Kline, R. B. (2019). Psychometrics. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), SAGE Research Methods Foundations. doi: 10.4135/9781526421036831959

SAGE researchmethods

Psychometrics

Foundation Entries

SAGE Research Methods Foundations

By: Rex B. Kline
Published:2020
Length: 10,000 Words
DOI: http://dx.doi.org/10.4135/9781526421036
Methods: Psychometrics
Online ISBN: 9781526421036
Disciplines: Psychology
Access Date: May 21, 2020
Publishing Company: SAGE Publications Ltd
City: London

Abstract

Psychometrics refers to the measurement of hypothetical constructs, such as cognitive abilities or attitudes, that are not directly observed but instead are inferred. Psychometrics also refers to special statistics about the properties of scores from psychological tests such as score reliability coefficients and validity coefficients, among other kinds. Described in this entry are the origins and characteristics of classical measurement theory, which was developed from about 1900–1960 and is still relevant to this day. Classical measurement theory is the basis for introducing students to psychological measurement. The same principles should also guide the selection of measures for use in research projects that involve psychological constructs, which are pertinent in many human studies. Also briefly considered are two approaches to psychometrics in modern measurement theory. These newer methods build on classical measurement theory or depend on statistical techniques that were impractical before the availability of relatively affordable fast computers with large memory capacities. One is generalizability theory, which extends the classical perspective on score reliability. The other modern approach is item response theory, also known as latent trait theory, which consists of families of mathematical models that relate observed responses on test items to underlying continuous variables, or latent traits, that represent target hypothetical constructs. Journal article reporting standards for psychometrics are also discussed.

Introduction

The product of measurement in science are scores, which are then analyzed. Formal theories and mathematical models for the measurement of physical attributes in disciplines such as physics, chemistry, and engineering date to the late 1800s (Hand, 1996). In the social sciences, the measurement of hypothetical constructs is emphasized, especially in disciplines such as psychology and education in which individual differences are often studied. Constructs such as “leadership ability,” “intelligence,” “quality of life,” “religiosity,” “attitudes toward abortion,” or “trait anxiety” are not directly observable and, thus, are latent variables. Principles of construct measurement involve the specification of quantifiable observed variables that are all supposed to indicate the same construct, given the theoretical definition of that construct. Next, these operational definitions yield quantities that are converted to scores in a particular metric, such as standardized scores. Finally, the precision and accuracy of these scores are evaluated. Score precision and accuracy correspond to, respectively, reliability and validity, which are critical concepts in measurement theory. Methods for evaluating score reliability and validity are thus a major topic of this entry.

Classical Measurement Theory

Classical theory about construct measurement has its origins in two key events at the beginning of the 1900s. The first was the development in France of the Binet-Simon intelligence test for children in 1905 (an English

translation of their work is available in Binet & Simon, 1916). At the time, items of the Binet–Simon Scale were relatively unique in that they directly involved complex mental abilities such as verbal reasoning, memory for words or objects, and visual-spatial analysis. In contrast, approaches to mental testing in the late 1800s were generally based on measuring sensory reactions or simple motor skills, but such indicators were later found to be essentially uncorrelated with indices of complex mental functioning, such as scholastic achievement (Jones & Thissen, 2007).

A second impetus for classical measurement theory was the development of statistical techniques for analyzing mental test scores in order to evaluate hypotheses about the nature and organization of human intelligence (Jones & Thissen, 2007). For example, C. Spearman (1904a, 1904b) devised a method of factor analysis that analyzes observed correlations from mental tests assuming that a general intelligence factor, or *g*, is common to all of the tests. He also described the basic idea that observed scores can have at least two sources of influence, including the latent variable the scores are intended to assess and random measurement error. Along with others such as L. L. Thurstone (1947) who developed a factor analytic method for detecting multiple factors of intelligence instead of a single general factor, Spearman articulated the basic statistical framework for classical measurement theory about score reliability and validity.

Outlined next are the two major families of psychometrics: score reliability coefficients and validity coefficients. In the discussion that follows, we assume that a test consists of $N_i \geq 2$ items where a total score, X, is the sum of the responses across all the items. These characteristics describe *multiple-item measures*, whereby a test comprises multiple items. In contrast, *single-item measures* rely on a single item or question, which is far from ideal. Imagine that a questionnaire about life satisfaction has the single item listed next:

> I am happy with my life (0 = *disagree*, 1 = *uncertain*, 2 = *agree*)

A researcher may object that the construct of life satisfaction cannot be adequately captured by a single, global question. For example, there are probably multiple facets of this construct, including satisfaction with family or social relationships, work-related prospects, and general health, among other possibilities. It is also true that people's responses to a single item can be inconsistent over time, depending on their mood on a particular day or other factors. These limitations are partially remedied by multiple-item measures, whereby the items as a set reflect somewhat different facets of the construct and total scores are subsequently analyzed. Total scores tend to be more reliable than item scores, which means less measurement error when total scores are analyzed compared with when item scores are analyzed. Given all these advantages, psychological tests are almost never single-item measures.

Score Reliability

Reliability concerns the precision of scores from a psychological test. It is critical to know that reliability is not a property of *tests*; instead, reliability concerns the precision of test *scores* in a particular sample. This is because a test is not reliable or unreliable across all possible samples or uses of the test. For example,

a reaction time task in which responses are collected using video game controllers that generate precise scores for teenagers may fail to do so for elderly adults who may be unfamiliar with the same hardware. Another example is a test that yields precise scores for respondents with no mental health problems, but scores from the same test are imprecise in samples of psychiatric patients. Bruce Thompson and Tammi Vacha-Haase (2000) speculated that the widespread but false belief that reliability is an attribute of tests, not scores in particular samples, leads to a kind of "black box" mentality, or the assumption that reliability can be established by others, such as test authors, and that, once established, the same level of score precision can be expected within any other sample. The same flawed mentality may also explain the generally poor state of reporting practices about score reliability apparent in the social science research literature. Standards for reporting information about psychometrics are considered later in this entry.

Score precision has a special meaning in reliability analyses: If scores for the same participants maintain their relative positions over variations in time, test versions, item selection, or raters (for tests that are subjectively scored), then those scores are precise. This means that participants with the highest scores in one variation also tend to have the highest scores in another variation, just as participants with lower scores in one variation also have lower scores in other variations. Scores with the property just described are reliable. Note that reliability does *not* mean that each participant obtains the *same* score over all conditions. Scores as just described have perfect precision, but so would two sets of scores with the same relative values even though no two scores are identical for any case. For example, the following are two sets of scores for participants S_1–S_5, respectively:

Set 1: 10, 15, 25, 32, 36

Set 2: 15, 20, 30, 37, 41

No scores for the same participant are identical across the two sets just listed, but the *relative values* of the scores in each set are identical over participants. The Pearson correlation between these two sets of scores is $r = 1.0$ and describes the same characteristic. The fact that some, but not all, kinds of score reliability coefficients are based on Pearson correlations helps to make reliability analysis more familiar.

A common symbol for a generic reliability coefficient is r_{XX}, where the character X designates a set of total scores and the whole subscript, XX, represents the idea of score precision or repetition over variations. In classical measurement theory, r_{XX} estimates the proportion of the total observed variance in scores that is *not* due to the type of random error (i.e., imprecision) measured by a particular kind of reliability coefficient. Because r_{XX} is a proportion of variance, its theoretical range is 0–1.0. For example, if $r_{XX} = .80$, then $1 - .80 = .20$, or *at least* 20% of total variance is random due to measurement error. But the remaining proportion of standardized variance, or .80 = 80%, may not all be systematic (nonrandom). This is because a particular type of reliability coefficient may estimate just a single source of error, and scores can be affected by multiple sources of error. This fact explains why score reliability is usually estimated over a series of studies, whereby a different type of random measurement error is estimated in each analysis.

As r_{XX} approaches zero, the scores are increasingly more like random numbers, and random numbers have no precision and measure nothing (i.e., they also have no validity). It can happen that the value of an empirical reliability coefficient is < 0. A negative coefficient is interpreted as though its value were zero, but such a result (r_{XX} < 0) indicates a serious problem with the scores. In general, r_{XX} > .90 is generally considered as excellent score precision, which is required in a *high-stakes testing* where the results from psychological testing have consequences, such as whether an examinee is eligible or ineligible for a scholarship or whether a student is eligible or not for remedial academic services. Values of r_{XX} between .70 and .90 are generally adequate for research applications in which group statistics, such as mean differences, are analyzed, but even lower values of r_{XX} may indicate potential problems. For example, the result r_{XX} < .50 says that more than half of total variation is random; that is, there is more noise than signal in the scores, and scores so imprecise should probably not be analyzed.

Reliability Methods and Coefficients

Listed in Table 1 are classical methods to estimate score reliability. All these methods involve administering some proportion of the test, k, within the same sample on ≥ 1 occasion(s). When the test proportion is k = 1.0, the whole (original) test is administered. In the *test–retest method*, the same test is administered within the same sample but at two different points in time. The test–retest reliability coefficient is the Pearson correlation between the two sets of scores over time, and it measures the degree to which scores maintain their relative positions over time. Thus, test–retest reliability coefficients measure *time sampling error*. It is crucial to specify an appropriate retest interval, given the definition of the corresponding construct. For example, a retest interval of 1 year may be appropriate when test scores are supposed to reflect enduring characteristics such as general cognitive ability among adults as measured by IQ scores. But a 1-year interval may be much too long if the scores are expected to measure something less enduring, such as state anxiety.

The *interrater reliability method* is for tests where the scoring is not completely objective; that is, the examiner must exercise judgment in scoring. The original test (k = 1.0) is administered once, but responses are independently scored by two different raters (see Table 1). The Pearson correlation between the two sets of scores, one from each rater, is the interrater reliability coefficient. If the value of this coefficient is relatively high, then the scores generally maintain their relative positions over the two raters. Thus, interrater reliability coefficients measure *rater-sampling error*, or whether score consistency is affected by idiosyncratic—and thus random—characteristics of individual examiners. Raters should be properly trained in the first place, and it may be necessary to subsequently retrain raters from time to time. This is because raters tend to become less precise after training unless they receive “booster” sessions that remind them of correct procedures. *Rater drift* describes the expected deterioration in precision after initial training. Meredith G. Warshaw, Ingrid Dyck, Jenifer Allsworth, Robert L. Stout, and Martin B. Keller (2001) describe a program to prevent rater drift in a longitudinal study of anxiety disorders.

In the *alternate* (*parallel*) *forms method*, the proportion of the test administered is k = 2.0 (Table 1). This means that two different forms (versions) of the test are created, each of which is supposed to measure the same

domain but with nonoverlapping sets of items. The availability of an alternate version is ideal for situations in which reassessment of examinees is mandated. For example, it is common for states or provinces to require the retesting of school children who receive special education services within a specified time frame, such as an obligatory 1-year follow-up period. A drawback of the alternate forms method is that it requires two equivalent forms of the test, which takes more effort than developing a single form. The alternate forms reliability coefficient is the Pearson correlation between the scores across the two forms. It measures *content-sampling error*, or whether the items of each form seem to have been selected from a common domain. If the value of the alternate forms reliability coefficient is low, then the two versions of the test do not measure the same construct.

In the *split-half reliability method*, the fraction of the test is $k = .5$ (Table 1). The whole test is administered just once, but then test items are partitioned into two equivalent halves, and a subtotal score is calculated for each half of the test. The Pearson correlation between these two sets of subtotal scores is r_{hh}, where the subscript indicates that each subtotal is based on half of the test's items. Next, the value of r_{hh} is corrected for the total number of items on the original test, and this corrected result, r_{11}, is the split-half reliability coefficient. The correction that generates the latter is the *Spearman–Brown prophesy formula*, or

(1)

$$r_{11} = \frac{2r_{hh}}{1 + r_{hh}}$$

where the constant "2" represents the fact that the original test has twice as many items as half of the test. For example, if $r_{hh} = .60$, then applying Equation 1 generates $r_{11} = .75$. In general, $r_{11} > r_{hh}$, which represents the fact that longer tests tend to generate total scores that are more reliable than shorter tests, in this case splits based on half as many items compared with the original test.

Because there are multiple ways to split test items, the value of r_{11} from a particular split is typically not unique. When items become increasingly difficult, a logical method is to assign the odd-numbered items to one set and the even-numbered items to the other set. The two sets formed this way should be roughly equal in difficulty; thus, an odd-–even split may be optimal in that the corresponding value of r_{11} is relatively high. For tests in which items are *not* presented in order of increasing (or decreasing) difficulty, other methods to split the items include first half–second half splits and various random selections of half the items to form one set while the rest are assigned to the other set. None of these options may be clearly optimal, so variation in r_{11} values is a kind of sampling error, in this case over different methods to split test items. Split-half reliability coefficients also measure content-sampling error (Table 1). If the subtotal scores do not maintain their relative values across the two halves of the test, then it cannot be said that those two sets of items measure a common domain.

Content-sampling error in the *internal consistency reliability method* is estimated at the level of individual items instead of at the whole-test level (alternate forms) or at the half-test level (split-half). The whole test is administered once, but the data are analyzed at the item level; thus, the test is conceptually split into as many parts as items, so $k = 1/N_i$ (Table 1). Consistency at the item level is measured by the $N_i \times N_i$ matrix of Pearson correlations between each pair of items. If values of these correlations are generally positive ($r_{ij} > 0$) over most pairs of items, then responses at the item level are consistent; that is, there is evidence for internal consistency. But values of $r_{ij} \leq 0$ indicate pairs of items where responses are not consistent. This is because $r_{ij} \leq 0$ indicates that responses did not maintain their relative values over the two items. Inconsistent responding at the item level may indicate heterogeneous content; that is, the items were not all drawn from a common domain.

The most widely reported internal consistency reliability coefficient is *Cronbach's α*, r_α, also called the *α coefficient*. A general equation for continuous test items is

(2)

$$r_\alpha = \frac{N_i \, \bar{c}_{ij}}{\bar{s}_i^2 + (N_i - 1)\bar{c}_{ij}}$$

where $\bar{s}_i^2$ is the average variance over all items, and $\bar{c}_{ij}$ is the average covariance for all pairs of items. A covariance for two continuous variables is the product of their Pearson correlation and standard deviations or

(3)

$$c_{ij} = r_{ij} \, S_i \, S_j$$

A covariance is an unstandardized measure of the linear association between two variables that preserves the original metrics of each variable. In contrast, the Pearson correlation r_{ij} is a standardized statistic where the maximum absolute value is 1.0. If $r_{ij} > 0$, then $c_{ij} > 0$ for the same two variables. Likewise, higher positive values of both r_{ij} and c_{ij} indicate greater stability in item responses, or internal consistency.

A calculational example follows. Suppose for $N_i = 3$ items that

$$s_1 = 2.5, \; s_2 = 5.0, \; s_3 = 4.5$$

$$s_1^2 = 6.25, \; s_2^2 = 25.0, \; s_3^2 = 20.25$$

$$r_{12} = .40,\ r_{13} = .60,\ r_{23} = .50$$

Given the item statistics just listed and using Equations (2) and (3), the reader should verify the results that follow:

$$c_{12} = 5.00,\ c_{13} = 6.75,\ c_{23} = 11.25$$

$$\overline{s}_i^2 = (6.25 + 25.0 + 20.25)/3 = 17.17$$

$$\overline{c}_{ij} = (5.0 + 6.75 + 11.25)/3 = 7.67$$

$$r_\alpha = \frac{3(7.67)}{17.17 + (3-1)7.67} = .71$$

Thus, r_α = .71 says that a total of 1 − .71 = .29, or 29% of the observed variation in total scores is due to inconsistent responding at the item level weighted by the total number of items (3).

There is a special version of Equation 2 called the *Kuder–Richardson Formula 20* (KR20) (Kuder & Richardson, 1937) for dichotomously scored items, such as true-false items or items scored as correct or incorrect, and also a version called *standardized alpha*, which is calculated after item responses are standardized (converted to normal deviates; see Thompson, 2003, for more information and examples). The two special versions just described are automatically calculated in some computer procedures for reliability analysis, such as the Reliability Analysis procedure in SPSS.

It can be shown that the value of r_α equals the average split-half reliability coefficient, r_{11}, computed for all possible splits where k = .5 for the same test, but only if the item variances are all equal (Cronbach, 1951). This means that values of some among all possible r_{11} values will exceed that of r_α, but values of r_{11} for other splits will be lower than that of r_α, again for the same test. If difficulty progressively increases from earlier to later items, then r_{11} based on an odd–even split may be preferred over r_α. But if item order is unrelated to difficulty—and for which no particular type of split would clearly be optimal—then r_α as the central tendency among all possible values of r_{11} provides an overall estimate of reliability at the item level.

There are some noteworthy limitations of r_α. It is a confounded measure of internal consistency and test length (Sijtsma, 2009; Tavakol & Dennick, 2011). This means that as either quantity increases, so does the value of r_α (Equation 2). Suppose that the average item covariance $\overline{c}_{ij}$ is positive but just slightly greater than zero, so items responses are not generally consistent. But if the test is very long, then test length could offset the near-zero value of $\overline{c}_{ij}$ so that the value of r_α is, say, .80, a result that looks nominal, but it masks a problem with

the items. Thus, higher values of r_α do not guarantee that the items are *unidimensional*, which means that total scores based on those items can be interpreted as varying along a single-underlying (latent) dimension. In fact, r_α assumes unidimensional measurement, so its value does not somehow evaluate the plausibility of this assumption.

If the items actually come from different domains—that is, measurement is multidimensional—then r_α will *underestimate* reliability in this case. This is why r_α is called a *lower-bound estimate of reliability* (Tavakol & Dennick, 2011). If a set of items corresponds to at least two different latent dimensions, then the total score computed over the whole set has no simple interpretation. Also, separate subtotal scores would be needed, one for each subset of unidimensional items that measure something in common within the larger set of multidimensional items, if total scores are to be derived as response summaries. If so, then it would make sense to calculate separate values of r_α for each subset of unidimensional items within the larger test.

Another problem with r_α is that many tests comprise items with *Likert-type scale response formats*, for which people rate their amount of agreement to a series of statements, or items. Gradations along a Likert-type scale are often represented with numbers, such as 0 for *disagree* versus 1 for *agree* for true-false items. Finer distinctions can be made using Likert-type scales with ≥3 levels, such as

0 = *strongly disagree*, 1 = *disagree*, 2 = *neutral*, 3 = *agree*, and 4 = *strongly agree*

In statistical analyses, items with Likert-type scales should not generally be treated as continuous variables, especially if the number of response alternatives is relatively small, such as ≤ 5. This is because such data are actually categorical, that is, either dichotomous for items with just two response options (e.g., 0/1 coding) or ordinal for items with ≥ 3 response categories. Pearson correlations are not the best measures of association when the data are categorical; specifically, their values can severely misestimate the true relation between continuous latent variables measured with responses to Likert-type scale items (Gadermann, Guhn, & Zumbo, 2012). This is especially true if the theoretical underlying dimensions generate skewed distributions of responses to Likert-type scale items. The coefficient r_α is based in part on average covariances over sets of continuous items (Equation 2), and covariances are for continuous variables.

Some alternatives to r_α are briefly described here. Gadermann and colleagues (2012) described *ordinal α*, which is an item-level internal consistency coefficient for categorical data. Ordinal α is based on *polychoric correlations* among Likert-type scale items. A polychoric correlation estimates the linear association between a pair of continuous and normally distributed latent variables each measured by observed variables that are ordinal, not continuous. Results of computer simulations by Gadermann and colleagues (2012) indicate that ordinal α is a better estimator of theoretical reliability for ordinal data than r_α calculated for the same ordinal data.

Two alternative internal consistency measures for continuous items were described by Gregory R. Hancock and Ralph O. Mueller (2001) and Tenko Raykov (2004). Both coefficients take account of the factor structure of a set of items, and their values are generally less affected by test length compared with r_α. The *composite*

reliability (CR) coefficient is

(4)

$$r_{\mathrm{CR}} = \frac{\left(\sum \hat{\lambda}_i\right)^2}{\left(\sum \hat{\lambda}_i\right)^2 + \left(\sum \hat{\varepsilon}_i\right)}$$

where $\hat{\lambda}_i$ is the sample standardized factor loading (estimated correlation with the factor) for the *i*th item and $\hat{\varepsilon}_i$ is the standardized error variance for the corresponding item, both results from a factor analysis where all items are specified to measure a single dimension. For any item,

$$\hat{\lambda}_i^2 = 1.0 - \hat{\varepsilon}_i$$

which says that the squared factor loading equals the difference between the total standardized variance (1.0) and error variance. The numerator in Equation 4 estimates the proportion of item variance explained by the common factor. Factor analytic results consistent with a single-factor model generate higher values of r_{CR}, unlike r_{α}. A computationally simpler alternative is the *average variance extracted*, which is just the average of the squared standardized factor loadings for items specified to measure a single factor (see Kline, 2016 for more information).

Multiple Reliability Estimates

A brief example illustrates potential advantages of estimating more than one type of score reliability for a hypothetical test of impulsivity. The format of the test is a structured interview where the responses to 50 questions are recorded but scored later by the examiner. Item responses are summed to form a total score, but scoring is not strictly objective. Across the 50 items, the average bivariate correlation is .15. The following are values of reliability coefficients for this hypothetical test:

Interrater reliability, .85; r_{α} = .90

Test-retest reliability: 3 months, .80; 1 year, .30

Given these results, we can say that test scoring is reasonably precise and that responses at the item level are generally consistent, but we cannot claim that the 50 items measure a single dimension. Of the observed variation in test scores, about 1 − .85 = .15, or 15%, is due to random scoring error and another 1 − .90 = .10, or 10%, is due to both content sampling error and test length. Test scores are generally stable over three months, for which 1 − .80 = .20, or 20%, of observed variation is due to time sampling error. But scores are

clearly inconsistent after 1 year, for which most of the observed variation, or 1 − .30 = .70, or 70%, is due to time sampling error.

Prophesy Formula Revisited

The general form of the Spearman–Brown prophesy formula (Brown, 1910; Spearman, 1910) estimates score reliability given a change in test length. Suppose that r_{XX} is a reliability coefficient for scores of an existing test, and k is the factor by which test length is theoretically changed. If k = 2.0, for example, then the modified test will have twice as many items as the original; if k = .5, then the modified test will have half as many items. The equation

(5)

$$\hat{r}_{XX} = \frac{k\, r_{XX}}{1 + (k-1)\, r_{XX}}$$

generates the estimated score reliability for the modified test. In general, if k > 1 (test is lengthened), then $\hat{r}_{XX} > r_{XX}$, but if k < 1 (test is shortened), then $\hat{r}_{XX} < r_{XX}$. Both results reflect the expected impact of test length on score reliability.

Suppose that r_{XX} = .50 for a test with 10 items. A researcher who is dissatisfied with this result uses Equation 4 to generate the predicted score reliability, if the test were twice as long. Given the observed reliability and k = 2.0, then $\hat{r}_{XX}$ = .67, which is the predicted reliability for scores from a 20-item version of the test (Equation 5 can be used to verify this claim). If even higher predicted reliability is sought, then a value of k > 2.0 could be substituted in Equation 5. An alternative is to select a target level of reliability, such as $\hat{r}_{XX}$ = .80 and then solve Equation 5 for the value of k. For example, given $\hat{r}_{XX}$ = .80 and r_{XX} = .50 for a 10-item test, the result k = 4.0 satisfies Equation 5. In words, the predicted reliability for a test with 40 items, or k = 4 times longer than the original test, is .80 when applying the prophesy formula.

The prophesy formula generates *estimated* reliabilities. The *empirical* (observed) reliability for a test in which items are either added or deleted must be calculated in a particular sample. The prophesy formula also assumes that any items added to a test (k > 1.0) have the same psychometric characteristics as the original items. But if new items are actually worse in quality than the original items, then the empirical reliability could be lower than that indicated by the prophesy formula. It could even be lower in value that of r_{XX} for the original test; that is, adding bad items can actually reduce score reliability even though the revised test is longer. Likewise, the prophesy formula assumes that any random selection of items to be dropped from a test (k < 1.0) would result in the same amount relative reduction in the value of $\hat{r}_{XX}$ compared with that of r_{XX} for the original test, but the actual effect of shortening a test must be estimated with sample data.

Factors That Affect Reliability

The following are characteristics of tests, samples, examiners, or test conditions that can affect values of reliability coefficients:

1. *Test length*. Scores from longer tests are generally more precise than scores from shorter tests (e.g., Equations 2, 5), assuming comparable item psychometrics.
2. *Item clarity and scoring*. Items should be well written and unambiguous in their meaning. Poorly written items can result in less precise scores due to greater random error in responding.
3. *Objective versus subjective scoring*. Objectively scored tests tend to generate more precise scores than tests for which scoring is subjective or requires examiner discretion.
4. *Speed*. In a pure *speed test*, items are trivially difficult but so numerous that it may be difficult, if not impossible, to complete all items within the time allotted. Because the number of items in a speed test that are completed by respondents will vary, the methods listed in Table 1 are not generally suitable for such tests. Special methods are needed to estimate the reliability of scores from speed tests (Gulliksen, 1950). The methods in Table 1 are best for *power tests*, whereby respondents generally have time to respond to all or nearly all items, which are of nontrivial difficulty.
5. *Examiner proficiency*. Examiners who administer or score tests should be well trained. If scoring is subjective, training should be periodically refreshed in order to avoid rater drift.
6. *Sample homogeneity*. Some types of reliability coefficients are Pearson correlations, and values of absolute correlations tend to be lower when calculated in homogeneous samples with reduced variances on test scores (i.e., range restriction). Likewise, values of reliability coefficients tend to be higher in more heterogeneous samples.
7. *Other sample characteristics*. Score precision for the same test can vary appreciably as a function of respondent age, gender, ethnicity, reading proficiency, or level of education, among other variables.
8. *Testing situation or temporary states*. Suboptimal testing conditions, such as distracting noise levels, can reduce score precision. Temporary states of respondents, such as illness, can also lower precision. Longer retest intervals in the test–retest method tend to reduce the temporal stability of scores.
9. *Test split*. How test items are split into two groups, such as an odd–even split versus random selection of items that comprise each half-test, affects values of split-half reliability coefficients.

Consequences of Low Reliability

There are many detrimental effects of low score reliability. In research, poor reliability reduces the effective power of statistical tests, which means that the probability of finding statistically significant results, given a real effect in the population, is reduced. Effects sizes are generally attenuated when scores on outcome variables are unreliable, but absolute effect sizes can be artificially either increased or decreased when measurement error affects both predictor and outcome variables, such as in the technique of multiple regression (see Williams, Grajales & Kurkiewicz, 2013 for examples). There are statistical techniques that

take direct account of score reliability, such as structural equation modeling when latent variable models are analyzed (Kline, 2016), but many classical methods, including multiple regression and the analysis of variance, make unrealistic assumptions about measurement error. In high-stakes testing, decisions based on imprecise test scores may result in harm through either denial of treatment resources to people who really need them or unnecessary exposure to treatments with potentially harmful side effects among people who actually need no treatment.

Values of Pearson correlations are generally truncated in absolute value when scores are not perfectly precise. The exact effect of unreliability on correlation magnitudes is spelled out in the following equation:

(6)

$$\max \left| r_{XY} \right| = \sqrt{r_{XX}\, r_{YY}}$$

That is, the maximum absolute value of the Pearson correlation between variables X and Y is limited by the square root of the product of their score reliabilities. Suppose that r_{XX} = .80 and r_{YY} = .30, which says that scores on variable X are reasonably precise but scores on variable Y are very imprecise. Given the reliability coefficients just listed, the greatest absolute value expected for r_{XY} is .49. This means that the theoretical range for r_{XY} is actually the interval

$$-.49 \leq r_{XY} \leq .49$$

and *not* the interval

$$-1.0 \leq r_{XY} \leq 1.0$$

Because scores analyzed in the social sciences are almost never perfectly precise, it is generally a myth that Pearson correlations range in value from −1.0 to 1.0. The danger in this myth is that a researcher may be unable to correctly judge the strength of an association, if that researcher falsely believes that the range for r_{XY} is always −1.0 to 1.0. For example, what might appear as a moderate association may actually be as strong as it could possibly be, once reliability is considered (Huck, 2016).

Computer Tools for Reliability Analyses

Some computer programs for general statistical analyses have special procedures for calculating reliability coefficients. For example, both IBM SPSS Statistics, a commercial software program, and GNU PSPP, a freely available computer tool that reads and writes SPSS data files, have reliability procedures that calculate various types of coefficients, including r_{α}, and also generate values of item statistics, such as Pearson

correlations between each item and the total score across all items on a scale.

There is no specific procedure in SAS/STAT for reliability analyses. Instead, the SAS/STAT user relies on more general procedures that can optionally compute values of reliability coefficients. For example, the CORR procedure in SAS/STAT has an option to compute r_{α} for a set of items, and the FREQ procedure has an option for calculating an index of agreement between two different raters who make judgments about a categorical (nominal) outcome.

The *α* command in Stata computes r_{α}, and various measures of interrater agreement can be generated using the intraclass correlation, or *icc*, command. A procedure for reliability and item analysis in STATISTICA calculates both reliability coefficients and various types of items statistics. It also provides a set of interactive "what-if" procedures that estimate reliability given particular changes to an existing test (e.g., it re-estimates reliability after adding or deleting items). There are freely available packages for reliability analyses in the *R* computing environment, including *ReliabilityTheory*, *coefficientalpha*, and *psych*, among others.

Validity

Reliable scores are required for validity (e.g., Equation 6). This is because imprecise scores measure nothing, but reliability does not guarantee validity. Thus, it can happen that scores are precise but not valid. Validity concerns the accuracy of interpretations of scores from psychological tests in a particular context, which involves both the population and purpose for which the test is intended. Just as reliability is not an immutable property of tests, the concept of validity does not refer to a test per se but instead to proposed interpretations of test scores (Reynolds & Livingston, 2012). A related concept is that of *interpretation-use arguments*, which concern the plausibility and appropriateness of the both the interpretation and proposed uses of scores (Kane, 2013). As the range of potential generalizations from test scores increases, such as from an observed sample of performances (test data) to predicted performances in other settings, more and more evidence is needed in order to support interpretations of scores.

A higher-order of validity is *construct validity*, or whether scores actually measure the target construct, given its operational definition. Another description is to say that construct validity entails the *meaning* of the scores in relation to the latent variable those scores are supposed to reflect. All other forms of validity described next can be seen as more specific instances of construct validity. Establishing construct validity usually requires multiple lines of evidence, each of which address a particular aspect of construct validity. So similar to reliability, a series of empirical studies is typically needed in order to evaluate construct validity.

One facet of construct validity is *content validity*, or whether test items are representative of the target domain. Content validity is critical for scholastic achievement measures, such as tests that should assess specific skills at a particular grade level (e.g., Grade 4 math). In educational settings, content validity may be referred to as *curricular validity*, or whether test content matches the specifications and objectives of a specific curriculum. Content validity is also important for symptom rating scales. For example, the items of a

depression rating scale should represent the symptom areas thought to reflect clinical depression.

Establishing content validity is more straightforward when the target domain is relatively well defined, such as knowledge of state or provincial traffic laws and signs. If the target domain is ill-defined, though, then establishing content validity is more difficult. Consider the construct of “leadership.” What exactly is meant by this term? Just what are the attributes, habits, attitudes, or skills of a leader? Do these characteristics change over settings, contexts, or environments (e.g., business, government, education) and, if so, how? Without clear answers, it can be hard to evaluate content validity. Stephen N. Haynes, David C. S. Richard, and Edward S. Kubany (1995) describe methods for collecting expert opinions about item content.

Whether test scores relate to an external variable against which the scores can be evaluated is a matter of c*riterion-related validity*. The external variable or criterion is some type of outcome or indication of success in a particular context. For example, whether an admissions test for graduate school predicts later graduation is a question of criterion-related validity. Another example is whether a screening test constructed to detect a particular medical disorder can actually do so with reasonable accuracy. Criterion-related validity is often assessed through the computation of regression coefficients that estimate the magnitude of the association between test scores and measures of the criterion. Unreliability in either test or criterion scores reduces the absolute values of bivariate validity coefficients (e.g., Equation 6). Scores on the criterion variable should be valid, too; that is, they should actually measure the target outcome.

There are various types of criterion-related validity depending on when test scores versus criterion scores are collected. The term *concurrent validity* is used when scores on the test and criterion are collected at the same time, or within a very brief interval between the two measurements. *Predictive validity* describes the case when the criterion is measured at a later point in time. Whether test scores can appreciably predict an outcome that occurs later is the main question in a predictive validity study, such as whether an admissions test given to applicants predicts graduation months or even years later. *Postdictive validity* concerns whether test scores can predict something that has already occurred; that is, status on the criterion is already known before administering the test. This is how screening tests are usually developed: An initial version is administered to people known to have some characteristic, such as a disorder where objective diagnosis is possible, but perhaps expensive or invasive. The same test is also given to a control (normal) sample without the target characteristic. Whether screening test scores can differentiate the clinical versus control samples is an example of postdictive validity.

Convergent validity and discriminant validity involve the evaluation of measures against each other instead of against an external criterion. There is evidence for *convergent validity* if correlations between scores from two tests presumed to measure the *same* construct are of appreciable magnitude. For example, if tests X and Y are both supposed to measure depression but $r_{XY} = .02$ (i.e., practically zero), it clear that these two tests measure nothing in common; that is, the correlation evidence speaks against convergent validity. Likewise, *discriminant validity* is supported if correlations between scores from two tests that are supposed to measure *different* constructs are not too high in magnitude. If $r_{XY} = .95$ (i.e., practically 1.0) and X is intended as a measure of, say, self-confidence and test Y is supposed to reflect abstract thinking, the evidence says that

the two tests measure the same thing, not different constructs. There is little support for the hypothesis of discriminant validity in this example.

Evidence about convergent validity or discriminant validity is stronger if each of the two tests are based on a different method of measurement. A method can be a *medium* for gathering the information, such as the self-report method, observational method (i.e., an observer scores the test), or interview method, whereby an interviewer poses the questions and records the responses. A method can also be a *source* of information, such as when adolescents, parents, and teachers complete rating scales about the adjustment status of the adolescents. Archival data, such as medical records, are another source of information. Finally, the *format* for giving a test, such as computer (online) administration versus paper-and-pencil administration in either a group setting or when examinees are individually tested, is another form of method. Elements of methods can be combined, such as when test items are presented to respondents on the computer.

Common method variance (CMV) is systematic variation in test scores attributable to a particular measurement method instead of due to the target construct. Method variance can spuriously inflate convergent validity coefficients or discriminant validity coefficients. For example, suppose that two different self-report tests, *X* and *Y*, are each supposed to measure depression. After administering both tests in the same sample, it is found that r_{XY} = .70, a value considered to be meaningfully high. Now, what explains this result? One possibility is that both tests measure a common trait, but another likelihood is that at least some part of the value of r_{XY} is due to CMV, in this case, a systematic effect of the self-report method that has nothing to do with depression. This concern is well founded because there are forms of response styles or biases that may be associated with self-report, such as the tendency to present oneself in a socially desirable light or the tendency to agree with or endorse questionnaire items regardless of content. Both response styles are sources of systematic variation that may have little to do with a target construct.

The best evidence for convergent validity occurs when measures of the same presumed trait are each based on a different method. Likewise, evidence for discriminant validity is stronger when measures that are supposed to assess different traits are each based on a different method. For instance, if r_{XY} = .95 and tests *X* and *Y* each rely on a different method, one could not claim that *X* and *Y* assess different constructs because the validity coefficient was not inflated by CMV. This logic provides the basic rationale for a *multitrait-multimethod study*—first described by Donald T. Campbell and Donald W. Fiske (1959)—in which ≥ 2 traits are measured by tests that rely on ≥ 2 methods. The goals are to (a) evaluate the convergent validity and discriminant validity of tests that vary in their measurement method and (b) derive separate estimates of the effects of traits versus methods on the observed scores.

The concept of construct validity was extended by Samuel Messick (1995) who emphasized both score meaning and social values in test use and interpretation. Social values concern whether a particular interpretation of test scores produces desired social consequences, such as the accurate assessment of scholastic skills among minority children. Such considerations and others, such as those about distributive justice, absence of bias, or fairness, correspond to what Messick (1995) described as *consequential validity*, which involves consideration of both psychometrics and the outcomes of test use in a particular context from

a social perspective. Not all measurement experts agree that social consequences are properly a part of the concept of validity, but Messick's (1995) ideas have been influential in education. There is also a literature about *test fairness*, which is also concerned with the social consequences of test use (Karami & Mok, 2013).

Statistical Methods for Validity Analyses

Many types of statistical techniques can be applied in validity analyses, but just a few of the major methods are mentioned here. Regression techniques are widely used in analyses that concern criterion-related validity. Test scores and other predictors, such as demographic variables, can be specified as predictions of a criterion of interest in a multiple regression analysis. Factor analysis, both exploratory and confirmatory, is also widely used to evaluate construct validity (see Brown, 2015 for more information).

Translating Tests

Most psychological tests are developed in the United States and are available in English only. Unfortunately, it is not a simple matter to translate a test into another language. Differences in language may also come with differences in culture—if a translated test is used in a different country—and there are cultural differences in constructs such as the definition or experience of depression (Ryder et al., 2009). Thus, it is generally unrealistic to expect that scores from a translated test will have the same meaning as scores from the original test.

It is a myth that anyone who knows both languages is capable of producing a suitable translation. There is variability among the skills of even professional translators, and different translators working from the same source will not typically produce the same translation. It is also challenging when professional translators are not familiar with more specialized or technical vocabulary related to construct definition. Use of a back-translation procedure, whereby the original test is translated into another language and the translated version is then translated back to the original language, early in the translation process may help to avoid mistranslations that appreciably change the meaning of test content (see van Widenfelt, Treffers, de Beurs, Siebelink, & Koudijs, 2005).

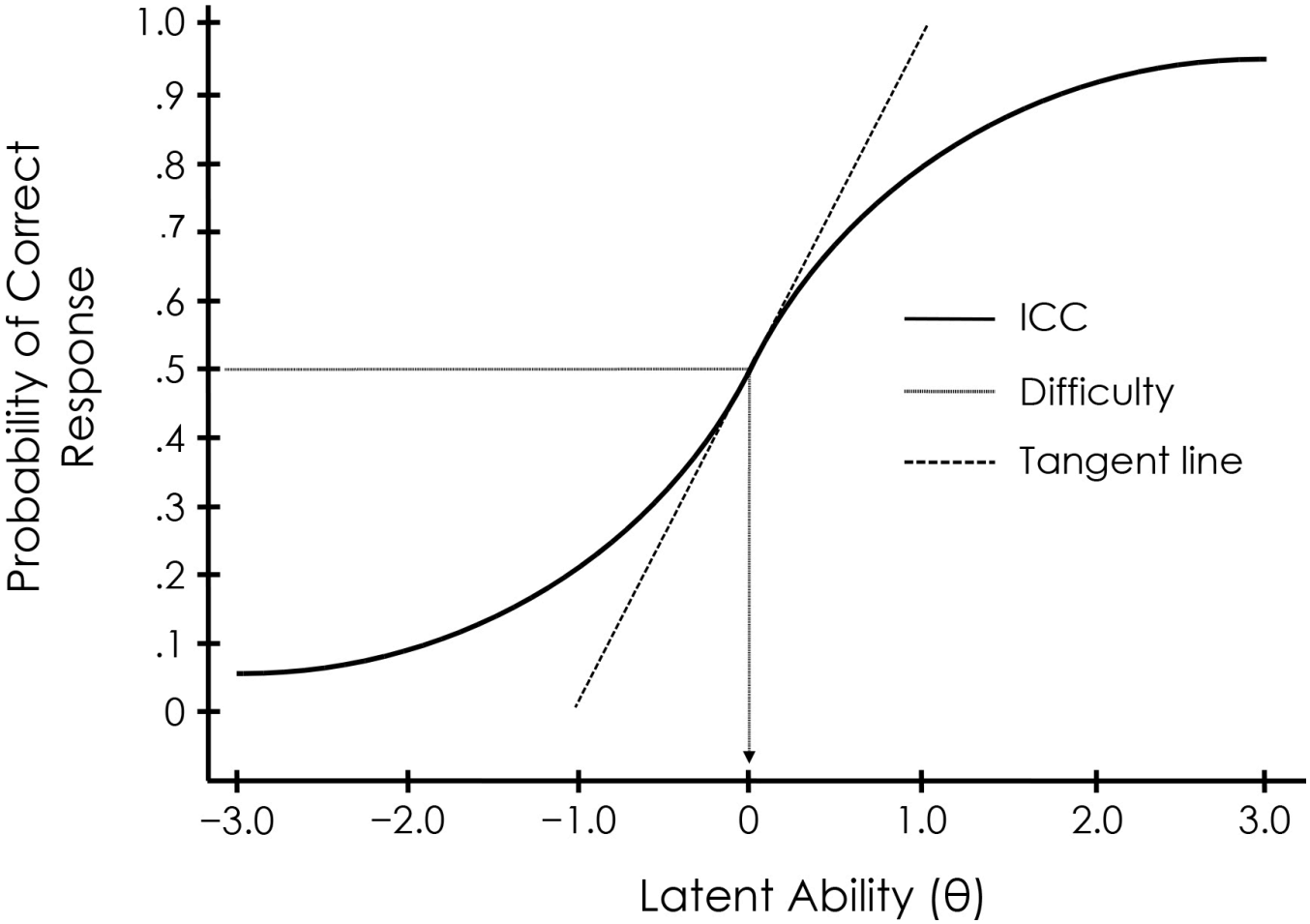
Up to three different types of bias may be introduced when translating a test. These include (a) *construct bias*, where a translation ignores cross-cultural differences in construct definition; (b) *method bias*, where the impact of how a test is administered (e.g., by an adult authority figure) is greater in one culture than another; and (c) *item bias*, where items fail to have constant statistical properties over translations or cultures (Van de Vijver & Hambleton, 1996). Some of the statistical methods for evaluating validity mentioned in the previous section can be applied to detect these forms of bias.

Modern Psychometrics

The test–retest, alternate forms, and interrater reliability methods in classical measurement theory are restricted to the study of, respectively, two times, forms, or raters at a time. Each type of measurement error—time, content, or rater sampling error—is usually estimated in separate studies (Table 1). But in *generalizability theory*, different sources of measurement error can be simultaneously estimated. Also, >2 times, forms, or raters can be included in a generalizability study. For example, the consistency of scores over four different occasions and over five raters can be analyzed together in a generalizability study. It is also possible to estimate *measurement error interaction effects*, which are joint effects between different sources of error that create yet more score imprecision than either source alone. For instance, interrater reliability could be higher or lower on a particular alternate form compared with other versions of the same test (see Thompson, 2003 for more information).

Item response theory (IRT) consists of mathematical models that relate responses on individual test items to a continuous latent variable θ. The graphical representation of this relation is an *item characteristic curve* (ICC), an example of which is presented in Figure 1 for a dichotomously scored item (0 = *incorrect*, 1 = *correct*). The ICC in the figure depicts a *three-parameter model*, in which the parameters are item difficulty, discrimination, and guessing. Difficulty is the level of θ that corresponds to a 50% chance of getting the item correct. This value in the figure is θ = 0, or the mean in a standardized metric. Discrimination is the slope of the tangent line on the ICC at the level of item difficulty. The steeper the slope, the more discriminating the item, which also indicates a stronger relation with θ. Guessing is the >0 probability that a respondent with a low standing on θ would get the item correct (see Figure 1). There are also *two-parameter models* with just item difficulty and discrimination parameters, where guessing effects are assumed to be nil, and *one-parameter models*—also called *Rasch models* after the Danish mathematician Georg Rasch—with just the difficulty parameter, whereby discrimination is assumed to be equal across the items.

Figure 1. Item characteristic curve for the predicted probability of a correct response for a dichotomously scored item in a three-parameter item response theory model. Difficulty is θ = 0, discrimination is the slope of the tangent line at θ = 0, and guessing is the probability of a correct response that is greater than zero at the lowest ability levels.



There are many applications of IRT. One is *tailored testing*, whereby the computer presents sets of items that are tailored for each respondent, given the pattern of correct versus incorrect responses to that point. The computer also estimates a reliability coefficient for each individual respondent, given the particular set of items administered. Another application is *test equating*, which is a statistical process for determining comparable scores on ≥ 2 different forms of a test. The IRT framework offers powerful and flexible methods for analyzing data from psychological tests, but very large samples are required (Baylor et al., 2011, offers a gentle introduction to IRT).

Reporting Practices and Standards

The state of reporting about psychometrics in published studies tends to be poor, especially regarding score reliability. For example, in a review of 47 meta-analyses of results from about 13,000 primary studies in which scores from psychological tests were analyzed, Vacha-Haase and Thompson (2011) found that 55% of authors did not even mention score reliability. This means that authors generally failed to reassure their readers that the scores analyzed were precise. In about 15% of studies reviewed by Vacha-Haase and Thompson (2011), authors merely reported values of reliability coefficients from other sources, such as test manuals. Inferring from reliability coefficients derived in other samples, such as the standardization sample for a published test, to a different population is called *reliability induction*. Such reasoning about

the generalizability of reliability coefficients requires explicit justification, especially if the composition of the researcher's sample is not comparable to other samples in which reliability coefficients were derived. Few researchers directly compare characteristics of their sample with those from cited studies of score reliability.

The cognitive error described earlier that reliability is a property of tests instead of scores in particular samples may explain part of the problem. Another is the *measurement crisis*, or the substantial decline in the quality of instruction about psychometrics since the mid-1980s, a time period during which courses on measurement disappeared from many undergraduate and graduate programs in psychology, education, and related disciplines (Lambert, 1991). For instance, Leona S. Aiken, Stephen G. West, Leo Sechest, and Raymond R. Reno (1990) found that one third of psychology PhD programs in North America offered no formal training in measurement, and the directors of only about one quarter of these programs judged that their students were competent in psychometrics. James Frederich, Evelyn Buday, and David Kerr (2000) found that measurement courses were not offered in most undergraduate psychology programs including programs at "elite" universities. This state of affairs puts researchers in a difficult spot: They may be expected to select measures for their research, but some may lack the skills needed in order to critically evaluate the scores generated by those measures.

The American Psychological Association (APA) Publications and Communications Board Working Group on Journal Article Reporting Standards (JARS; 2008) published a set of reporting standards for manuscripts submitted to psychology research journals that report new data collections or meta-analyses. Other professional groups or associations have also established reporting standards for various types of research, including the Consolidated Standards of Reporting Trials (CONSORT) group for randomized clinical trials (Schulz et al., 2010) and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses group for summaries of data from primary studies (Moher et al., 2009), among others. Reporting standards are summaries of best practices for describing empirical results from scientific studies in complete and transparent ways.

Recently the APA Publications and Communications Task Force Working Group described updated reporting standards for quantitative studies, which are now called JARS—Quant (Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu, & Rao, 2018). There is stronger emphasis on the importance of complete reporting about psychometrics in the revised standards. Shown in Table 2 are the standards from the psychometrics section of JARS—Quant, and these standards are consistent with principles for reporting about psychometrics discussed to this point. Briefly summarized, the standards in Table 2 call on authors to report values of reliability coefficients for the scores analyzed, describe the specifics of those reliability coefficients (e.g., state the length of the retest interval for test–retest reliability coefficients) and, if reporting values of reliability or validity coefficients derived in other samples, describe the demographic characteristics of those other samples.

Table 1. Summary of classical methods for estimating score reliability.

Method	Fraction of Test (k)	No. of Occasions	Description	Type of Error Variance Estimated
Test–retest	1.0	2	Repeat same test on two different occasions	Time sampling
Interrater	1.0	1	Test scored by two different raters	Rater sampling
Alternate forms	2.0	1	Administer two equivalent forms of test	Content sampling
Split-half	.5	1	Split test items into two equivalent halves	Content sampling, type of split
Internal consistency	$1/N_i$	1	Split test into as many parts as items, N_i	Content sampling, test heterogeneity, test length

Table 2. American Psychological Association revised journal article reporting standards for psychometrics.

Estimate and report values of reliability coefficients for the scores analyzed (i.e., the researcher's sample), if possible.
Provide estimates of convergent and discriminant validity where relevant. Report estimates related to the reliability of measures, including
Interrater reliability for subjectively scored measures and ratings.
Test–retest coefficients in longitudinal studies in which the retest interval corresponds to the measurement schedule used in the study.
Internal consistency coefficients for composite scales in which these indices are appropriate for understanding the nature of the instruments being employed in the study.
Report the basic demographic characteristics of other samples if reporting reliability or validity coefficients from those sample(s), such as those described in test manuals or in the norming information about the instrument.

Source: Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu, & Rao (2018, pp. 3–25).

Final Thoughts

Elazar Pedhazur and Liora Schmelkin (1991) claimed that "measurement is the Achilles' heel of sociobehavioral research" (p. 2). Without proper attention to psychometrics, this statement is justified. Best practice is to analyze the reliability of the scores analyzed in a particular sample and to report the values of reliability coefficients along the rest of the results. Values of reliability coefficients derived in other samples,

such as test standardization samples, can be reported, too, but explicit comparison of the researcher's sample and those other samples should be made. Authors of empirical studies often say nothing about score reliability, which is a serious omission. Validity concerns the accuracy of interpretations of test scores in a particular context of use. Test scores may be valid for one purpose or setting but not in another, so validity—just as reliability—is not a fixed characteristic of tests. Both reliability and validity are usually established over a series of empirical studies, not in any single investigation.

References

Aiken, L. S., **West, S. G.**, **Sechrest, L.**, & **Reno, R. R.** (1990). Measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, 45, 721–734. doi:10.1037/0003-066X.45.6.721

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63, 839–851. doi:10.1037/0003-066X.63.9.839

Appelbaum, M., **Cooper, H.**, **Kline, R. B.**, **Mayo-Wilson, E.**, **Nezu, A. M.**, & **Rao, S. M.**, (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist*, 73, 3–25. doi:10.1037/amp0000191

Baylor, C., **Hula, W.**, **Donovan, N. J.**, **Doyle, P. J.**, **Kendall, D.**, & **Yorkston, K.** (2011). An introduction to item response theory and Rasch models for speech-language pathologists. *American Journal of Speech–Language Pathology*, 20, 243–259. doi:10.1044/1058-0360(2011/10-0079)

Binet, A., & **Simon, T.** (1916). New methods for the diagnosis of the intellectual level of subnormals. In E. S. Kite (Trans., Ed.), *The development of intelligence in children* (pp. 191–244). Vineland, NJ: Publications of the Training School at Vineland.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322. doi:10.1111/j.2044-8295.1910.tb00207.x

Campbell, D. T., & **Fiske, D. W.** (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56, 81–105. doi:10.1037/h0046016

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi:10.1007/BF02310555

Frederich, J., **Buday, E.**, & **Kerr, D.** (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology*, 27, 248–257. doi:10.1207/S15328023TOP2704_02

Gadermann, A. M., **Guhn, M.**, & **Zumbo, B. D.** (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, 17. Retrieved from http://pareonline.net/

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika*, 15, 259–269. doi:10.1007/BF02289042

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society*, *Series A*, 159, 445–492. doi:10.2307/2983326

Hancock, G. R., & **Mueller, R. O.** (2001). Rethinking construct reliability within latent variable systems. In **R. Cudeck**, **S. du Toit**, & **D. Sörbom** (Eds.), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Haynes, S. N., **Richard, D. C. S.**, & **Kubany, E. S.** (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, 7, 238–247. doi:10.1037/1040-3590.7.3.238

Huck, S. W. (2016). *Statistical misconceptions*. New York, NY: Routledge.

Jones, L. V., & **Thissen, D.** (2007). A history and overview of psychometrics. In **C. R. Rao** & **S. Sinharay** (Eds.), *Handbook of statistics* (Vol. 26, pp. 1–27). Amsterdam, the Netherlands: Elsevier.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73. doi:10.1111/jedm.12000

Karami, H., & **Mok, M. M. C.** (Eds.). (2013). Fairness issues in educational assessment [Special issue]. *Educational Research and Evaluation*, 19.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Kuder, G. F., & **Richardson, M. W.** (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160. doi:10.1007/BF02288391

Lambert, N. M. (1991). The crisis in measurement literacy in psychology and education. *Educational Psychologist*, 26, 23–35. doi:10.1207/s15326985ep2601_2

Messick, S. (1995). Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749. doi:10.1037/0003-066X.50.9.741

Moher, D., **Liberati, A.**, **Tetzlaff, J.**, **Altman, D. G.**, & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology*, 62, 1006–1012. doi:10.1016/j.jclinepi.2009.06.005

Pedhazur, E. J., & **Schmelkin, L. P.** (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, 35, 299–331. doi:10.1016/S0005-7894(04)80041-8

Reynolds, C. R., & **Livingston, R. B.** (2012). *Mastering modern psychological testing: Theory & methods*. Boston, MA: Pearson.

Ryder, A. G., **Yang J.**, **Zhu, X.**, **Yao, S.**, **Yi, J.**, **Heine, S. J.**, & **Bagby, R. M.** (2009). The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? *Journal of Abnormal Psychology*, 117, 300–313. doi:10.1037/0021-843X.117.2.300

Schulz K., F., **Altman, D. G.**, **Moher, D.**, & the CONSORT Group. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine*, 152, 726–732. doi:10.7326/0003-4819-152-11-201006010-00232

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74, 107–120. doi:10.1007/S11336-008-9101-0

Spearman, C. (1904a). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201–293. doi:10.2307/1412107

Spearman, C. (1904b). The proof and measurement of the association between two things. *American Journal of Psychology*, 15, 72–101. doi:10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295. doi:10.1111/j.2044-8295.1910.tb00206.x

Tavakol, M., & **Dennick, R.** (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. doi:10.5116/ijme.4dfb.8dfd

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: SAGE.

Thompson, B., & **Vacha-Haase, T.** (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, 60, 174–195. doi:0.1177/00131640021970448

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago, IL: University of Chicago Press.

Vacha-Haase, T., & **Thompson, B.** (2011). Score reliability: A retrospective look back at 12 years of reliability generalization. *Measurement and Evaluation in Counseling and Development*, 44, 159–168. doi:10.1177/0748175611409845

Van de Vijver, F., & **Hambleton, R. K.** (1996). Translating tests: Some practical guidelines. *European Psychologist*, 1, 89–99. doi:10.1027/1016-9040.1.2.89

van Widenfelt, B. M., **Treffers, P. D. A.**, **de Beurs, E.**, **Siebelink, B. M.**, & **Koudijs, E.** (2005). Translation

and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical Child and Family Psychology Review*, 8, 135–147. doi:10.1007/s10567-005-4752-1

Warshaw, M. G., **Dyck, I.**, **Allsworth, J.**, **Stout, R. L.**, & **Keller, M. B.** (2001). Maintaining reliability in a long-term psychiatric study: An ongoing inter-rater reliability monitoring program using the longitudinal interval follow-up evaluation. *Journal of Psychiatric Research*, 35, 297–305. doi:10.1016/S0022-3956(01)00030-9

Williams, M. N., **Grajales, C. A. G.**, & **Kurkiewicz, D.** (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, 18. Retrieved from http://pareonline.net/