes knowledge as opposed to reasoning. Suppose teachers become less creative and less spontaneous in response to high-stakes student testing. Will teachers change their classroom assessments to match the format used in a high-stakes test? If so, is this desirable? What is the long-term effect?

Local Classroom Assessment

On a more informal level, teachers use the concept of consequential validity continuously. For example, teachers may assess student understanding during instruction with some questions and use the results of the "assessment" to form small groups of students. The consequence is forming the small group of students, and the evidence comes in when the performance of the students is examined. If student performance is maximized, then there is evidence that the use of the informal assessment was valid. Or a teacher may decide, on the basis of informal assessment, that the entire class needs remediation before moving on in the textbook. The consequence is doing the remediation. The evidence is whether the remediation helped.

At the classroom level, student motivation and learning processes are often influenced by the nature of the assessment. A consequence of heavily using objective items is to encourage students to learn for recognition, whereas essay items motivate students to learn in a way that stresses the organization of information, principles, and application. An important effect of essay items is to engage students in reasoning skills, something that is much more difficult with objective items.

SUGGESTIONS FOR ENHANCING CLASSROOM ASSESSMENT VALIDITY

In large-scale testing, there are established procedures for obtaining correlation coefficients as validity evidence. In local classroom assessments, however, teachers must rely largely on nonstatistical procedures to establish the validity of their uses and inferences. Here is a list of