**Reliability**

‘Reliability’ of any research is the degree to which it gives an accurate score across a range of measurement. It can thus be viewed as being ‘repeatability’ or ‘consistency’. In summary:

1. [**Test-retest**](http://changingminds.org/explanations/research/design/types_reliability.htm#tes)**:** Same people, different times. (The same test over **time**)
2. [**Inter-rater**](http://changingminds.org/explanations/research/design/types_reliability.htm#int)**:** Different people, same test. (The same test conducted by different **people**)
3. [**Parallel-forms**](http://changingminds.org/explanations/research/design/types_reliability.htm#par)**:** Different people, same time, different test. (**Different versions** of a test which are designed to be equivalent)
4. [**Internal consistency**](http://changingminds.org/explanations/research/design/types_reliability.htm#int)**:** Different questions, same construct. (The **individual items** of a test)
5. **Test-Retest Reliability (also called Stability)**

An assessment or test of a person should give the same results whenever you apply the test.

Test-retest reliability evaluates reliability across *time*.

Reliability can vary with the many factors that affect how a person responds to the test, including their mood, interruptions, time of day, etc. A good test will largely cope with such factors and give relatively little variation. An unreliable test is highly sensitive to such factors and will give widely varying results, even if the person re-takes the same test half an hour later.

Generally speaking, the longer the delay between tests, the greater the likely variation. **Better tests will give less retest variation with longer delays.** Of course the problem with test-retest is that people may have learned and that the second test is likely to give different results. This method is particularly used in experiments that use a no-treatment control group that is measure pre-test and post-test.

**Examples**

Various questions for a personality test are tried out with a class of students over several years. This helps the researcher determine those questions and combinations that have better reliability. In the development of national school tests, a class of children is given several tests that are intended to assess the same abilities. A week and a month later, they are given the same tests. With allowances for learning, the variation in the test and retest results are used to assess which tests have better test-retest reliability.

**2.** [**Inter-rater**](http://changingminds.org/explanations/research/design/types_reliability.htm#int) **Reliability**

When **multiple people** are giving assessments of some kind or are the subjects of some test, then similar people should lead to the **same resulting scores**. It can be used to calibrate people, for example those being used as observers in an experiment.

Inter-rater reliability thus evaluates reliability *across different people*.

Two major ways in which inter-rater reliability is used are (a) testing how similarly people *categorize*items, and (b) how similarly people *score*items.

This is the best way of assessing reliability when you are using observation, as observer bias very easily creeps (steals) in. It does, however, assume you have multiple observers, which is not always the case. Inter-rater reliability is also known as ***inter-observer reliability* or *inter-coder reliability*.**

### Examples

Two people may be asked to categorize pictures of animals as being dogs or cats. A perfectly reliable result would be that they both classify the same pictures in the same way. Observers being used in assessing prisoner stress are asked to assess several ‘dummy’ people who are briefed to respond in a programmed and consistent way. The variation in results from a standard gives a measure of their reliability.

In a test scenario, an IQ test applied to several people with a true score of 120 should result in a score of 120 for everyone. In practice, there will be usually some variation between people.

## 3. Parallel-Forms Reliability

One problem with questions or assessments is knowing what questions are the best ones to ask. A way of discovering this is do two tests in parallel, using different questions.

Parallel-forms reliability evaluates different questions and question sets that seek to assess the same construct.

Parallel-Forms evaluation may be done in combination with other methods, such as *Split-half*, which divides items that measure the same construct into two tests and applies them to the same group of people.

### Examples

An experimenter develops a large set of questions. They split these into two and administer them each to a randomly-selected half of a target sample.

In development of national tests, two different tests are simultaneously used in trials. The test that gives the most consistent results is used, whilst the other (provided it is sufficiently consistent) is used as a backup.

## 4. Internal Consistency Reliability

When asking questions in research, the purpose is to assess the response against a given construct or idea. Different questions that test the same construct should give consistent results.

Indicates the extent to which items on a test measure the same thing. Internal consistency reliability evaluates individual questions in comparison with one another for their ability to give consistently appropriate results.

*Average inter-item correlation* compares correlations between all pairs of questions that test the same construct by calculating the mean of all paired correlations. ***Average item total correlation***takes the average inter-item correlations and calculates a total score for each item, then averages these. ***Split-half correlation*** divides items that measure the same construct into two tests, which are applied to the same group of people, then calculates the correlation between the two total scores. ***Cronbach's alpha*** calculates an equivalent to the average of all possible split-half correlations and is calculated thus:

a = (N . r-bar) / (1 + (N-1) . r-bar)

Where N is the number of components, and r-bar is the average of all Pearson correlation coefficients.

**Factors Affecting Reliability**

1. Test length
2. Test-retest interval
3. Variability of scores
4. Guessing

Variation within the test situation

### Standard Error of Measurement

Test manuals report a statistic called the **standard error of measurement (SEM)**. It gives the margin of error that you should expect in an individual test score because of imperfect reliability of the test. The SEM represents the degree of confidence that a person’s “true” score lies within a particular range of scores. For example, an SEM of “2” indicates that a test taker’s “true” score probably lies within 2 points in either direction of the score he or she receives on the test. This means that if an individual receives a 91 on the test, there is a good chance that the person's “true” score lies somewhere between 89 and 93. The SEM is a useful measure of the accuracy of individual test scores. The smaller the SEM, the more accurate the measurements.

**Validity**

Validity refers to the degree to which an item is measuring what it’s actually supposed to be measuring.

1. [**Construct validity**](https://www.scribbr.com/methodology/types-of-validity/#construct-validity): Does the test measure the concept that it’s intended to measure?
2. [**Content validity**](https://www.scribbr.com/methodology/types-of-validity/#content-validity): Is the test fully representative of what it aims to measure?
3. [**Face validity**](https://www.scribbr.com/methodology/types-of-validity/#face-validity): Does the content of the test appear to be suitable to its aims?
4. [**Criterion validity**:](https://www.scribbr.com/methodology/types-of-validity/#criterion-validity) Do the results correspond to a different test of the same thing?
5. [**Construct validity**](https://www.scribbr.com/methodology/types-of-validity/#construct-validity) Construct validity evaluates whether a measurement tool really represents the thing we are interested in measuring. It’s central to establishing the overall validity of a method.

### What is a construct?

A construct refers to a concept or characteristic that can’t be directly observed, but can be measured by observing other indicators that are associated with it.

Constructs can be characteristics of individuals, such as intelligence, obesity, job satisfaction, or depression; they can also be broader concepts applied to organizations or social groups, such as gender equality, corporate social responsibility, or freedom of speech.

### What is construct validity?

Construct validity is about ensuring that the method of measurement matches the construct you want to measure. If you develop a questionnaire to diagnose depression, you need to know: does the questionnaire really measure the construct of depression? Or is it actually measuring the respondent’s mood, self-esteem, or some other construct?

To achieve construct validity, you have to ensure that your indicators and measurements are carefully developed based on relevant existing knowledge. The questionnaire must include only relevant questions that measure known indicators of depression. The other types of validity described below can all be considered as forms of evidence for construct validity.

1. **Content validity**

Content validity assesses whether a test is representative of all aspects of the construct.

To produce valid results, the content of a test, survey or measurement method must cover all relevant parts of the subject it aims to measure. If some aspects are missing from the measurement (or if irrelevant aspects are included), the validity is threatened.

1. **Face validity**

It is the extent to which the measurement method appears “on its face” to measure the construct of interest. Face validity considers how suitable the content of a test seems to be on the surface. It’s similar to content validity, but face validity is a more informal and subjective assessment.

1. **Criterion validity**

Criterion validity evaluates how closely the results of your test correspond to the results of a different test.

### What is a criterion?

The criterion is an external measurement of the same thing. It is usually an established or widely-used test that is already considered valid.

### What is criterion validity?

To evaluate criterion validity, you calculate the [correlation](https://www.scribbr.com/methodology/correlational-research/) between the results of your measurement and the results of the criterion measurement. If there is a high correlation, this gives a good indication that your test is measuring what it intends to measure.

Some other types of validity are hereunder:

**External Validity:** It is the extent to which the results of a research study can be generalized to different situations, different groups of people, different settings, different conditions, etc.

**Internal Validity:** It is basically the extent to which a study is free from flaws and that any differences in a measurement are due to an independent variable and nothing else.

**Factors Affecting Validity**

**1. History:-** Events that occur besides the treatment (events in the environment).

**2. Maturation:-** Physical or psychological changes in the participants.

**3. Testing:-** Effect of experience with the pretest - become test wise.

**4. Instrumentation:-** Learning gain might be observed from pre to posttest simply due to nature of the instrument. Particularly a problem in observation studies when observers more likely to give ratings based on expectations (conscious or subconscious).

**5. Statistical Regression:-** Tendency for participants whose scores fall at either extreme on a variable to score nearer the mean when measured a second time.

**6. Differential Selection:-** Effect of treatment confounded with other factors because of differential selection of participants, problem in non-random samples.

**7. Experimental Mortality:-** Participants lost from the study, attrition.