Reliability

In: The SAGE Encyclopedia of Social Science Research Methods

By: Peter Y. Chen & Autumn D. Krauss
Edited by: Michael S. Lewis-Beck, Alan Bryman & Tim Futing Liao
Book Title: The SAGE Encyclopedia of Social Science Research Methods
Chapter Title: "Reliability"
Pub. Date: 2011
Access Date: September 10, 2018
Publishing Company: Sage Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780761923633
Online ISBN: 9781412950589
DOI: http://dx.doi.org/10.4135/9781412950589
Print pages: 953-957

Theories of reliability have been primarily developed in the context of psychology-related disciplines since Charles Spearman laid the foundation in 1904. Nevertheless, the emphasis on reliability in these areas does not imply that this issue is of little relevance to other disciplines. Reliability informs researchers

about the relative amount of random inconsistency or unsystematic fluctuation of individual responses on a measure. The reliability of a measure is a necessary piece of evidence to support the CONSTRUCT VALIDITY of the measure, although a reliable measure is not necessarily a valid measure. Hence, the reliability of any measure sets an upper limit on the construct validity.

An individual's response on a measure tends to unsystematically vary across circumstances. These unsystematic variations, referred to as RANDOM ERRORS or random errors of MEASUREMENT, create inconsistent responses by a respondent. Random errors of measurement result from a variety of factors that are unpredictable and do not possess systematic patterns. These factors include test-takers' psychological states and physiological conditions, characteristics of the test environments, different occasions of test administration, characteristics of the test administrators, or different samples of test items in which some are ambiguous or poorly worded. As a result, one or more of the above factors contributes to the unpredictable fluctuation of responses by a person should the individual be measured several times.

Types of Measurement Errors

It is safe to say that most, if not all, measures are imperfect. According to a recent development in measurement theory, the scores for any measured variable likely consist of three distinguishable components: the construct of interest, constructs of disinterest, and random errors of measurement (Judd & McClelland, 1998). Of these, the latter two components are measurement errors, which are not desirable but are almost impossible to eliminate in practice. The part of the score that represents constructs of disinterest is considered SYSTEMATIC ERROR. This type of error produces orderly but irrelevant variations in respondents' scores. In contrast, a measure's random errors of measurement do not impose any regular pattern on respondents' scores. It should be noted that errors of measurement in a score are not necessarily a completely random event for an individual (e.g., Student Smith's test anxiety); however, across a very large number of respondents, the causes of errors of measurement likely vary randomly and act as RANDOM VARIABLES. Hence, errors of measurement are assumed to be random. It is the random errors of measurement that theories of reliability

attempt to understand, quantify, and control.

Effects of Random Errors on a Measure

Inconsistent responses on measures have important implications for scientific inquiries as well as practical applications. Ghiselli, Campbell, and Zedeck (1981) pointed out that inconsistent responses toward a measure have little value if researchers plan to compare two or more individuals on the same measure, place a person in a group based on his or her ability, predict one's behavior, assess the effects of an intervention, or even develop theories. Suppose a university admissions officer wants to select one of two applicants based on an entrance examination. If applicants' responses on this test tend to fluctuate 20 points from one occasion to another, an actual 10-point difference between them may be attributed to errors of measurement. The conclusion that one applicant has a stronger potential to succeed in college than another could be misleading. Similarly, a decrease of 10 points on an anger measure after a person receives anger management treatment may result from unreliable responses on the measure. The amount of errors of measurement resulting from one or more factors can be gauged by a variety of reliability coefficients, which will be discussed in a later section.

Classical True Score Theory

There have been two major theories of reliability developed since 1904: classicaltrues core theory, based on the model of parallel tests, and generalizability theory, based on the model of domain sampling. Historically, Spearman (1904, 1910) started to develop classical true score theory, and it was further elaborated on by Thurstone (1931), Gulliksen (1950), and Guilford (1954). Generalizability theory is a newer theory, proposed by Jackson (1939) and Tryon (1957), and later advanced by Lord and Novick (1969) as well as Cronbach and his colleagues (1972). Conventional reliability coefficients such as parallel-forms and alternate-forms RELIABILITY, TEST-RETEST RELIABILITY, INTERNAL consistency RELIABILITY, SPLIT-HALF RELIABILITY, INTERRATER RELIABILITY, and INTRA-CODER RELIABILITY are estimated based on classical true score theory, although these coefficients can also be estimated by generalizability theory (see a demonstration in Judd & McClelland, 1998). Because generalizability theory and its applications are described elsewhere in the Encyclopedia, we will focus only on classical true score theory in the remaining section.

Classical true score theory is a simple, useful, and widely used theory to describe and quantify errors of measurement. In general, classical theory operates under the following assumptions:

1.

- Xi =T +Ei.

2.

- E(Xi) =T.

3.

- F(X1) =F(X2) =F(X3) =⋯=F(Xk).

4.

- ρX1X2 =ρX1X3 =ρX1Xk =⋯=ρX(k−1)Xk .

5.

- ρX1Z =ρX2Z =ρX3Z =⋯=ρXkZ.

6.

- ρE1E2 =ρE1E3 =ρE1Ek = ⋯ =ρE(k−1)Ek =0.

7.

- ρE1T =ρE2T =⋯=ρEkT = 0.

Assumption 1,Xi =T +Ei, states that an observed score (or fallible score) of a test i(Xi), consists of two components: true score (T) and random errors of measurement, Ei. Errors of measurement, caused by a combination of factors described earlier, can be either positive or negative across different parallel tests (defined later), and therefore make the observed scores either higher or lower than the true score.

A true score on a test does not represent the true level of the characteristic that a person possesses. Instead, it represents the consistent response toward the test, resulting from a combination of factors. Note that different combinations of factors can lead to a variety of consistent responses. For instance, consistent responses on a measure of intelligence in an environment with humidity, noise, and frequent interruptions are different from those on the same measure under a cool, quiet, and controlled environment. The former consistent response represents the true score for intelligence under the hardship environment, whereas the latter response is the true score for intelligence under the pleasant condition.

Assumption 2,E(Xi) =T, states that the true score is the expected value (mean) of an individual's observed scores, should the person be tested over an infinite number of parallel tests. For instance, assuming an individual receives a large number of k parallel tests, the observed score of each test for this person can be expressed as

$$
\begin{aligned}
X_1 &= T + E_1, \\
X_2 &= T + E_2, \\
&\vdots \\
X_k &= T + E_k.
\end{aligned}
$$

Because errors of measurement are either positive or negative, the sum of these errors in the

long run is zero, E(Ei) = 0. As a result, the expected value of observed scores equals the true score.

Assumptions 3 through 5 are about the characteristics of parallel tests, which have theoretical as well as practical implications. Assumption 3, F(X1) = F(X2) =F(X3) = ⋯ =F(Xk), states that distributions of observed scores across k parallel tests are identical for a large population of examinees. According to the assumption of identical distributions, we expect T1 =T2 = ⋯ =Tk and σ 2X1=σ2X2= ⋯ =σ2Xk. In addition, CORRELATIONS among observed scores of the k parallel tests are the same, which is expressed as ρX1X2 =ρX1X3 =ρX1Xk = ⋯ =ρX(k−1)Xk (Assumption 4). Finally, Assumption 5, ρX1Z = ρX2Z =ρX3Z =⋯=ρXkZ, states that there are equal correlations between observed scores and any other measure (Z), which is different from k parallel tests.

Assumptions 6 and 7 essentially state that errors of measurement are completely random. Consequently, correlations among errors of measurement across k parallel tests in a large population of examinees equal zero, as do the correlations between errors of measurement and the true score. Based on Assumption 7, a variance relationship amongσ 2X,σσ2Xσ2T, and σ2E can further be expressed by σ2X=σ2T+σ2E. Specifically, the observed score variance in a population of examinees equals the sum of the true score variance and the errors of measurement variance. Deduced from the above relationship in conjunction with Assumption 3, errors of measurement variances across k parallel tests are equal to one another, σ2E1=σ2E2=⋯σ2Ek

---

Theoretical Definition of Reliability

According to the variance relationshipσ2X=σ 2T+σ2E, a perfectly reliable measure occurs only when random errors do not exist (i.e., σ 2E=0 orσ 2X=σ2T). Whenσ2E increases, observed differences on a measure reflect both true score differences among the respondents and random errors. Whenσ 2X=σ2E, all of the observed differences reflect only random errors of measurement, which suggests that responses toward the measure are completely random and unreliable.

If we divideσ 2X on both sides of the above equation, the variance relationship can be modified as

$$1 = \frac{\sigma_T^2}{\sigma_X^2} + \frac{\sigma_E^2}{\sigma_X^2}.$$

Reliability of a measure is defined as

$$\frac{\sigma_T^2}{\sigma_X^2},$$

which can be interpreted as the squared correlation between observed and true scores (ρ2XT). Conceptually, reliability is interpreted as the proportion of the observed score variance that is explained or accounted for by the true score variance. Following the above assumptions, it further proves that a correlation between two parallel tests equals

$$\frac{\sigma_T^2}{\sigma_X^2}.$$

That is,

$$\rho_{X_1X_2} = \frac{\sigma_{T_1}^2}{\sigma_{X_1}^2} = \frac{\sigma_{T_2}^2}{\sigma_{X_2}^2}.$$

If a correlation between two parallel tests is 0.80, it suggests that 80% (not 64%) of the observed score variance is accounted for by the true score variance. We will hereafter use the general notation, ρXX' and rXX', to represent reliability for a population and a SAMPLE, respectively.

Procedures to Estimate Reliability

A variety of reliability coefficients are based on classical true score theory, such as test-retest reliability, internal consistency reliability, interrater reliability, or intracoder reliability, which are reviewed elsewhere in the Encyclopedia. We will focus on three additional reliability coefficients: alternate-forms reliability, reliability of difference scores, and reliability of composite scores.

An alternate-forms reliability coefficient, also referred to as coefficient of equivalence, is estimated by computing the correlation between the observed scores for two similar forms that have similar means, STANDARD DEVIATIONS, and correlations with other variables. Generally, a group of respondents receives two forms in a random order within a very short period of time, although the time period between administrations may be longer.

When conducting interventions or assessing psychological/biological development, researchers are often interested in any change in scores on the pre and posttests. The difference between scores on the same or a comparable measure at two different times is often referred to as a change score or difference score (denoted as D). The reliability of the difference scores in a sample can be estimated by

$$r_{DD'} = \frac{s_X^2 r_{XX'} + s_Y^2 r_{YY'} - 2 s_X s_Y r_{XY}}{s_X^2 + s_Y^2 - 2 s_X s_Y r_{XY}},$$

where s2X and s2Y are variances of the pre-and posttests, rXX' and rYY' are the reliabilities of the tests, and rXY is the correlation between the tests. If the variances of both measures are the same, the above equation can be rewritten as

$$r_{DD'} = \frac{\frac{r_{XX'} + r_{YY'}}{2} - r_{XY}}{1 - r_{XY}}.$$

Given that rXX' and rYY' are constant, stronger rXY results in weaker rDD'. Furthermore, rDD' moves toward zero as rXY approaches the average reliability of the measures. In reality, the size of rXY is often strong because of two measures' similarity. As a result, difference scores tend to have low reliability, even when both measures are relatively reliable.

In contrast to difference scores, researchers sometimes need to utilize a composite score based on a linear combination of raw scores from k measures. For example, a personnel psychologist may select job applicants based on a composite score of an intelligence test, a psychomotor ability test, an interest inventory, and a personality measure. The reliability of this composite score (C) can be estimated by

$$r_{CC'} = 1 - \frac{\sum_{i=1}^{k} s_i^2 - \sum_{i=1}^{k} (r_{ii'} s_i^2)}{s_C^2},$$

where s2C and s2i are variances of the composite score, and rii' is the reliability of measure i. In practice, researchers may create the composite score based on a linear combination of STANDARDIZED scores from k measures. The reliability can be estimated by the simplified formula,

$$r_{CC'} = 1 - \frac{k - k\bar{r}_{XX'}}{k + (k^2 - k)\bar{r}_{XY}},$$

where $\bar{r}$ rXX' is the average reliability across k measures, and $\bar{r}$rXY is the average intercorrelation among the tests. Assuming the average intercorrelation is constant, the reliability of the composite score is greater than the average reliability of the measures. Hence, the reliability of a composite score tends to be relatively more reliable. respectively.

Standard Errors of Measurement

As noted earlier, reliability coefficients only inform about the relative amount of random inconsistency of individuals' responses on a measure. A test with a Cronbach's alpha of .90 suggests that the test is more internally consistent than another test with a Cronbach's alpha of

.70. However, both reliability estimates do not provide absolute indications regarding the precision of the test scores. For instance, the Stanfordbinet Intelligence Scale is a very reliable measure with an internal consistency reliability of .95 to .99. Although the test is highly reliable, we are not confident in saying that a score of 60 is truly higher than the average of 50 by 10 points. More realistically, the 10 points reflect the true difference in intelligence and random errors of measurement.

In order to interpret test scores accurately, researchers need to take unreliability into consideration, which is gauged by the square root of the errors of measurement variance, or standard error of measurement (σE). Standard error of measurement, derived from

$$1 = \frac{\sigma_T^2}{\sigma_X^2} + \frac{\sigma_E^2}{\sigma_X^2},$$

is expressed asσE =σX√1 −ρXX' in a population, which can be estimated by ^σE =sE =sX√1 −rXX' from a sample. Conceptually, the standard error of measurement reveals how much a test score is expected to vary given the amount of error of measurement for the test. WhenσE is available, researchers can use it to form confidence intervals of the observed scores. Specifically, the observed score (X) in a sample has approximately 68%, 95%, and 99% confidence to fall in the range of X ± 1 ^σE, X ± 2 ^σE, and X ± 3 ^σE, respectively.

Factors that Affect Reliability Estimates

While estimating the reliability of a measure, researchers cannot separate the respondents' characteristics from the measure's characteristics. For instance, test performance on an algebra test in part depends on whether the test items are difficult or easy. At the same time, a determination of the test items as easy or difficult relies on the ability of the test taker (e.g., college student vs. junior high school student; homogeneous group vs. heterogeneous group). When a group is so homogeneous that members possess similar or the same algebraic ability, the variance of the observed scores could be restricted, which leads to a very low reliability even though the test could be perfectly reliable. The above examples clearly demonstrate that reliability estimates based on classical true score theory change contingent upon the group of test takers and/or the test items. Specifically, traditional procedures to estimate reliability are group dependent and test dependent, which suggests that the standard error of measurement (and reliability) of a measure is not the same for all levels of respondents. Empirical evidence has shown that the standard error of measurement tends to be smaller for the extreme scores than the moderate scores.

There are other characteristics of measures that affect reliability. First, according to the

Spearman Brown Prophesy formula, the reliability of a measure tends to increase when the length of the measure increases, given that appropriate items are included. The Spearman-Brown Prophesy formula is defined as

$$\hat{\rho}_{XX'} = \frac{n \times \rho_{XX'}}{1 + (n-1)\rho_{XX'}},$$

where n is the factor by which an original test is lengthened, and ^ρXX' and ρXX' are reliability estimates for the lengthened test and the original test, respectively. Second, empirical findings have shown that reliability coefficients tend to increase when response categories of a measure increase. A MONTE CARLO study has revealed that overall reliability estimates of a 10-item measure, such as test-retest reliability or CRONBACH'S ALPHA, increase when response categories increase from a 2-point scale to a 5-point scale. After that, the reliability estimates become stable. Test content can also affect reliability. For instance, the reliability of power tests can be artificially inflated if time limits are not long enough. Power tests refer to tests that examine respondents' knowledge of subject matter. When time limits are short, most respondents cannot answer all the items. As a result, their responses across items become artificially consistent.

Finally, reliability can be affected by the procedures researchers opt to use. As described earlier, reliability coefficients based on classical true score theory estimate one or two types of errors of measurement. For instance, internal consistency estimates response variations attributed to item heterogeneity or inconsistency of test content, whereas test-retest reliability estimates variations attributed to changes over time, including carry-over effects. In contrast, alternate-forms reliability with a long lapse of time (e.g., 4 months) estimates two types of random errors, namely, those attributed to changes over time and inconsistency of test content. Although all reliability coefficients based on classical true score theory estimate the relative amount of errors of measurement, the errors estimated by different reliability coefficients are qualitatively different from each other. Hence, it is not logical to compare two different types of reliability estimates.

The fact that different random errors are being estimated by different reliability coefficients has an important implication when researchers correct for attenuation due to unreliability. Derived from the assumptions of classical true score theory, a population correlation between two true scores (ρTXTY) can be adversely affected by unreliability of the measures (X and Y). The adverse effect can be deduced from the relationship between a population correlation and population reliabilities of measures X and Y (ρXX' and ρYY'), which is defined as

$$\rho_{T_X T_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}}.$$

Often, researchers want to show what the sample correlation (^rXY) would be if both measures were perfectly reliable. Following the modified formula from above, the disattenuated correlation can be obtained by

$$\hat{r}_{XY} = \frac{r_{XY}}{\sqrt{r_{XX'}r_{YY'}}}.$$

The difference between ^ rXY and rXY increases when one or both measures become less reliable. For instance, r^XY equals rXY when rXX'=rYY' = 1. However, the size of ^rXY becomes twice as much as rXY when rXX =rYY'=0.50. In practice, the correction for attenuation tends to overestimate the actual correlation. Foremost, it does not seem tenable to assume that both rXX' and rYY' deal with the same type(s) of errors of measurement. Furthermore, there are many different types of errors of measurement that are not adequately estimated by classical true score theory. For instance, researchers may be interested in finding out how reliable an achievement test is when it is administrated by both male and female administrators at two different times and at three different locations.

- true scores
- errors

Peter Y.Chen and **Autumn D.Krauss**

http://dx.doi.org/10.4135/9781412950589.n845
10.4135/9781412950589.n845

References

Crocker, L., & **Algina, J.(1986).** Introduction to classical & modern test theory.New York: Harcourt Brace Jovanovich.

Feldt, L. S., & **Brennan, R. L.(1989).** Reliability. In Edited by: **R. L. Linn** (Ed.), Educational measurement (3rd ed., pp. 105–146).New York: Macmillan.

Ghiselli, E. E., **Campbell, J. P.**, & **Zedeck, S.(1981).** Measurement theory for the behavioral sciences.San Francisco: W. H. Freeman.

Hambleton, R. K., **Swaminathan, H.**, & **Rogers, H. J.(1991).** Fundamentals of item response theory.Newbury Park, CA: Sage.

Judd, C. M., & **McClelland, G. H.(1998).** Measurement. In Edited by: **D. T. Gilbert**, **S. T. Fiske**, & **G. Lindzey** (Eds.), The handbook of social psychology (4th ed., Vol. 1).Boston: McGraw-Hill.

Nunnally, J. C., & **Bernstein, I. H.(1994).** Psychometric theory (3rd ed.).New York: McGraw-Hill.