II.2 Technical Concerns About Assessment

How good are the assessment procedures outlined in Procedures and Instruments for Assessing Learning Problems (p. 341)? Some are terrible. Some are useful. Even the best are not good enough.

Many widely used assessment devices, even those that are commercially marketed, have technical deficiencies. To understand why this is so requires knowledge of what is meant by the concepts of reliability, validity, norms, and standards.

Assessment procedures (tests, rating scales, interview and observation schedules) produce descriptive data that vary in their degree of reliability and validity. In judging and interpreting the meaning of data, assessors use norms and standards that also vary in their reliability and validity.

What does it mean for an assessment procedure to be reliable and valid? What are norms and standards? Let's look at each of these matters.

Reliability

In common usage the term reliability refers to dependability, accuracy, and precision. As related to assessment, the concept is used to indicate how consistent and reproducible assessment data are.

Just as assessment is not limited to tests, the concept of reliability is relevant to all assessment activity. Furthermore, it is used to determine whether findings are consistent and reproducible over time, in different settings, despite differences in assessors. The concept is also used frequently to describe the consistency among procedures that claim to measure the same thing, such as several measures of intelligence.

Reliability, however, says nothing about what the procedure is measuring. The concept of validity deals with what the findings mean.

In technical terms, reliability refers to the degree a procedure is free of random error. Mathematically, reliability is presented as a number that varies from zero to one. This number is called a correlation or reliability coefficient. The coefficient for a totally unreliable procedure would be .00, perfect reliability is indicated by 1.0.

A common way to arrive at a reliability coefficient is to analyze the findings from several comparable administrations of an assessment procedure. Assuming what is being measured has not changed, the findings should be similar.

When the findings are not similar, it may be because of deficiencies in the procedure or because what is being measured has changed. On the other hand, highly reliable findings do not always mean that a procedure is technically sound; highly reliable findings can also be produced by biased administration or interpretation (nonrandom error).

No psychological assessment procedure is completely dependable or free from random error. With regard to tests, it has been pointed out that unless the test is perfectly reliable, and such instruments do not exist to the authors' knowledge, there is likely to be some margin of error in every test score. Indeed test theorists concede that, were a hypothetical individual given the same test a great many times (assuming that learning and practice effects did not systematically change scores), there would be a range of score values observed, with the range being narrow for reliable tests and broad for unreliable tests. The examinee's "true" score would be the average of all the observed scores. This conception is an elaboration of the idea of reliability, and is presented in an easily comprehended manner in materials developed by test publishers such as the Educational Testing Service. Users of many of the tests published by this organization are advised to make decisions about individual students not in terms of the precise score obtained, but rather in terms of a band or interval extending on each side of the score. This "confidence band" is calculated on the basis of the reliability coefficient for the test and is to be interpreted as having a sufficiently high probability of including the student's true score. (Skager & Weinberg, 1971, pp. 121–22)

Validity

Valid data and decisions are the main concerns of assessors. The concept of validity applies both to
assessment processes and to decisions based on assessment findings. If data are not highly valid, we cannot be very certain about what they mean. This leads to controversy and to a great many errors in interpreting the findings and making decisions.

Discussions of the validity of assessment procedures can be confusing. The concept relates to whether assessment and decision-making procedures are leading to meaningful

- descriptions of a phenomenon (Does the procedure measure what it says it does? Only what it says it does? All of what it says it does?)
- interpretations and judgments of a phenomenon (Are the inferences justified? How appropriate are the norms and standards?)
- decisions (How relevant are the data for the decisions that must be made?)

To understand the concept in more specific terms, we need to look at several of the related types of validity discussed in the assessment literature. Furthermore, while reliability is established in a relatively technical and objective way, validity usually is determined through a great deal of rational and subjective activity.

Three basic types of validity are

- content (How well does the procedure assess the knowledge, skills, and behavior it claims to measure?)
- criterion-related—including predictive, concurrent, and diagnostic validity (How well does the procedure assess the relationship between currently measured phenomena and future, concurrent, or past phenomena?)
- construct (How well does the procedure assess some theoretical concept, such as intelligence or anxiety?)

**Content Validity**

Content validity refers to how well the items on such procedures as standardized achievement tests and course exams actually sample learned skills, knowledge, and behavior. For example, several items that require adding two single-digit numbers may be used to test whether a youngster has learned this skill.

How do we know if a procedure has high content validity? We know because we or someone else makes a judgment that it does. In the example of adding single-digit numbers, the content validity of the items seems so evident that they are seen by most people as appropriate on their “face value”—thus the term face validity. However, when a great many possible items can be included on a test, judgments about the content validity of many of them may be challenged. (Think about the items on any recent test you have taken; chances are you thought some were not a valid assessment of your knowledge and abilities.)

When more than face validity is necessary, judgments made by experts are used to establish content validity. However, the more comprehensive the content area to be assessed, the harder it is to sample and the more likely experts are to disagree.

Take any standardized achievement tests as an example. The content validity of such tests is almost always judged to be high. That is, a group of experts has designed and judged the items to be a good, representative sample of skills in a particular area such as reading. At the same time, inspection of the popular reading tests on the market shows they vary markedly in their content. All claim high content validity, yet they differ in many important ways. The experts clearly have different opinions about what the content of this type of test should be, and one suspects that each would disagree with the judgments of others. And since reading is an area in which content has been relatively well defined, the problems can only be greater in areas where knowledge and skills have not been clarified very well.

Even when there is expert consensus about the high content validity of a test, it may be a poor measure for some individuals and groups. For instance, a math test that requires reading directions or has a short time limit is a poor measure of what a person knows about arithmetic for any individual with a reading problem or anyone who works slowly.

In sum, statements about content validity reflect someone’s subjective judgment. For most consumers, this means they must choose from among expert opinions. In doing so, it is helpful to have clearly in mind what one wants to assess and why as bases for discriminating among conflicting expert views.

**Predictive, Concurrent, and Diagnostic Validity**

Criterion-related (including predictive, concurrent, and diagnostic) validity refers to how well current assessment data can be used to help understand some future, present, or past concerns.

An example is assessment designed to predict
future school failure. To establish predictive validity in such a case, research must show a strong relationship between the data used to predict and some criterion representing subsequent school failure, such as poor grades or low achievement scores. As with reliability, correlation coefficients are used to indicate the strength of the relationship.

Criterion-related validity coefficients for comprehensive psychoeducational assessment procedures generally are not high. Coefficients as high as .60 are rare. Some assessment experts recommend that procedures with coefficients as low as .30 can still be useful (Garrett, 1954; Guilford, 1956).

Decisions as to whether a procedure is valid enough depend on how much new and unique information the data represent. If a procedure has the potential to add a piece of important information that will otherwise not be available, it may be worth using even if it has a low coefficient. On the other hand, it makes little sense to use a procedure that adds no new information even if it has a relatively high coefficient.

Ultimately, the point is to make good decisions. Every criterion-related procedure results in some errors. In screening for future reading problems, for instance, some individuals will be identified as future problems but turn out not to be; this type of error is called a false-positive error. Some individuals will not be identified as future problems but turn out to have difficulty learning to read; these are called false-negative errors.

The number of errors can be used as another way to look at the validity of the assessment and decision-making procedures. That is, computation of the number of correct predictions and the number of errors expressed as proportions provides another index of validity.

One major complication related to establishing criterion-related validity is the selection of appropriate criteria with which to correlate assessment data. For instance, what is the most appropriate criterion of school success or failure? Grades? Amount learned? Ability to apply what has been learned? Positive attitudes toward learning? Obviously, each is debatable. In choosing a criterion, four major qualities have been stressed: (1) relevance, (2) freedom from bias, (3) reliability, and (4) availability (Thordike & Hagen, 1977).

Another complication is that the validation process requires a criterion measure that is highly reliable and valid. For the measurement of complex psychoeducational criteria, such measures simply do not exist. At best, what is available has moderate validity. Thus, the irony often is that researchers must use criterion data gathered by measures that have rather limited reliability and validity; this includes most of the procedures cited in this text.

A variety of other factors can confound the validation process related to criterion-referenced assessment. They need not be reviewed here. The point is that all these complications and confounding factors make it evident why so many of the available predictive and diagnostic assessment procedures should be used with great care and discretion.

In choosing such procedures, the question is not which will avoid making errors—they all make errors. The question is which will produce the smallest number of errors. And although the procedure that produces the smallest number of errors is the best that is available, it may not be good enough. Certainly, it will not be good enough for individuals for whom the findings are in error.

Finally, we note that there is a tendency to infer causation from criterion-related measures. For example, a test may predict failure, but this is not sufficient evidence that what the test measures is the cause of the failure. This jumping at conclusions reflects the common mistake of observing correlates and believing one is seeing cause and effect. The establishment of cause-effect connections requires construct validation.

**Construct Validity**

Construct validity refers to how well an assessment procedure measures something that is not directly observable, such as theoretical concepts or abstractions (intelligence, anxiety, perception, motivation). It also is used in theorizing about relationships among concepts and things that are observable (Nunnally, 1978).

The term construct is used to clarify that what we are trying to measure is a hypothetical idea constructed by scientists to help organize thinking about something that can’t be seen directly. In effect, a construct is a myth that is found to be a useful and convenient way to understand and communicate about a theoretical notion. Of course, it is rarely discussed as a myth as long as enough scientists find it useful.

Because constructs are not directly observable,
they are not directly measurable. What then is measured?

Let's take intelligence as an example. Intelligence is an idea that has been constructed to reflect an attribute of people that many of us find interesting and important. Theories have been formulated to suggest the nature of this attribute and the behaviors that should be directly and indirectly associated with it. For instance, it has been suggested that one of the abilities that reflect intelligence is the ability to reason using verbal, graphic, and mathematical symbols. It also is argued that some groups of persons, such as the majority of university students, are likely to be more intelligent than others (for example, the majority of those who drop out of high school).

Efforts to develop a test of intelligence will include—among other items—some that measure verbal reasoning. Efforts to determine the construct validity of the measure of intelligence, then, might involve comparing groups of university students and high-school dropouts as to their performance on tests of verbal reasoning ability. If the majority of university students are found to score higher, the findings will be seen as providing some evidence for the construct validity of the measure.

Obviously, no one set of findings is enough. Construct validation requires consistent findings from many different studies.

Construct validity is established through a long-continued interplay between observation, reasoning, and imagination. . . . The process of construct validation is the same as that by which scientific theories are developed. (Cronbach, 1970, p. 142)

As with content and criterion-related validity, reasoning and judgment play an important role in construct validation. Also, as with content validity, there is no one correlation coefficient that describes how valid a construct measure is. However, unlike content validation, expert judgment alone cannot be used to validate a construct measure. Conclusions must be justified by an extensive body of research.

A major complication in the validation process is that constructs, such as intelligence, motivation, and self-concept, often are defined differently by those who develop the procedures used to measure the phenomena. Thus procedures that claim to be measuring the same construct may be very different from each other. This means the findings from validation studies using different measures often are not comparable. Therefore, even though many studies may have been done, it may not be appropriate to combine them as evidence for construct validity.

There is wide recognition of the difficulty of validating measures of constructs such as visual perception, psycholinguistic ability, and minimal cerebral dysfunctioning. Thus it is obvious that many assessment and decision-making procedures used in diagnosing such factors require a substantial amount of additional validation.

---

**Norms and Standards**

One of the most basic aspects of human behavior is that people try to make sense out of what they see. Complex things are observed and conclusions are made: the new neighbor appears to be wholesome and friendly; the person walking behind you in an unlit parking lot at night looks like a mugger or worse; students are judged by their teachers to be smart or not too smart; teachers are judged by students to be good or terrible.

What is the basis for such conclusions? Obviously people are influenced by the information available to them in a given situation. They also draw on past experiences. In addition, they may be influenced by some general ideas, theories, attitudes, values, and beliefs they have developed.

Because of differences in the information available and factors influencing how the information is processed, people often arrive at different conclusions about what seems to others to be the same phenomenon. In interpreting the same assessment data on an individual, one assessor may arrive at a diagnosis of learning disabilities, and another may conclude the person has no disability; among a group of students experiencing the same program, some may adore it, others may hate it, and the rest may think it’s merely OK.

Given that people are seeing the same things, the differences in their conclusions probably reflect the use of different norms and standards in making judgments. For our purposes, the term standard refers to the use of values or a theoretical idea to make judgments about what has been assessed. The judgments often are about whether what is assessed is a problem, is “good” or “bad,” or is consistent with some theory.

**Norms** are empirical standards. They are not value statements or theoretical statements. Formal
norms are based on research and systematic observation. For example, a set of previous findings can be used as a set of norms in judging assessment data. Are the current findings higher, lower, or the same as the previous findings? How much higher or lower?

Assessment norms reflect data gathered on various research samples. The score for someone subsequently assessed is compared to the scores for the previous samples. The individual's score will be described in terms of how it compares to the average (mean or median) score for the research samples. It may be found to be above or below average or typical or atypical.

After norms are used, it is commonplace to apply some set of standards to make judgments. This can happen so quickly that it often is not apparent that judgments have been made. For example, a score above or below average may be quickly translated into a judgment that the performance was acceptable or unacceptable, passing or failing. This is understandable, especially with tests of achievement. However, it is important to understand that a value judgment—a standard—has been used. The use of norms by themselves does not lead to a judgment.

Why the distinction between norms and standards is important can be seen by taking a closer look at judgments based on commonplace assessments of behavior and performance. For example, we may have norms that show that an individual's activity level is well above the average found in research samples. On the basis of this information, should we conclude that the individual is highly active or hyperactive? The latter judgment carries with it the implication of a problem and pathology. Such a judgment may not be necessary based on the norms—especially if many of those studied previously who scored at the same level were not found to have problems. Thus the judgment of hyperactivity would reflect someone's standards (the assessor's standards based on theory). Somebody else might use the same findings and norms, but a different set of standards, and decide the individual is not hyperactive.

In general, the decisions that follow assessment activity usually are influenced by the norms and standards that have been used in making judgments about the meaning and significance of the assessment findings. Unfortunately, adequate norms are frequently unavailable, and the standards used in making judgments often are extremely controversial.

To illustrate the problem, we can look at the matter of norms for intelligence tests. The individual intelligence test for children developed by Wechsler is among the best assessment devices available to LD professionals. With regard to available norms, there is more data on this instrument than on almost any other major procedure. But, as we have suggested already, the statement "best available" should not be too readily interpreted as meaning that the instrument is unquestionably a "good" procedure. As published in 1949, the sample used to standardize and establish norms for the test included only white children. In the years prior to the test's 1974 revision, countless individuals were tested and decisions were made based on norms from this extremely narrow sample. The inadequacy of the norms (and apparent bias in the test's content) with regard to the types of youngsters included in the original samples led to enormous criticism, including judicial action. In an attempt to correct the situation, the 1974 and 1991 revisions gathered norms on samples representative of the population in the United States with respect to socioeconomic status, race/ethnicity, and geographical region distribution based on the most recent census data. For example, the 1991 revision used 1988 census data in selecting a validation sample consisting of 2200 children between ages 6 and 16, with an equal number of males and females. However, considerable criticism still is directed both at the content and the fact that the construct validity for many subgroups in the population remains to be established—including persons with learning disabilities. Furthermore, there is controversy over the widespread tendency to use IQ levels as standards for judging people as "good" or "bad" in a variety of contexts.

---

**Summing Up**

Understanding the concepts of reliability, validity, norms, and standards allows one to appreciate the current state of the art and the hurdles that must be overcome if psychoeducational assessment and decision-making practices are to improve. Such hurdles are not insurmountable. To do so, however, will be costly and require the talents of many.