IMRF Measurement Module: Validity and Reliability

Validity: Valid Test or Valid Scores and Valid Use?

For a concept that is the foundation of virtually all aspects of our measurement work, it seems that the term validity continues to be one of the most misunderstood or widely misused of all. Of course, to say definitively what validity should mean suggests that there is some absolute truth about what validity really is. In the absence of such truth—which is where we seem to be—we generally depend on consensus in the field to reach a common understanding and a standard use of the term. The most current version of the *Standards for Educational and Psychological Testing* (*Test Standards*) (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) has generally been a trustworthy source of consensus in the field regarding conceptual understandings, the meanings of specialized terms, and expectations for acceptable practice.

It is clear from the current *Test Standards*, that validity is *not* primarily about instruments themselves, but it is about score interpretations and score use. To speak of a "valid test" begs the question about what such an instrument is. If we quickly say that such a test is one that measures what the test maker intended, that response begs still another question: Can a "valid test" yield scores that do *not* represent the type of meaning the test maker intended, or scores that should *not* be used in the way the maker originally intended? In other words, can a "good test" be used to give us "bad" information, or can the information from a "good test" be used in "bad" ways? Surely the answers to both questions are affirmative. Consequently, failure to realize that well-developed instruments can give us meaningless scores, and, therefore, scores that should not be used as originally intended, is a certain indication of misunderstanding on the user's part about what validity is.

Here are some examples from a variety of sources that demonstrate the kind of misunderstanding about validity that have occurred:

1. "... the term 'screening reading assessment' means an assessment that is valid . . . and based on scientifically based reading research; . . ." (From a law passed by Congress)
2. ". . . you can help ensure that the test will be valid and equitable for all students." (From an examiner's manual for a statewide assessment program)
3. "Evidence of test validity . . . should be made publicly available." (From a major publication of a prominent testing organization)
4. "In the assessment realm, this is referred to as the validity of the test." (From an introductory assessment textbook)
5. "[Test name] has proven itself in use for more than 50 years as a ... valid test . . .." (From the web site of a prominent test publisher)

The authors of this module have been evasive intentionally in attributing these quotations because the point is not to condemn the source. Rather, it is to demonstrate that the term *validity* is consistently used in print in a variety of sources in ways that are counter to our consensual understanding of what the concept means.

This misunderstanding can be illustrated further with a spelling test from a standardized achievement battery. A sixth-grade spelling test might contain 38 exercises, like the one shown in Figure 1, in which students are presented four words per exercise and are asked to identify the one that is misspelled. A fifth option allows them to indicate that there are no mistakes—all four are spelled correctly. Assume that tryout and standardization results indicate that these words are appropriate to use to measure the spelling achievement of students in sixth grade. By almost any standard, this is probably a good test. The assessment purpose is to use the score a student obtains to generalize beyond the 152 words in this test (38 items × 4 options = 152) about how well a student can spell words that students in sixth grade usually encounter.

1. A diamond
B model
C tribe
D anxous
E (No mistakes)

FIGURE 1. Sample spelling item.

Now suppose that in the administration of this spelling test in a certain sixth-grade classroom, any one or more of the following happened:

a. The teacher allowed only 8 minutes instead of the prescribed 12—a timing error occurred.

b. Two students copied answers from other students' papers—cheating occurred.

c. A student coded his document as Test Form B, but he actually took Form A— a scoring error resulted.

In a very simple way, these scenarios point out why a "good test" may not be a very "valid test".

What are some potential consequences of the inconsistencies in usage in the case of either "valid scores" or "valid use" versus "valid test"?

1. *Weak validation*. The misinformation conveyed by our coarse precision can lead test developers and users to unintentionally short-circuit the validation process. If validation is about the test, they might reason, then evidence for validity should come only from the test and how it was developed. Consequently, the evidence gathered would focus on the content-related type (see below), but other potential sources of invalidity might be ignored. In fact, the possibility that the administration or scoring of the test could have an impact on validity may seem inappropriate to some, or even unnecessary to consider by others.

2. *Failure to consider* score use *in the validation process*. When the validity argument is limited to score meaning only, we may end up using good scores from good tests in bad ways. Consequently, scores used to satisfy one purpose might be used blindly to make decisions for another context without giving much thought to its appropriateness.

3*. Incomplete directions for test administration*. If we think only about test validity, we can easily overlook some of the things that can cause test scores to fall short of the meaning we originally intended them to have. For example, if we think the directions for administration affect only the standardization of the measurement process and the applicability of norms that may go with the scores, we can easily fail to guide test users to practices that will help preserve the meaning we had hoped to attribute to the scores. As an example, failure to provide guidance in the use of accommodations or modifications for individuals with special needs can certainly lead to inappropriate practices in achievement testing situations and, therefore, to less meaningful scores. And providing guidance in the directions about whether to permit calculators to be used on a math test, or about what students should be told about random guessing or about proofreading written responses, can lead to more valid and useful scores.

4. *Interpreting scores as though only the quality of the test matters*. Educators and researchers tend to be accepting and unquestioning in their use of assessment tools. If it is regarded widely as a good instrument, or if it has been around for a lot of years, then it *must* give dependable results. "Measurement you can trust" is a slogan that goes with some commercial tests. When the focus is mainly on the test, we can, for example, inappropriately interpret the scores of random responders as though they

> actually tried. Or we may overlook such things as "overzealous test preparation," inappropriate use of accommodations, or poorly developed scoring rubrics for performance assessments.

No doubt these consequences, and others that could be identified, suggest that the damage done by using such terms as *test validity* or *valid test* is far greater than we might realize. Researchers who measure various attributes of the subjects in their study must be able to demonstrate that the scores they have gathered through their measurements have the meaning to justify both the conclusions they draw and the generalizations they offer. In reporting their research findings, researchers should provide enough information so that the consumer of their work, the reader, can decide whether those findings are likely to be relevant or applicable to the user's context. Are the findings generalizable beyond the research setting?

Reliability: Reliable Test or Reliable Scores and Reliable Scoring

Some of what was noted above about validity has a parallel with reliability. Not only do many researchers using measurements confuse validity and reliability with each other, but they also develop a very limited view of how to use the two ideas. Among the many points of misunderstanding or misuse by practitioners and newcomers alike, two propositions about reliability have been selected here to probe more deeply. First, consider this relationship:

Reliability is a property of a set of scores, not of the assessment that produced the scores.

As mentioned previously, the *Test Standards* is regarded as the most authoritative source regarding the standard usage and meaning of the concepts used in the field, as well as the general expectations for appropriate practice. The first paragraph of the validity chapter in the *Test Standards* makes the point that the focus on validity is interpretations and uses of scores rather than on the tools that produced the scores. It seems that a similar statement needs to be made about reliability, and with the same rigor and emphasis. That is, reliability is about scores and not about an instrument per se and not about the individuals who obtained the scores. Certainly, scores are affected by the instrument used and by the individuals assessed—and more, but that is the point. Why should reliability be characterized as describing anything other than the scores that result from a measurement process?

Turning to the most current *Test Standards*, the opening paragraph of the reliability chapter begins with this definition:

> *Reliability refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups.* (p. 25)

The glossary definition in the *Test Standards* is compatible with this one and refers the reader to "generalizability theory," which provides for generalizations "beyond the items, persons, and observational conditions" of the situation. The *Test Standards* commentary and the *Test Standards* themselves do not explicitly caution against use of the terms *test reliability* or *reliable test*, as they do for *valid test*. This is unfortunate because making the distinction explicit could be a preventative measure that might serve to reduce the frequency of occurrence of usage like the following:

1. "Such assessments shall be used for purposes for which such assessments are... reliable ... ." (From a law passed by Congress)
2. "The contractor will perform psychometric analyses to monitor the content validity, construct validity and reliability of the tests." (From an RFP from a state education department)
3. "Because test reliability is greatly influenced by the number of items in a test,.. ." (From a technical manual of a prominent achievement test battery)

Here is a second proposition about reliability to consider:

> *Score reliability differs from scorer reliability, and the need for one kind of estimate cannot be satisfied by the other.*

These ideas may represent a greater source of confusion for measurement practitioners and students than for measurement specialists. The distinctions between these terms have been made clearly in the *Test Standards*. In the current edition, Standard 2.10 says, in part:

> *When subjective judgment enters into test scoring, evidence should be provided on both inter-rater consistency in scoring and within-examinee consistency over repeated measurements.* (p. 33)

There is recognition and general understanding among measurement professionals that these two types of reliability evidence reflect the effects of different sources of measurement error and that these sources are independent of one another. But for many assessment practitioners or non-specialists, this distinction often is not recognized.

The limited view, which some practitioners hold, is that reliability is a unitary notion that can be estimated by, or represented by, a single number. Further, some may believe that there are several ways to obtain the number, but each method yields essentially the same result. Of course, this means, they additionally think that you can just choose the method that will be easiest to use in your circumstances. Better yet, some might reason, if someone already has rater-reliability information for the performance assessment I'm using, I can just use *their* value to show how reliable my scores are. This line of thinking is closely associated with another: "If someone else has shown how reliable this *test* is, there's no need for me to duplicate their fine efforts." For those who hold these misunderstandings, the important connections between sources of measurement error and reliability estimation methods remains elusive.

The consequences of maintaining a blurred view of score reliability and scorer reliability are likely to show up in the processes of instrument selection and development, as well as in establishing the technical adequacy of a specific assessment procedure. Here are some illustrations:

- In making a decision to adopt a particular instrument for a research study, the researcher allowed a high level of reported rater agreement to compensate for a relatively low reliability coefficient.
- In developing a performance assessment in math, the user neglected the notion of content sampling error in deciding on the number of tasks or on working-time requirements. Instead, effort was directed at developing rubrics that could generate high levels of scorer consistency.
- In documenting the technical adequacy of scores from the *most recent* use of a performance assessment, the user reported only percent-of-agreement indices from *previous* scoring sessions that had taken place.

Some Implications for Research Proposal Development

Researchers in education and the behavioral sciences typically conduct studies that involve the measurement of traits of human subjects. When there is no existing instrument for use in a proposed study, part of the research process requires the construction of the needed instrument. In other cases, several different instruments may exist already, and the researcher must select the one that is most appropriate for the researcher's situation. Whether constructing or selecting an instrument, the burden is on the researcher to identify how the validity and reliability of the scores, the measurements, will be determined. What kind of evidence will the researcher obtain to demonstrate that the scores are sufficiently valid and reliable to warrant drawing the types of conclusions that are hypothesized?

<u>Validity Evidence.</u> For proposed research, the process of gathering evidence to support intended score interpretations or score uses should be provided. The type of evidence to be obtained should depend on a number of contextual variables. The *Test Standards* identifies several categories of evidence that might be considered.

a. <u>Content-related evidence</u> generally comes from documentation of the instrument development process. Is each of the items or statements relevant to the trait being measured? How do you know? Are all of the key facets of the trait represented by the collection of items/statements to which subjects will respond? What conceptual framework for the trait will be used to make judgements about such aspects of the instrument and its items/statements?

b. Criterion-related evidence generally is described as either of two types—concurrent or predictive. If, for example, I can show that the scores from my cognitive ability test correlate highly with scores from some other widely-accepted measure of cognitive ability, this would be evidence to support the claim that my instrument measures much the same trait as the generally-accepted one (concurrent). If, however, I intend to use a certain instrument to predict future behavior, I might need to gather predictive evidence. For example, if I were using a screening device to predict who might benefit from a certain training program, I would need scores from my screening device and from some measure of success in the training program for a group of subjects who have completed training. The correlation between screening scores and success scores would be predictive evidence for the former. (Of course, there still would be a need to gather evidence to support the validity of the success scores.)

c. Construct-related evidence is needed when the user wants to claim that scores on a certain psychological measure do in fact represent the responders actual standing on the trait of interest and not some other similar trait or a closely-related one. A variety of quantitative or qualitative evidence might be used for this purpose. For example, if I am using an instrument to measure "Attitude toward School," l might gather judgements from a group of teachers about the content of the statements to which subjects would respond, using a continuum from "Strongly Agree" to "Strongly Disagree." Does each statement seem to belong on such an instrument? Are there significant aspects of "school" missing from the set of statements presented? Maybe factor analysis would be used to show the extent to which the items together measure a single trait. Some types of reliability estimates might also be used for supporting the stability of this attitude over time, as measured by this instrument.

Reliability Evidence

As with validity, the *Test Standards* provide useful guidance regarding the type of evidence to obtain to support the claim that a set of scores is sufficiently reliable to warrant the use of the scores proposed by the user. The large number of methods available for this purpose is one indication that the user must decide what kind of reliability estimate should be selected for their specific situation, or intended score use. Some factors to consider are: the types of measurement errors that are most likely to influence the accuracy of the scores, the stability of the trait in individuals in the population that will be measured, and the nature of the theory (e. g., classical vs. latent trait) that is being used. Some methods of estimating score reliability are:

a. Test-retest involves administering the same instrument to the same group on separate occasions. The correlation between the two sets of scores is an indication of the extent to which the scores of individuals are in the same relative rank order on the two occasions, and thus, an indicator of consistency of measurement.

b. Equivalent forms, sometimes called Parallel forms, involves administering two equivalent forms of an instrument to a single group on a single occasion. A high correlation between scores from the two forms can be interpreted to mean that the content of the items/statements of either form is representative of the same domain—content sampling errors are not likely interfering with the measurements obtained. That is, a subject's score on this instrument does not seem to be dependent on which specific items/statements in the domain of interest the person responded to.
c. Internal consistency estimates, like Coefficient Alpha and Kuder-Richardson, are convenient ways to obtain equivalent-forms estimates. In effect, these estimates are indicators of the extent to which content sampling errors might be influencing the scores, and thus, the interpretations made with them. The advantage, however, is that only one form of the instrument need be administered to obtain the reliability estimate.

Other methods of estimating score reliability, such as a generalizability coefficient, or the use of an information function, might be needed for other circumstances. The reader can consult other resources for further details about them.

Closing Comments

Validity and reliability are terms used with the results of a measurement process, whether for research purposes or for obtaining an educational or psychological assessment. Measurements result in numbers, and it's these numbers that we raise questions about when we speak of validity or reliability. What do the numbers mean? What are some appropriate ways of using these numbers? Are the numbers fairly reproducible if we could repeat the measurement process with these same subjects or a randomly-equivalent group?

Iowa Measurement Research Foundation
Research Committee
December 2022