



A Testing Brief
Reliability

Timothy R. Vansickle, Ph.D.



Reliability

If one talks about a reliable person or that a car is reliable, we all tend to know what is meant. When reliability is attached to a test, do we really understand what that means?

Generally, the term reliability, when used in testing, has a very similar meaning as it does when used with cars or people. That is, the more reliable a test, the more consistent a person's test score will be. However, when we think of a reliable car, we typically think it starts all the time, it runs well, and the brakes work all the time. Some of us also want the radio to work and the AC or heat to be functioning all the time too. Reliability of tests is a bit different than that.

One way to think of reliability of a test score is in the amount of error associated with a test score. For now, let us ignore the fact that reliability and the associated error for a test varies along the score scale, and focus on the general concept of reliability and associated error. In the most general form, a person's true score on a test is the sum of their observed score and error (error can be either positive or negative). Thus the observed score may be greater or less than a person's true score.

$$T_{score} = O_{score} + E_{error}$$

You may have taken an IQ test either as part of your school experiences or as part of applying for a job; or maybe you took one of those online IQ tests. Typically, you receive a score in the 100s. For example, you might have received an IQ score of 120 on one of the major IQ tests. This is a norm-referenced score that has a mean of 100 and a standard deviation of 15 or 16. So, your score of 120 is 1+ standard deviations above the mean. That is, you are above average in intelligence. If you took the same IQ test multiple times you would receive a different score each time but all the scores would be somewhere near or around the score of 120 (assuming 120 is your true intelligence). For most of the IQ tests in production the reliability statistics are quite good (e.g., high .9s). Still the standard error of measure for most IQ tests is about 7 scale score points. That is, we have 68 percent confidence that your IQ is somewhere between 113 and 127, and we have 95 percent confidence that your IQ is between 106 and 133. This is an example of the amount of error associated with your IQ score.

What does all of that mean? Well, the more reliable a test, the smaller the error associated with the test score; or put another way, the more confidence we have in the test score being the correct or true estimate of the person's trait being tested. In our work, it is the amount of learning or achievement demonstrated by a student. Or in the case of your IQ score, it is your general mental ability, in which case you are above average!

A Testing Briefing
Reliability

There are a number of factors that can make tests more or less reliable. One such factor is test length. Generally, the longer the test, the more reliable the test will be. Another factor is the stability of the trait being measured. For example, measurement of your interests is not as stable a trait as is intelligence. Hence, the reliability of measures of your interests is less reliable than are those of intelligence, relatively speaking. Do not take this to mean measures of interest, especially vocational, are unreliable; rather they tend to have more error associated with the scores because interests in most folks is not a permanent condition (i.e., it is readily changeable). Another factor that can impact the reliability of a test is the opportunity to learn the content of the test. If students are not given ample opportunity and/or good instruction to learn the content that will be covered on the test, the amount of error in student responding will increase and test reliability will suffer.

There are a number of types of test reliability. Each controls for or estimates different types of error associated with a test. The following table provides a quick overview of the types of reliability and characteristics for each type.

Types and Characteristics of Reliability			
Types of Reliability	Number of Test Administrations	Number of Test Forms	Type of Error Variance
Test Retest	2	1	Stability Over Time
Alternate From			
Immediate	1	2	Content Stability
Delayed	2	2	Content Stability & Stability Over Time
Split Half	1	1	Content Stability
Internal Consistency	1	1	Content Stability - Heterogeneity
Score/Inter-rater	1	1	Stability Across Scorers/Raters

As can be seen in the table, Test Retest reliability requires two test administrations of the same test form across a specified period of time (e.g., two weeks) while Internal Consistency requires only one form administered at one time point.

Reliability is typically shown as a correlation coefficient. Correlation coefficients have a range of values from -1.00 to 1.00. Reliability coefficients

A Testing Briefing
Reliability

are typically positive values and should be in the high .8s or greater. In most state testing programs, internal consistency is the primary method of computing reliability along with scorer/inter-rater reliability when constructed-response items are included on the test.