Test Reliability Indicates More than Just Consistency

by Dr. Timothy Vansickle
April 2015

Introduction

Reliability is the extent to which an experiment, test, or measuring procedure yields the same results on repeated trials.[1] A reliable car, for example, works well consistently. It starts every time and has trustworthy brakes and tires. A reliable test is similar in that it works well consistency by producing consistent results.

A reliable test also considers random error, which results from outside influences that can affect a student's test score. A test that produces highly consistent, stable results (i.e., mostly free from random error) is considered highly reliable. The less random error, the more reliable a test is. The more reliable a test, the more consistent a student's test score will be.

The term "reliability," much like validity, has a very technical meaning that is often misconstrued in the assessment industry by the media and public. This assessment brief tries to explain reliability in simple terms, but keep in mind that analyzing reliability requires a lot of psychometric and statistical expertise. Claiming that a test is reliable and valid requires proof found through psychometric analyses and a body of evidence that supports the use of the test for specific purposes.

Random Error

Outside influences that cause random error include guessing, illness, direction errors, and mistakes when marking an answer. For example, a student who has a bad cold might not do as well on a test compared to if he or she was not sick at all. The random error stems from the student not answering to the best of his or her ability due to the cold.

These factors or outside influences are called measurement error (or errors of measurement), which can be reduced by using clear and unambiguous items, standardizing testing conditions, and including a sufficient number of items in the test. Increasing reliability by minimizing error is an important goal for any test.

Random error is associated with classical test theory, a psychometric model used to analyze and report test scores. In classical test theory, the theoretical true score reflects an error-free assessment of a student's ability. However, because obtaining a true score is impossible because of random error, a student's observed score is used instead when evaluating results:

$$X = T + e$$

where X is the observed score, T is the true score, and e is the measurement error.

1 reliability. In *Merriam-Webster.com*. Retrieved June 9, 2014, from http://www.merriam-webster.com/dictionary/reliability.

Measuring a Test's Reliability

Reliability is typically shown as a reliability coefficient created in a calculation to determine the reliability, or consistency, of scores, such as a measure of the amount of consistency between two sets of scores from different administrations from the same group of students. There are a number of reliability measures, but all of them produce a value that is similar to a correlation.

Correlation coefficients range from –1.0 to 1.0. Reliability coefficients are usually positive values and should be in the high 0.8s or higher. The higher the value of a reliability coefficient, the greater the reliability of the test will be.

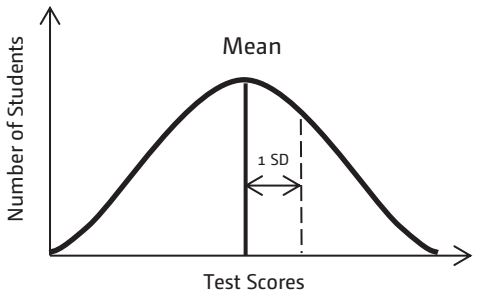
Reliability measurements for an assessment program typically include the following:

- Standard deviation or variance
- Standard error of measurement (SEM)
- Internal consistency (e.g., Cronbach's Alpha)
- Inter-rater reliability for handscored items (e.g., the kappa coefficient)
- Classification accuracy and consistency

Standard Deviation

The standard deviation is a measure of a test's variability that describes the distribution of scores, as shown in Figure 1. The further the scores are away from each other in value, the higher the standard deviation and, thus, the higher the test score variation. If all of a test's scores are the same, the standard deviation is 0. If there is no variation in scores, the reliability of the test goes down. Conversely, if the variation is too high, the reliability can also decrease.

Figure 1. Standard Deviation (SD)



Standard Error of Measurement (SEM)

Because perfect measurement of ability does not exist due to error, it is important to analyze the amount of measurement error on any assessment. Error can still exist even if the reliability statistics are high (i.e., close to 1.0), as indicated by the standard error of measurement (SEM).

The SEM is estimated using information about the reliability and standard deviation for a set of test scores, and it represents the amount of variability that can be expected in a student's test score due to the inherent imprecision of the test (e.g., if the student were tested again, he or she would likely obtain a slightly different score).

The smaller the SEM (close to 0), the greater the accuracy of the scores will be and, thus, the greater reliability of the scores and the more precise the estimate of the student's true ability. Measurement error is commonly expressed in terms of standard deviation, and the SEM can be thought of as the standard deviation of the distribution of measurement error.

The SEM is calculated as follows:

$$SEM = SD\sqrt{1 - r_{xx}} \Leftrightarrow s_e = s_x\sqrt{1 - \frac{s_t^2}{s_x^2}}$$

where *SEM* (= s_e) refers to the standard error of measurement, *SD* (= s_x) is the standard deviation unit of the scale for a test, r_{xx} is the reliability coefficient for a sample test (or estimate of ρ_{XX}, which is a population reliability coefficient), s_t^2 is the estimate of σ_T^2, and s_x^2 is the estimate of σ_X^2.

Internal Consistency

Internal consistency (a reliability measure) measures whether different items on an assessment measure the same general construct[2] and produce similar scores.

Cronback's Alpha Cronbach's Alpha (α), first named as alpha by Lee Cronbach in 1951[3], is the most commonly used measure of internal consistency. Cronbach's Alpha ranges from 0.0 to 1.0, where 1.0 refers to a perfectly consistent test. Tests are typically considered of sound reliability when Cronbach's alpha (a type of reliability coefficient) ranges from 0.8 and above. Cronbach's Alpha is calculated as follows:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}\right)$$

2 A construct is an attribute or ability that cannot be directly measured, such as intelligence. More generally, it is what tests are designed to measure.

3 Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(3)*, 297–334.

Inter-rater Reliability

Inter-rater reliability is a measure of a scorer's consistency in handscoring constructed-response items. It also measures the consistency of two scorers in regards to scoring a student's constructed-response item.[4] Inter-rater reliability usually refers to the degree of agreement between the scorers of an item for the same student.

Statistical calculations of inter-rater reliability are represented by the percent of exact and adjacent scores, as well as other statistical procedures that measure reliability in the scoring process: the kappa coefficient and intraclass correlation. Kappa is easily interpreted and most often used.

Kappa Coefficient Kappa statistics typically ranges from 0.0 to 1.0, with 1.0 indicating perfect agreement. On the rare occasions that Kappa is negative, it indicates that the two ratings agreed less than would be expected by chance. One way to interpret Kappa was provided in Altman (1991)[5]:

- Poor agreement = less than 0.20
- Fair agreement = 0.20 to 0.40
- Moderate agreement = 0.40 to 0.60
- Good agreement = 0.60 to 0.80
- Very good agreement = 0.80 to 1.00

Classification Accuracy & Classification Consistency

Classification refers to any group or category students are placed in based on test scores, such as performance levels (e.g., Below Basic, Basic, Proficient, and Advanced). The less reliable a test, the more error there is in the classification of students into these groups or categories.

Classification accuracy refers to the accuracy of decisions (e.g., the accuracy of students' assignments to those performance levels), or the extent to which decisions would agree with those that would be made if each student could somehow be tested with all possible forms of the assessment.

Classification consistency refers to the consistency of decisions, or the extent to which decisions would agree with those that would have been made if the students had taken a parallel form of the assessment that is equal in difficulty and covers the same content as the form the students actually took. In other words, classification consistency is the degree of consistency of student performance between parallel forms of a test. It is important to analyze regarding cut scores[6] that have decisions

4 A constructed-response item is a test question, often worth more than one point, that requires students to produce their own written or verbal response rather than select a response from given options.

5 Altman, D. G. (1991). *Practical statistics for medical research*. Chapman and Hall/CRC. 404.

6 A cut score is the minimum score a student must get on a test in order to be placed in a certain performance level such as Below Basic, Proficient, and Advanced. Cut scores usually differ by content area and are determined between two adjacent performance levels (e.g., the cut score between the Basic and Proficient performance levels is the minimum score a student must get in order to be placed in the Proficient performance level).

associated with those scores. Classification consistency looks at the amount of error at the cut score: the smaller the error, the more consistent the classification of the student to a given performance level.

Reliability Example: IQ Test

An intelligence test, or IQ test (IQ = intelligence quotient), typically has a mean score of 100 and a standard deviation of 15. Therefore, a score of 130 is 2 standard deviations above the mean, which indicates the person receiving that score is above average in intelligence. If the same person took the same IQ test multiple times, he or she would receive a different score each time but all the scores would be somewhere near or around the score of 130 (assuming 130 is his or her true intelligence).

Most IQ tests have good reliability (e.g., reliability statistics in the high 0.9s), but the SEM is still about 7 points. Therefore, if a person had an IQ of 130, his or her score would be between 123 and 137 points 68% of the time (i.e., there is 68% confidence that the person's IQ is somewhere between 123 and 137) and between 116 and 144 points 95% of the time. The person is still above average, but because all measurement has error, a confidence interval might be put around the obtained score to better understand where the person's actual true score may fall on the IQ scale.

Other Factors That Affect Reliability

Other factors such as test length can make tests more or less reliable. 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 a student's interests is not as stable a trait as his or her intelligence—thus, measurements of the student's interests is less reliable than measurements of his or her intelligence. This does not mean that measures of interest are unreliable; rather, they tend to have more error associated with the scores because interests are changeable.

Another factor that can affect the reliability of a test is the opportunity to learn the test's content. If students are not given many opportunities or do not receive good instruction, the amount of error in the students' responses will increase and test reliability will suffer.

Types of Test Reliability

Each type of test reliability estimates different types of error associated with a test. Table 1 provides an overview of the types of reliability and characteristics for each type. Because reliability is a common word that also has a technical meaning, it is important for anyone speaking about, using, or making decisions based on test scores to understand that reliability is not a single concept and that there are multiple ways to determine the reliability of a test.

Table 1. Types of Reliability

Types of Reliability	# of Test Administrations	# of Test Forms	Type of Error Variance
Test Retest	2	1	Stability over time
Alternate Form			
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
Inter-rater	1	1	Stability across raters (i.e., scorers)

As shown in Table 1, "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. In most state testing programs, internal consistency is the primary method of computing reliability, along with inter-rater reliability when constructed- response items are included on the test. However, other types of reliability may be more useful as new and different item types are created and assessments move toward performance assessment, simulations, and games.

Conclusion

Reliability in testing indicates more than just consistency—it also indicates the amount of random error associated with the test score. In other words, reliability refers to the confidence placed in a test score being the correct or true estimate of a student's trait or construct being tested, such as his or her level of proficiency in English or, in the case of an IQ score, his or her general mental ability.

Reliability & Validity

While test reliability is important, a reliable test is of little use if it is invalid. Reliability is necessary but it is not sufficient for great tests. Therefore, a reliable test must also have high validity in order for it to be considered psychometrically sound. Validity is covered in a separate assessment brief.