# Validity

The validity of any measuring instruments depends upon the accuracy with which it measures what it purports to measure when compared with a standard criterion. A test is valid when the performance which it measures corresponds to the same performance as otherwise independently measured or objectively defined. At this point, a distinction needs to be made between validity and reliability. Suppose, a clock is set forward 30 minutes. If the clock is a good timepiece, the time it "shows" will be reliable, that is, consistent, but will not be valid as judged by "standard time". The reliability of a test is determined by making reported measurements of the same facts; and validity is found by comparing the data obtained from the test with standard (and sometimes arbitrary) measures. Since independent standards (that is, criteria) are hard to get in mental measurement, the validity of a mental test can never be estimated as accurately as can the validity of a physical instrument.

Validity is a relative term. A test is valid for a particular purpose; it is not generally valid.

## Determining Validity by means of Judgements

Content validity becomes more of an issue for tests of achievement or ability and less a concern for tests of personality where high content validity may limit the overall usefulness/applicability of the test. Further, it is useful for tests related with cognitive skills that require an assessment of a broad range of skills in a given area.

The concept of "content validity" is employed in the selection of items for a test. Standard educational achievement examination represent the consensus of many educators as to what a child of a given age or grade should know about arithmetic, reading, spelling, history, and other subjects. A test of English history, for instance, would be valid if its content consists of questions covering this area. The validation of content through competent judgements is most satisfying under two conditions, (i) when the sampling of items is wide and judicious, and (ii) when adequate Standardization groups are utilized.

Less defensible than content validity is the judgement process called "face validity".

A test is said to have face validity when it appears to measure whatever the author had in mind. Content validity is generally confused with "face validity". Face validity does not deal with what a test generally measures but rather with what a scale appears to measure based on the reading of various items. Rating scales for various hypothesized traits, neurotic inventories, attitude scales, and even intelligence tests often claim face validity. Judgement of face validity is very useful in helping an author decide whether his test items are relevant to some specific situation (e.g., the industry) or to specialized occupational experiences. For example, arithmetic problems dealing with bank operations are more relevant to bank jobs than are fictitious problems dealing with men rowing against a river, or the cost of preparing wall. However, face validity should never be more than a first step in testing an item; it should not be the final word.

## Criterion Based Validity

This kind of validity deals with the ability of test scores to predict human behaviour, either with the help of other test scores, observable behaviour, or other accomplishments such as grade point averages.

Experimentally, the validity of a test is determined by finding the correlation between the test and some independent criterion, may be an objective measure of performance, or a quantitative measure such as a judgement of the character or excellence in work done. Intelligence tests were first to be validated against school grades, ratings for aptitude by teachers and other indices of ability. Personality, attitude and interest inventories are validated in a variety of ways. The best way is to check test predictions against actual outcomes. A high correlation between a test and a criterion is evidence of validity, provided that, (i) the criterion was setup independently, and (ii) both the test and the criterion are reliable.

Criterion validity can be categorised into two types — concurrent and predictive. Concurrent validity involves prediction to an alternative method of measuring the same characteristics of interest, while predictive validity attempts to show a relationship with future behaviour.

Both predictive and concurrent validities are accepted by deciding the appropriate level of validity coefficient or correlation between a test score and some criterion variable. The appropriate acceptance level depends upon the intended use of the test. For instance, if one is interested to predict group membership, a classification analysis, or a similar technique that determines placement, accuracy based on test scores would be appropriate. This is known as the non-correlational method of validation.

The index of reliability is sometimes taken as a measure of validity. The correlation coefficient gives the relationship between obtained scores and their theoretical true counterparts. Suppose the reliability coefficient of a test is 0.81, it is $\sqrt{0.81}$ or 0.90. This would mean that the test measures true ability to the extent expressed by $r$ equal to 0.90.

## Construct Validity

Construct validity approach is much more complex than the other forms of validity and is based on accumulation of data over a long period of time. Construct validity requires study of test scores in relationship not only to variables that the test is intended to assess, but also study of those variables that have no relationship to the domain underlying the instrument.

Therefore, one builds a nomothetic net or inferential definition of the characteristics that a test is intended to assess. Another approach includes predictions to other tests that are assumed to measure the same underlying trait as well as those that describe unrelated traits. Hence, we may find or predict that a specific intellectual skill should have a moderate correlation with a test of general IQ, little or no correlation with a measure of hypochondriasis, and a strong correlation to another test assessing the same intellectual skill. One should keep in mind that while examining such interrelationship the efficacy of the research depends on the accuracy of the original hypothesis. This hypothesis is related with the researcher's comprehension of the traits under study. One should be careful and not confuse a researcher's misunderstanding of either the intention of an instrument or the underlying theory with the inefficiency of the instrument itself.

## Factorial Validity

Another method to study construct validity is with the help of factor analysis. One may postulate a factorial structure for a specific test given one's assumptions about both the characteristics that are being assessed and the theory from which they are derived. A confirmatory factor analysis is then performed to test the hypothesis. In case of tests in which a limited number of scores or a single score is generated, other variables with meanings that are more completely understood may be included in the analysis. Factorial relationships with such marker variables can then be used to determine the meaning of the new test scores. In such an analysis and all factor analytic procedures, it is helpful to perform a series of factor analyses in order to determine whether the factor structure and the factorial relationships are stable across time and across groups.

Factor analysis, a specialized statistical technique, is widely used, and is highly important in modern test construction. The intercorrelations of a large number of tests are examined and if possible accounted for in terms of a much

smaller number of more general "factor" or trait categories. The factors presumably run through the often complex abilities measured by the individual tests. It is sometimes found, for example, that three or four factors account for the intercorrelation obtained among 15 or more tests. The validity of a given test is, thus, defined by its factor loading, and these are given by the correlation of the test with each factor. For instance, a vocabulary test may correlate 0.85 with the verbal factor extracted from the entire test battery. This coefficient becomes the test's factorial validity.

Unlike reliability, which is influenced only by unsystematic errors of measurement, the validity of a test is affected by both unsystematic and systematic (constant) errors. This correctly implies that a test may be reliable without being valid, but it cannot be valid without being reliable. Another way of stating the same point is that reliability is a necessary but not a sufficient condition for validity. Technically speaking, the criterion-related validity of a test, as indicated by the correlation between the test and an external criterion measure, can never be greater than the square root of the parallel forms reliability coefficient.

## Factors Affecting Validity

(a) *Group Differences*: The characteristics of a group of people on which the test is validated affect the criterion-related validity. Differences between group of people on variables coefficient between group of people on variables like sex, age and personality traits may affect the correlation coefficient between the test and the selected criteria. Like reliability coefficient, the magnitude of validity coefficient depends on the disagree of heterogeneity of the validation group on the test variable. In a group having a narrower range of test scores that is in a more homogeneous group, the validity coefficient tends to be smaller. Since the size of a correlation coefficient is a function of two variables, a narrowing of the range of either the predictor or the criterion variable will tend to lower the validity coefficient.

*(b) Correction for Attenuation*: An unreliable test cannot be very valid. A test of low reliability also has low validity. There is a formula that can be employed to estimate what the validity coefficient would be if both the test and the criterion are perfectly reliable. This correction for attenuation formula is,

$$r = \frac{V_{12}}{\sqrt{r_{11}}\sqrt{r_{22}}}$$

where $r$ is an estimate of what the validity coefficient would be if the predictor and criterion variables are perfectly reliable.

$V_{12}$ is the validity coefficient,

$r_{11}$ is the reliability of the predictor, and

$r_{22}$ is the reliability of the criterion.

For instance, assume that the validity coefficient is 0.40, the reliability of the test is 0.70, and the reliability of the criterion is 0.50. The validity coefficient would be given as follows if both the test and the criterion are perfectly reliable:

$$V_{12} = 0.40$$

$$r_{11} = 0.70$$

$$r_{22} = 0.50$$

$$r = \frac{V_{12}}{\sqrt{r_{11}}\sqrt{r_{22}}} = \frac{0.40}{\sqrt{0.70}\sqrt{0.50}} = 0.68$$

This shows a substantial increase in validity coefficient.

Therefore, researchers are cautioned at this point about employing the correction for attenuation because tests are never perfectly reliable and validity coefficients that are generally corrected for attenuation do not exist.

(c) *Criterion contamination*: The validity of a test is also dependent upon the validity of the criterion itself as a measure of the particular cognitive or affective characteristic of interest. Sometimes the criterion is contaminated or rendered invalid due to the method by which criterion scores are determined. Teachers have been known to test students scores on academic achievements test (AAT) before deciding what course grades to assign. Since AAT are also taken into consideration by the admission office to select students who are predicted to make satisfactory grades. This method of assigning grades contaminates the criterion and hence results in an inaccurate validity coefficient. Therefore, if AAT scores are to be used for predicting grades then grades should be arrived at independently without reference to AAT scores.

(d) *Test Length*: Like reliability, validity coefficient varies directly as the test length, that is, longer a test the greater its validity, and vice-versa.

Increasing a test's length effects the validity coefficient. This effect can be measured by the following formula:

$$V_n = \frac{K V_0}{\sqrt{K + K(K-1)\,\bar{r}}} \qquad (1)$$

where,

$V_n$ = the validity of the lengthened test

$V_o$ = the validity of the original test

$r$ = the reliability coefficient of the test

$K$ = number of parallel forms of test $X$, or the number of times it is lengthened. Let us explain with the help of an example:

# What is validity in research?

Validity is how researchers talk about the extent that results represent reality. Research methods, quantitative or qualitative, are methods of studying real phenomenon – validity refers to how much of that phenomenon they measure vs. how much "noise," or unrelated information, is captured by the results.

Validity and reliability make the difference between "good" and "bad" research reports. Quality research depends on a commitment to testing and increasing the validity as well as the reliability of your research results.

Any research worth its weight is concerned with whether what is being measured is what is intended to be measured and considers the ways in which observations are influenced by the circumstances in which they are made.

The basis of how our conclusions are made play an important role in addressing the broader substantive issues of any given study.

For this reason we are going to look at various validity types that have been formulated as a part of legitimate research methodology.

Here are the 7 key types of validity in research:

1. Face validity
2. Content validity
3. Construct validity
4. Internal validity
5. External validity
6. Statistical conclusion validity
7. Criterion-related validity

# 1. Face validity

Face validity is how valid your results seem based on what they look like. This is the least scientific method of validity, as it is not quantified using statistical methods.

Face validity is not validity in a technical sense of the term.  It is concerned with whether it seems like we measure what we claim.

Here we look at how valid a measure appears on the surface and make subjective judgments based off of that.

For example,

- Imagine you give a survey that appears to be valid to the respondent and the questions are selected because they look valid to the administer.
- The administer asks a group of random people, untrained observers,  if the questions appear valid to them

In research it's never enough to rely on face judgments alone – and more quantifiable methods of validity are necessary in order to draw acceptable conclusions.  There are many instruments of measurement to consider so face validity is useful in cases where you need to distinguish one approach over another.

# 2. Content validity

Content validity is whether or not the measure used in the research covers all of the content in the underlying construct (the thing you are trying to measure).

This is also a subjective measure, but unlike face validity we ask whether the content of a measure covers the full domain of the content. If a researcher wanted to measure introversion, they would have to first decide what constitutes a relevant domain of content for that trait.

Content validity is considered a subjective form of measurement because it still relies on people's perception for measuring constructs that would otherwise be difficult to measure.

Where content validity distinguishes itself (and becomes useful) is through its use of experts in the field or individuals belonging to a target population.  This study can be made more objective through the use of rigorous statistical tests.

For example you could have a content validity study that informs researchers how items used in a survey represent their content domain, how clear they are, and the extent to which they maintain the theoretical factor structure assessed by the factor analysis.

# 3. Construct validity

A construct represents a collection of behaviors that are associated in a meaningful way to create an image or an idea invented for a research purpose. Construct validity is the degree to which your research measures the construct (as compared to things outside the construct).

Depression is a construct that represents a personality trait which manifests itself in behaviors such as over sleeping, loss of appetite, difficulty concentrating, etc.

The existence of a construct is manifest by observing the collection of related indicators. Any one sign may be associated with several constructs.  A person with difficulty concentrating may have A.D.D. but not depression.

Construct validity is the degree to which inferences can be made from operationalizations (connecting concepts to observations) in your study to the constructs on which those operationalizations are based.  To establish construct validity you must first provide evidence that your data supports the theoretical structure.

You must also show that you control the operationalization of the construct, in other words, show that your theory has some correspondence with reality.

- **Convergent Validity –** the degree to which an operation is similar to other operations it should theoretically be similar to.
- **Discriminative Validity** -– if a scale adequately differentiates itself or does not differentiate between groups that should differ or not differ based on theoretical

- **Convergent Validity –** the degree to which an operation is similar to other operations it should theoretically be similar to.
- **Discriminative Validity** -– if a scale adequately differentiates itself or does not differentiate between groups that should differ or not differ based on theoretical reasons or previous research.
- **Nomological Network –** representation of the constructs of interest in a study, their observable manifestations, and the interrelationships among and between these.  According to Cronbach and Meehl, a nomological network has to be developed for a measure in order for it to have construct validity

- **Multitrait-Multimethod Matrix –** six major considerations when examining Construct Validity according to Campbell and Fiske. This includes evaluations of the convergent validity and discriminative validity. The others are trait method unit, multi-method/trait, truly different methodology, and trait characteristics.

# 4. Internal validity

Internal validity refers to the extent to which the independent variable can accurately be stated to produce the observed effect.

If the effect of the dependent variable is only due to the independent variable(s) then internal validity is achieved. This is the degree to which a result can be manipulated.

# 5. External validity

External validity refers to the extent to which the results of a study can be generalized beyond the sample. Which is to say that you can apply your findings to other people and settings.

Think of this as the degree to which a result can be generalized. How well do the research results apply to the rest of the world?

A laboratory setting (or other research setting) is a controlled environment with fewer variables. External validity refers to how well the results hold, even in the presence of all those other variables.

# 6. Statistical conclusion validity

Statistical conclusion validity is a determination of whether a relationship or co-variation exists between cause and effect variables.

This type of validity requires:

- Ensuring adequate sampling procedures
- Appropriate statistical tests
- Reliable measurement procedures

This is the degree to which a conclusion is credible or believable.

# 7. Criterion-related validity

Criterion-related validity (also called instrumental validity) is a measure of the quality of your measurement methods  The accuracy of a measure is demonstrated by comparing it with a measure that is already known to be valid.

In other words – if your measure has a high correlation with other measures that are known to be valid because of previous research.

For this to work you must know that the criterion has been measured well.  And be aware that appropriate criteria do not always exist.

What you are doing is checking the performance of your operationalization against a criteria.

The criteria you use as a standard of judgment accounts for the different approaches you would use:

- **Predictive Validity –** operationalization's ability to predict what it is theoretically able to predict. The extent to which a measure predicts expected outcomes.
- **Concurrent Validity –** operationalization's ability to distinguish between groups it theoretically should be able to. This is where a test correlates well with a measure that has been previously validated.

When we look at validity in survey data we are asking whether the data represents what we think it should represent.

When we look at validity in survey data we are asking whether the data represents what we think it should represent.

We depend on the respondent's mind set and attitude in order to give us valid data.

In other words we depend on them to answer all questions honestly and conscientiously. We also depend on whether they are able to answer the questions that we ask. When questions are asked that the respondent can not comprehend or understand, then the data does not tell us what we think it does.