

VALIDITY OF THE ASSESSMENT TOOLS

Validity

The validity of an assessment tool is the degree to which it measures for what it is designed to measure. For example if a test is designed to measure the skill of addition of three digit in mathematics but the problems are presented in difficult language that is not according to the ability level of the students then it may not measure the addition skill of three digits, consequently will not be a valid test. Many experts of measurement had defined this term, some of the definitions are given as under.

According to Business Dictionary the “Validity is the degree to which an instrument, selection process, statistical technique, or test measures what it is supposed to measure.”

Cook and Campbell (1979) define validity as the appropriateness or correctness of inferences, decisions, or descriptions made about individuals, groups, or institutions from test results.

Howell’s (1992) view of validity of the test is; a valid test must measure specifically what it is intended to measure.



Methods of Measuring Validity

Validity is the appropriateness of a particular uses of the test scores, test validation is then the process of collecting evidence to justify the intended use of the scores. In order to collect the evidence of validity there are many types of validity methods that provide usefulness of the assessment tools. Some of them are listed below.

1. Content Validity

The evidence of the content validity is judgmental process and may be formal or informal. The formal process has systematic procedure which arrives at a judgment. The important components are the identification of behavioral objectives and construction of table of specification. Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. For example, a test of the ability to add two numbers, should include a range of combinations of digits. A test with only one-digit numbers, or only even numbers, would not have good coverage of the content domain. Content related evidence typically involves Subject Matter Experts (SME's) evaluating test items against the test specifications. It is a non-statistical type of validity that involves "the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured" (Anastasi & Urbina, 1997).

For example, does an IQ questionnaire have items covering all areas of intelligence discussed in the scientific literature? A test has content validity built into it by careful selection of which items to include (Anastasi & Urbina, 1997). Items are chosen so that they comply with the test specification which is drawn up through a thorough examination of the subject domain. Foxcraft et al. (2004, p. 49) note that by using a panel of experts to review the test specifications and the selection of items the content validity of a test can be improved. The experts will be able to review the items and comment on whether the items cover a representative sample of the behavior domain.

- **Face Validity**

Face validity is an estimate of whether a test appears to measure a certain criterion; it does not guarantee that the test actually measures phenomena in that domain. Face validity is very closely related to content validity. While content validity depends on a theoretical basis for assuming if a test is assessing all domains of a certain criterion (e.g. does assessing addition skills yield in a good measure for mathematical skills? - To answer this you have to know, what different kinds of arithmetic skills mathematical skills include) face validity relates to whether a test appears to be a good measure or not. This judgment is made on the "face" of the test, thus it can also be judged by the amateur. Face validity is a starting point, but should NEVER be assumed to be provably valid for any given purpose, as the "experts" may be wrong.

For example- suppose you were taking an instrument reportedly measuring your attractiveness, but the questions were asking you to identify the correctly spelled word in each list. Not much of a link between the claim of what it is supposed to do and what it actually does.

- **Curricular Validity**

The extent to which the content of the test matches the objectives of a specific curriculum as it is formally described. Curricular validity takes on particular importance in situations where tests

are used for high-stakes decisions, such as Punjab Examination Commission exams for fifth and eight grade students and Boards of Intermediate and Secondary Education Examinations. In these situations, curricular validity means that the content of a test that is used to make a decision about whether a student should be promoted to the next levels should measure the curriculum that the student is taught in schools. Curricular validity is evaluated by groups of curriculum/content experts. The experts are asked to judge whether the content of the test is parallel to the curriculum objectives and whether the test and curricular emphases are in proper balance. Table of specification may help to improve the validity of the test.

2. Construct Validity

Before defining the construct validity, it seems necessary to elaborate the concept of construct. It is the concept or the characteristic that a test is designed to measure. A construct provides the target that a particular assessment or set of assessments is designed to measure; it is a separate entity from the test itself. According to Howell (1992) Construct validity is a test's ability to measure factors which are relevant to the field of study. Construct validity is thus an assessment of the quality of an instrument or experimental design. It says 'Does it measure the construct it is supposed to measure'. Construct validity is rarely applied in achievement test.

Construct validity refers to the extent to which operationalizations of a construct (e.g. practical tests developed from a theory) do actually measure what the theory says they do. For example, to what extent is an IQ questionnaire actually measuring "intelligence"? Construct validity evidence involves the empirical and theoretical support for the interpretation of the construct. Such lines of evidence include statistical analyses of the internal structure of the test including the relationships between responses to different test items. They also include relationships between the test and measures of other constructs. As currently understood, construct validity is not distinct from the support for the substantive theory of the construct that the test is designed to measure. As such, experiments designed to reveal aspects of the causal role of the construct also contribute to construct validity evidence. Construct validity occurs when the theoretical constructs of cause and effect accurately represent the real-world situations they are intended to model. This is related to how well the experiment is operationalized. A good experiment turns the theory (constructs) into actual things you can measure. Sometimes just finding out more about the construct (which itself must be valid) can be helpful. The construct validity addresses the construct that are mapped into the test items, it is also assured either by judgmental method or by developing the test specification before the development of the test. The constructs have some essential properties the two of them are listed as under: 1. Are abstract summaries of some regularity in nature? 2. Related with concrete, observable entities.

For Example - Integrity is a construct; it cannot be directly observed, yet it is useful for understanding, describing, and predicting human behaviour.

- **Convergent Validity**

Convergent validity refers to the degree to which a measure is correlated with other measures that it is theoretically predicted to correlate with. OR Convergent validity occurs where measures of constructs that are expected to correlate do so. This is similar to concurrent validity (which looks for correlation with other tests).

For example, if scores on a specific mathematics test are similar to students scores on other mathematics tests, then convergent validity is high (there is a positively correlation between the scores from similar tests of mathematics).

- **Discriminant Validity**

Discriminant validity describes the degree to which the operationalization does not correlate with other operationalizations that it theoretically should not be correlated with. OR Discriminant validity occurs where constructs that are expected not to relate with each other, such that it is possible to discriminate between these constructs.

For example, if discriminant validity is high, scores on a test designed to assess students skills in mathematics

3. Criterion Validity

Criterion validity evidence involves the correlation between the test and a criterion variable (or variables) taken as representative of the construct. In other words, it compares the test with other measures or outcomes (the criteria) already held to be valid. For example, employee selection tests are often validated against measures of job performance (the criterion), and IQ tests are often validated against measures of academic performance (the criterion). If the test data and criterion data are collected at the same time, this is referred to as concurrent validity evidence. If the test data is collected first in order to predict criterion data collected at a later point in time, then this is referred to as predictive validity evidence.

For example, the company psychologist would measure the job performance of the new artists after they have been on-the-job for 6 months. He or she would then correlate scores on each predictor with job performance scores to determine which one is the best predictor.

4. Concurrent Validity

According to Howell (1992) “concurrent validity is determined using other existing and similar tests which have been known to be valid as comparisons to a test being developed. There is no other known valid test to measure the range of cultural issues tested for this specific group of subjects”. Concurrent validity refers to the degree to which the scores taken at one point correlates with other measures (test, observation or interview) of the same construct that is measured at the same time. Returning to the selection test example, this would mean that the tests are administered to current employees and then correlated with their scores on performance

reviews. This measure the relationship between measures made with existing tests. The existing test is thus the criterion.

For example, a measure of creativity should correlate with existing measures of creativity. For example: To assess the validity of a diagnostic screening test. In this case the predictor (X) is the test and the criterion (Y) is the clinical diagnosis. When the correlation is large this means that the predictor is useful as a diagnostic tool.

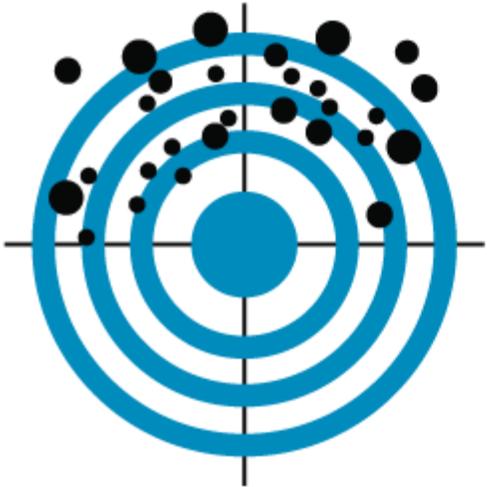
5. Predictive Validity

Predictive validity assures how well the test predicts some future behavior of the examinee. It validity refers to the degree to which the operationalization can predict (or correlate with) other measures of the same construct that are measured at some time in the future. Again, with the selection test example, this would mean that the tests are administered to applicants, all applicants are hired, their performance is reviewed at a later time, and then their scores on the two measures are correlated. This form of the validity evidence is particularly useful and important for the aptitude tests, which attempt to predict how well the test taker will do in some future setting. This measures the extent to which a future level of a variable can be predicted from a current measurement. This includes correlation with measurements made with different instruments. For example, a political poll intends to measure future voting intent. College entry tests should have a high predictive validity with regard to final exam results. When the two sets of scores are correlated, the coefficient that results is called the predictive validity coefficient.

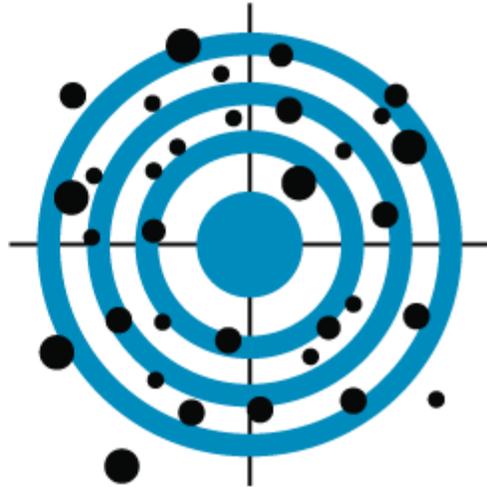
Examples: 1. If higher scores on the Boards Exams are positively correlated with higher G.P.A.'s in the Universities and vice versa, then the Board exams is said to have predictive validity. 2. We might theorize that a measure of math ability should be able to predict how well a person will do in an engineering-based profession.

Relationship between Validity and Reliability

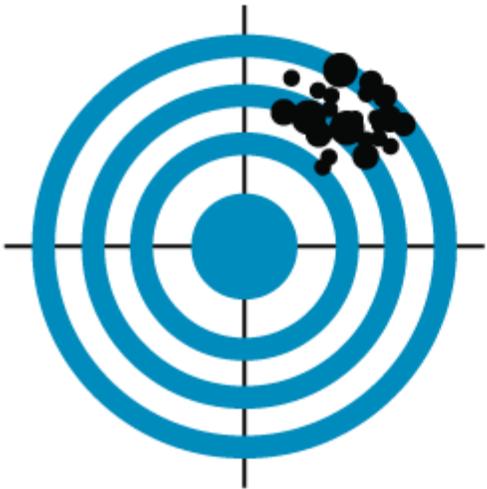
Reliability and validity are two different standards used to gauge the usefulness of a test. Though different, they work together. It would not be beneficial to design a test with good reliability that did not measure what it was intended to measure. The inverse, accurately measuring what we desire to measure with a test that is so flawed that results are not reproducible, is impossible. Reliability is a necessary requirement for validity. This means that you have to have good reliability in order to have validity. Reliability actually puts a cap or limit on validity, and if a test is not reliable, it cannot be valid. Establishing good reliability is only the first part of establishing validity. Validity has to be established separately. Having good reliability does not mean we have good validity, it just means we are measuring something consistently. Now we must establish, what it is that we are measuring consistently. The main point here is reliability is necessary but not sufficient for validity. In short we can say that reliability means noting when the problem is validity.



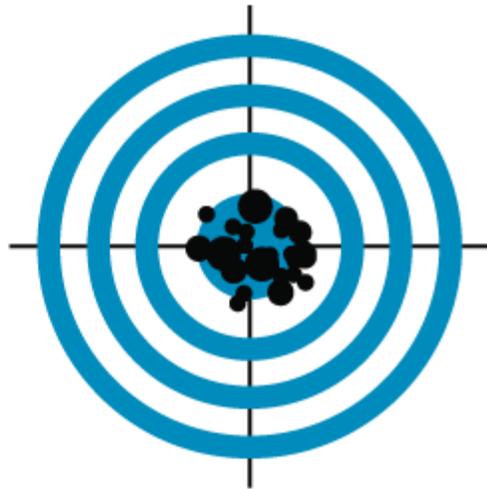
unreliable & invalid



unreliable, but valid



reliable, not valid



both reliable and valid