

6.1.3 Validity versus Reliability

A test can be reliable but may not be valid. If test scores are to be used to make accurate inferences about an examinee's ability, they must be both reliable and valid. Reliability is a prerequisite for validity and refers to the ability of a test to measure a particular trait or skill consistently. In simple words we can say that same test administered to same students may yield same score. However, tests can be highly reliable and still not be valid for a particular purpose. Consider the example of a thermometer if there is a systematic error and it measures five degrees higher. When the repeated readings has been taken under the same conditions the thermometer will yield consistent (reliable) measurements, but the inference about the temperature is faulty.

This analogy makes it clear that determining the reliability of a test is an important first step, but not the defining step, in determining the validity of a test.

There are different methods of assuring the validity of the assessment tools. Some of the important methods namely, content, construct, predictive, and criterion validity are discussed in section 6.4.

6.2 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.

6.2.1 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 behavioural 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 behaviour domain.

For Example - In developing a teaching competency test, experts on the field of teacher training would identify the information and issues required to be an effective teacher and then will choose (or rate) items that represent those areas of information and skills which are expected from a teacher to exhibit in classroom.

Lawshe (1975) proposed that each rater should respond to the following question for each item in content validity:

Is the skill or knowledge measured by this item?

- Essential
- Useful but not essential
- Not necessary

With respect to educational achievement tests, a test is considered content valid when the proportion of the material covered in the test approximates the proportion of material covered in the course.

Activity 6.1: Make a test from any chapter of science book of class 7th and test whether it is valid or not with the reference to its content?

There are different types of content validity; the major types face validity and the curricular validity are as below.

1 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.

Possible Advantage of Face Validity...

- If the respondent knows what information we are looking for, they can use that “context” to help interpret the questions and provide more useful, accurate answers.

Possible Disadvantage of Face Validity...

- If the respondent knows what information we are looking for, they might try to “bend & shape” their answers to what they think we want

Activity 6.2: Make an objective type test and discuss its face validity with at three experts of the subject considering the grade level of the students.

2. 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.

Activity 6.3: Curricular validity affects the performance of the examinees, how can you measure the curricular validity of tests, discuss the current practice followed by the secondary level teachers with two or three SST in your town.

6.2.2 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.

Activity 6.4: Make a tests for a child of class 4th which measures the shyness construct of his personality, and valid this test with reference to its construct validity.

There are different types of construct validity; the convergent and the discriminant validity are explained as follows.

1. 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).

2. 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 should not be positively correlated with scores from tests designed to assess intelligence.

Convergence and discrimination are often demonstrated by correlation of the measures used within constructs. Convergent validity and Discriminant validity together demonstrate construct validity.

6.2.3 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.

Activity 6.5: Administer any test of English to grade 9th and predict the performance of the students for future on the basis of that test. Compare its results after a month with their monthly English test to check the criterion validity of that test with reference to the prediction made about his performance on English language.

6.2.4 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.

6.2.5 Predictive Validity

Predictive validity assures how well the test predicts some future behaviour 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.