



A Testing Brief
Validity

Timothy R. Vansickle, Ph.D.



Validity

What does it mean when we say a test is valid? Actually, that phrase means nothing. Tests are neither valid nor invalid. The use of a test score can be valid or invalid. Of course, in our requests from clients we see statements such as "...valid and reliable..." What is really meant is that the inferences or interpretations made from the test scores will be valid for a set of specific situations. As an example, you see your physician and he asks for some common blood tests to be performed. The results of those blood tests tell your doctor about some very specific things that might be happening to you. For example, the results may indicate that you have an infection but not a more serious disease. Making an inference about an infection is valid. Making an inference from a simple blood test to a far more serious disease is invalid. Same test results but different levels, if you will, of validity.

Validity is usually described something like this: "...a test is valid, if it measures what it purports to measure..." That is, if a test is designed to be a measure of algebra and the items are algebraic in nature and have been deemed to be relevant to knowing algebra, the results from the test are said to be valid. There is a caveat here. The results are valid when making inferences about a student's algebra ability or achievement. The results are less valid when speaking of a student's math ability and even less valid when speaking of a student's arithmetic ability or achievement. Obviously, the algebra test results tell us little or nothing about the student's English Language Arts knowledge, skills, and abilities.

There are a number of types of validity that have been conceptualized over the years. Each type has a purpose when validating test results. Not all types are needed for every test but some form of validity should be presented in a technical document for every test used in an operational setting. There is also a need for a connection to test development when we discuss validity. That is, early in the design and development process, there should be a statement of purpose for the test. For example: "The XYZ Test is designed to measure the college readiness of 11th and 12th grade students in the content areas of English Language Arts, Mathematics, Science, and Social Studies." Although broadly defined we now have some indication of what the purpose of the test may be and what types of validity may be useful. Many times this statement is missing from the design process and often leads to the misuse of test results later in the process of implementing a testing system. Essentially, the lack of foreknowledge of the purpose of the test leads to invalid use of test results.

The basic types of validity include: Content, Criterion, and Construct. In addition, there are new concepts of validity such as validity generalization, consequential validity, and convergent/discriminate validity. Some of these are more methodologies than new concepts of validity but are mentioned

here as these terms are seen in reports, meeting notes, and other documents you will come in contact with. Finally, there is Face validity. Although not truly a type of validity, Face validity impacts how the test taker may interact with the test being administered. Face validity refers to whether or not the test appears to the test taker to be relevant or related to what type of test they believe they are taking. That is, if the test taker believes the test should be about history and the test looks and feels like a mathematics test, it will have poor face validity and will most likely have a negative impact on the test taker's responses.

The most common type of validity used for tests, like those created under NCLB, is content validity. Although, construct, consequential, and convergent/discriminate validity are being used more often now that NCLB has placed an emphasis on alignment and validity and confused these two concepts.

Content Validity: This form of validity is not a statistical form of validity but rather a process of building validity into the test from the very beginning of the design process. A documentation trail should accompany the test that describes the purpose, the content to be tested, the number and type of items for the whole test and any sub part or scales, the people involved in writing, reviewing, and evaluating test items, the analysis of item quality, and people involved in setting standards, if present in the final scoring and reporting. Most often there is a portion of the technical manual that covers validity specifically but other sections may well contain validity information related to content validity. In addition, other documents may provide support for content validity including: test blue prints or test specifications, item specifications, content standards or frameworks, statements of test purpose, legislative documents and meeting notes from various types of item and test reviews.

Construct Validity: This form of validity is most often provided in a statistical or numeric form. This type of validity argument concerns the construct being measured (e.g., reading, emotional intelligence, or depression). Typically, the test to be validated is administered along with a well known and documented test measuring the same construct, such as reading, to a set of examinees. A correlation coefficient is computed for the scores from the two tests. A high correlation is evidence that the two tests are measuring the same construct. In the case of validity coefficients we hope to see coefficients greater than .65 but do not typically see correlations of .90 or greater.

Criterion-related Validity: This form of validity is typically depicted in terms of a statistical index. Often criterion-related validity is a correlation computed between the test score(s) and some external criterion score. Most often the goal in criterion-related validity is to provide an estimate of how well a test predicts future success on a different indicator such as job or college success. Criterion-related validity comes in two basic flavors: concurrent and

A Testing Briefing

Validity

predictive. This refers to when the criterion measures are collected. In concurrent both the test and criterion measures are collected at approximately the same time. In predictive validity, the test is administered first, all examinees are allowed into the program or go to work in the job the test was designed for, and the criterion is collected at a point in the future (e.g., after the first school year or after a year on the job).

Consequential Validity: Is a loose group of methods that document the intended and unintended consequences of a testing program. Typically, information is gathered from teachers, supervisors, and others through surveys or focus groups to determine whether using a test has had positive or negative effects on, for example, how teachers teach, what is taught, or how the curriculum focus has changed.

The goal in validity research is to acquire several pieces of documentation that support the use of the test's scores in specific manner for a specified group of examinees. If the test scores are used for purposes not supported by validity evidence, that use is said to be invalid.