ALN
ASSESSMENT LEARNING NETWORK

LEARNING POINT

Reliability and validity: How do these concepts influence accurate student assessment?

There are two concepts central to accurate student assessment: reliability and validity. Very briefly explained, **reliability** refers to the consistency of test scores, where **validity** refers to the degree that a test measures what it purports to measure. These very basic definitions have some utility; but these two concepts are so fundamental to educational measurement that assessment literate individuals should have a deeper understanding of each concept.

What do we mean by reliability?

A real-world example makes it easy to see why reliability, or consistency of scores, is important. Suppose someone bought a new bathroom scale. Upon unpacking the scale and using it for the first time, imagine the owner's surprise in finding that the scale reads 50 pounds heavier than expected. A second reading shows a weight that is 60 pounds lighter than expected. Clearly, there is something wrong with this scale! It is not measuring the person's weight accurately or consistently. This is the basic idea behind reliability for educational measurements.

In practice, educational assessment is not quite as straightforward as the bathroom scale example. There are different ways to think of, and express, the reliability of educational tests. Depending on the format and purpose of the test, different types of reliability will be more or less appropriate to report.

Test-retest reliability is simply the correlation between the test scores of people who take the same test, twice. Ideally, a person would receive the same score on the same test each time they take it (if we assume that they have not learned any new, relevant content or forgotten any relevant content between the two administrations. This assumption is often not realistic.)

Internal consistency reliability can be thought of as the degree to which responses to all the items on a test "hang together." This is expressed as a correlation between 0 and 1, with values closer to 1 indicating higher reliability or internal consistency within the measure. Internal consistency values are impacted by the structure of the test. Tests that consist of more homogeneous items (i.e., all measure the same thing) will have higher internal consistency values than tests that have more divergent content. For example, we would expect a test that contains only American history items to have higher reliability (internal consistency) values than a test that contains American history, African history, and political science questions—even if both tests are equally well built.

Inter-rater reliability describes the degree of agreement between scores of different raters on the same student performance. Inter-rater reliability is particularly important for rubric-scored items such as written response or performance assessments. This type of reliability is important because if a written response item or performance task doesn't have inter-rater reliability, who evaluates a performance may have a bigger impact on the score than the

actual performance itself. No one wants to be judged by the harshest scorer, particularly if the assessment is high stakes.

The calculation of the various reliability statistics, the requirements for test construction, and considerations relevant to the judgement of appropriateness of reliability values are beyond the scope of this Learning Point. Understanding that there is nuance in calculation and interpretation of reliability is what is important.

To learn more

Assessment Literacy for Educators in a Hurry, by James Popham ASCD (2018).
http://bit.ly/2Gb8uID

Standards for educational and psychological testing.
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985).
http://apa.org/science/programs/testing/standards

"Scoring rubric development: Validity and Reliability," by Barbara M. Moskal and Jon A. Leydens. *Practical Assessment, Research & Evaluation*, 7(10), (2000).
https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1093&context=pare

What do we mean by validity?

The assessment of a test's validity is even more complicated and nuanced than the calculation of reliability. Stated in an over-simplified way, reliability is like a spreadsheet in which we can calculate a correlation value. Validity, on the other hand, is like a courtroom in which the test maker has to make a case for the validity of a test.

> *"A test is not valid or invalid, per se. A test is valid or invalid only for a stated purpose."*

Perhaps the most important thing to know is this: a test is not valid or invalid, per se. A test is valid or invalid only for a stated purpose. One cannot assess the validity of a test unless the purpose of that test is made clear. Additionally, a test may be valid for one purpose but invalid for another. As an example, a test might provide valid estimates of student achievement of content, but it might not be a valid test for judging the quality of a school. A validity argument must be made for each stated purpose or use of a test, and supportive evidence must be collected to "validate" each intended use of the assessment.

Historically, it was thought that there were different types of validity. Using a biology test as a context, historical notions of validity included:

- Face Validity - The biology test looks like a biology test
- Content Validity - Biology experts agree that the test is a biology test
- Criterion Validity - Scores on the biology test correlate with other aspects of knowledge of biology
- Predictive Validity - Scores on the biology test correlate and predict scores on some other test of biology or other factor
- Construct Validity - Results of the test help describe the construct of biology knowledge (usually supported by use of a statistical method such as factor analysis)
- Consequential Validity - The decisions made based on the results of the biology test are appropriate and constructive

As time and research progressed, more unified theories of validity caught on. Rather than different types of validity, specialists came to feel that different types of evidence are required to establish validity. This means that the intended uses of an assessment have to be supported by evidence that supports the use(s).

Validity is thought of now in terms of the appropriateness of decisions based on test results. Basically, it is currently thought that consequential validity subsumes all the other notions of validity. This interpretation is aided by a knowledge of the historical notions of validity in that they can help guide one in making the case for appropriate use of the test results for a specific purpose. Someone tasked with developing a validity argument can use the historical notions of validity to determine what type of evidence should be collected and brought to bear in support of the proposed use for the assessment.

The Michigan Assessment Consortium's Assessment Learning Network (ALN) is a professional learning community consisting of members from MI's professional education organizations; the goal of the ALN is to increase the assessment literacy of all of Michigan's professional educators.