

The Validity & Reliability of Assessment Methods

Assessment methods and tests should have validity and reliability data and research to back up their claims that the test is a sound measure.

Reliability is a very important concept and works in tandem with Validity. A guiding principle for psychology is that a test can be reliable but not valid for a particular purpose, however, a test cannot be valid if it is unreliable.

Validity

Assessment methods including personality questionnaires, ability assessments, interviews, or any other assessment method are valid to the extent that the assessment method measures what it was designed to measure.

There are different aspects of validity and they differ in their focus. The aspects of validity that have an impact on the actual scientific application of the assessment are concurrent validity, predictive validity and construct validity. The two less relevant aspects of validity are face and content validity.

The three aspects of validity that do have an impact on the practical usefulness of the assessment method are as follows:

Construct validity is the theoretical focus of validity and is the extent to which performance on the test fits into the theoretical scheme and research already established on the attribute or construct the test is trying to measure.

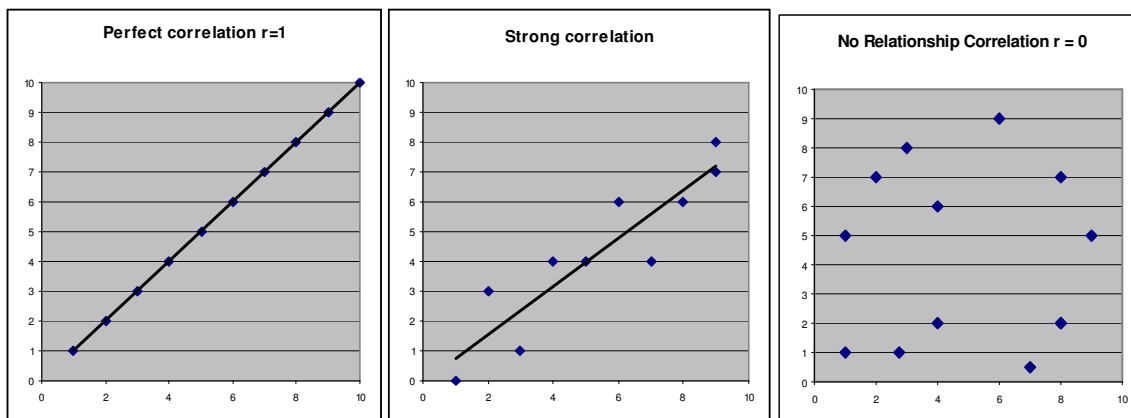
- ◆ In essence, it is the extent to which a test fits into the wider research picture the more we are able to confer construct validity onto the assessment method or test.
- ◆ Typically test makers research data from the same participants on a number of tests attempting to measure similar constructs.
- ◆ Creating a picture of construct validity can take considerable time and complex statistical analysis such as factor analysis.

Concurrent validity is the relationship between test scores and some criterion measure of job performance or training performance at the same time.

- ◆ Both the test scores and the job aspect being measured being collected at the same time.
- ◆ This type of validity is usually used with internal employees and can be useful to assess skill status and future training requirements.

Predictive validity (Criterion Related Validity) is the extent to which a test or questionnaire predicts some future or desired outcome, for example work behaviour or on-the-job performance. This validity has obvious importance in personnel selection, recruitment and development.

- ◆ Predictive validity is of particular interest to psychologists and HR professionals as it allows us to extrapolate the results of the test taken today to a meaningful outcome of what we want to know about the future behaviour of an employee.
- ◆ Predictive validity is usually measured by the correlation between the test score and some appropriate criterion. The criterion could be performance on the job, training performance, counter-productive behaviours, manager ratings on competencies or any other outcome that can be measured.
- ◆ A validity coefficient (denoted by r) is a correlation between a test score and some criterion measure (such as performance).
- ◆ A test may appreciably improve predictive efficiency if it shows any significant correlation with the criterion, however low.
- ◆ The significance of the correlation is a measurement of the probability that the relationship between the two sets of data is due to chance.
- ◆ A larger pool of related data decreases the probability that a correlation is due to chance and therefore the cut-off for the level of a significant r is usually less as the sample size increases.

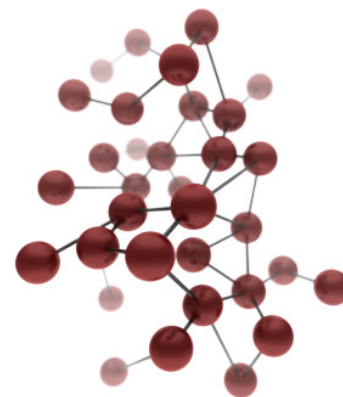


Face validity of a test or method concerns the look and feel of the assessment items and whether an applicant can see any relevance of the test or assessment method to the job or role concerned.

- ◆ Whilst a test with high face validity may make the person taking the test feel more comfortable with the test as it seems related to the job or role, it is not related at all to the test being a good measure or sound test.
- ◆ High Face Validity does not in any way infer that the test is actually predictive of something useful, like on the job performance.
- ◆ Unfortunately, some test makers push that high face validity is ideal, without pointing out the drawbacks of high face validity.
- ◆ The drawback has a serious impact on the information you get from the individual and its usefulness - it makes the test or assessment much easier to fake or manipulate.

Content validity of a test is concerned with how well a test samples the behavioural domain it is trying to measure.

- ◆ For example should you want to measure general numerical ability and the test items were only multiplication equations, this would have poor content validity as the items are not representative of all the aspects that make up general numerical ability.
- ◆ Often a detailed job analysis is required to establish content validity and this is something that is usually not available.
- ◆ It also becomes more complicated for more complex constructs such as intelligence and self esteem, as it is not easy to decide on the criteria that constitute content validity. So, like face validity, content validity is not the main focus of psychologists.



The following table summarises some of the general research findings around the predictive validity of the different selection methods available:

Assessment Method	Predictive Validity
Assessment Centres (multiple methods)	0.65
Behavioural Interviews	0.4 – 0.6
Work-sample Tests	0.54
Ability Tests	0.53
Modern Personality Tests	0.39
Biographical data	0.38
References	0.23
Traditional Interviews	0.05 – 0.19

Source: British Psychological Society/Accord Group

Assessment method	Predictive validity	Criterion measure
Integrity Tests	0.58	counter-productive work behaviour
Integrity Tests	0.51	Overall job performance



Where assessment expertise is part craft and part science

Source – Comprehensive meta-analysis of integrity test validities by Ones, Viswesvaran & Schmit (1993)

Test Length and Validity

The following table illustrates how validities increase as test length increases. The calculations are based upon typical reliability and validity figures of .70 and .40 respectively for a 5 minute test. The difference in validity between a 5 minute test and a test of infinite length is only a .078 difference (.478-.400).

Test Time (Minutes)	Validity
1	.270
2	.332
3	.365
4	.386
5	.400
6	.410
7	.410
8	.418
9	.424
10	.430
11	.434
12	.437
13	.440
14	.443
15	.445
Test of infinite length	.478

Multiple Assessments effects on Predictive Validity

When we combine assessments in a battery we can increase the validity of the testing if the tests are of approximately the same validity and have low inter-correlations. Guilford & Fruchter (1978) summed up the different effects of lengthening tests and including more tests in a battery as follows:

- ◆ “In general, if there is a choice between lengthening of tests in a battery to make them more reliable and adding more tests of different kinds that contribute unique valid variances, the decision should certainly go to the second alternative.”
- ◆ We can therefore increase the validity of testing by using a battery of different assessments or methods. This also explains why assessment centres that have multiple measures then do have higher validity.

Reliability of Assessment Tools & Methods

An aptitude or personality assessment needs to measure each factor it is attempting to measure reliably, for the given population (e.g., customer service applicants, males, females).

Reliability is the consistency or precision with which the test or assessment method measures what it claims to measure.

- ◆ Any assessment method or test needs to be a consistent measure – this means if the test was used repeatedly on the same candidate it would produce similar results.
- ◆ Assessments or tests with lower reliability are of little practical use. Reliability is a very important concept and works in tandem with validity.
- ◆ A guiding principle for psychology is that a test can be reliable but not valid for a particular purpose, however, a test cannot be valid if it is unreliable.
- ◆ Most psychological test makers provide a reliability coefficient which establishes the reliability of the test and is usually based on test-retest methodology or split half technique (otherwise known as internal consistency reliability).
- ◆ The reliability coefficient is used to set up a band for error around individual scores that is acceptable and renders the results reliable.

Test retest reliability is when the same test is administered to a sample group of people twice.

- ◆ The limitations of this method are the impact on performance of any information the subject remembers or has learnt from the first testing session that may impact how well they answer the test the second time around.
- ◆ To get around this issue, many test developers design an alternative form of the test and administer this – this second form however does have to measure the issue exactly as the first measure does for reliability to be assessed.

Alternatively, reliability is measured through a **split half technique**.

- This is where subjects may have half of their test answers correlated with the other half (say all odd numbered items compared to all even numbered items).
- If both halves correlate highly with each other, the test is considered reliable.



Test reliability is also represented by a correlation coefficient (r). As with validity coefficients, the closer the correlation coefficient is to 1 the better. While many personality tests are considered to have acceptable levels of reliability if they have reliability coefficients greater than $r=0.7$, ability tests should have reliability coefficients greater than $r=0.8$.