Reliability from α to ω: A Tutorial

William Revelle and David M. Condon
Northwestern University and University of Oregon

Abstract

Reliability is a fundamental problem for measurement in all of science. Although defined in multiple ways, and estimated in even more ways, the basic concepts seem straight forward and need to be understood by practioners as well as methodologists. Reliability theory is not just for the psychometrician estimating latent variables, it is for everyone who wants to make inferences from measures of individuals or of groups. For the case of a single test administration, we consider multiple measures of reliability, ranging from the worst (β) to average (α, λ_3) to best (λ_4) split half reliabilities, and consider why model based estimates (ω_h, ω_t) should be reported. We also address the utility of test-retest and alternate form reliabilities. The advantages of immediate versus delayed retests to decompose observed score variance into specific, state, and trait scores is discussed. But reliability is not just for test scores, it is also important when evaluating the use of ratings. Estimates that may be applied to continuous data include a set of intraclass correlations while discrete categorical data needs to take advantage of the family of κ statistics. Examples of these various reliability estimates are given using state and trait measures of anxiety given with different delays and under different conditions. An online supplement is provided with more detail and elaboration. The online supplement is also used to demonstrate applications of open source software to examples of real data, and comparisons are made between the many types of reliability.

Public Significance: A tutorial on the estimation of the reliability of test scores considers classical and model based approaches. Examples using open source software applied to several real world data sets are provided.

Keywords: Reliability; Generalizability; Classical Test Theory; R packages

Reliability

Reliability is a fundamental problem for measurement in all of science for "(a)ll measurement is befuddled by error" (p 294 McNemar, 1946). Perhaps because psychological measures are more befuddled than those of the other natural sciences, psychologists have long studied the problem of reliability (Spearman, 1904b; Kuder & Richardson, 1937; Guttman, 1945; Lord, 1955; Cronbach,

contact: William Revelle revelle@northwestern.edu
Preparation of this manuscript was funded in part by grant SMA-1419324 from the National Science Foundation to WR. This is the revised version as submitted for review to *Psychological Assessment* on June 11, 2019.

1951; Feldt & Brennan, 1989; McDonald, 1999) and it remains an active topic of research (Sijtsma, 2009; Revelle & Zinbarg, 2009; Bentler, 2009; McNeish, 2017; Wood, Harms, Lowman, & DeSimone, 2017). Unfortunately, although recent advances in the theory and measurement of reliability have gone far beyond the earlier contributions, much of this literature is more technical than readable and is aimed for the specialist rather than the practitioner. (This is not a new problem, e.g., Anastasi, 1967; Glass, 1986, bemoan this tendency). We hope to remedy this issue somewhat, for an appreciation of the problems and importance of reliability is critical to the activity of measurement across many disciplines. Reliability theory is not just for the psychometrician estimating latent variables, but also for the baseball manager trying to predict how well a high performing player will perform the next year, for accurately estimating agreement among doctors in patient diagnoses, and in evaluations of the extent to which stock market advisors under-perform the market.

Issues of reliability are fundamental to understanding how correlations between observed variables are (attenuated) underestimates of the relationships between the underlying constructs, how observed estimates of a person's score are biased estimates of their latent score, and how to estimate the confidence intervals around any particular measurement. Understanding the many ways to estimate reliability as well as the ways to use these estimates allows one to better assess individuals and to evaluate selection and prediction techniques. This is not just a problem for measurement specialists but for all who want to make theoretical inferences from observed data. Schmidt & Hunter (1996) discuss 26 ways that not correcting for the effects of reliability and measurement error can hinder progress in many areas of psychological research. However, Borsboom & Mellenbergh (2002) take a contrary position and suggest that using classical measures of reliability for such corrections is an error.

The fundamental question in reliability is to what extent do scores measured at one time and place with one instrument predict scores at another time and/or place and perhaps measured with a different instrument? That is, given a person's score on test 1 at time 1, what score should be expected at a second measurement occasion? The naive belief is that if the tests measure the same construct, then people will do just as well on the second measure as they did on the first. This mistaken belief contributes to several errors including the common view that punishment improves and rewards diminish subsequent performance (Kahneman & Tversky, 1973) and other popular phenomena like the "sophomore slump" and the "Sports Illustrated jinx" (Schall & Smith, 2000). More formally, the expectation for the second measure is just the regression of observations at time 2 on the observations at time 1. If both the time 1 and time 2 measures are equally "befuddled by error" then the observed relationship *is* the reliability of the measure: the ratio of the latent score variance to the observed score variance.

Reliability as a variance ratio

The basic concept of reliability seems to be very simple: observed scores reflect an unknown mixture of signal and noise. To detect the signal, we need to reduce the noise. Reliability thus defined is a function of the ratio of signal to noise. The signal might be something as esoteric as a gravity wave produced by a collision of two black holes, or as prosaic as estimating the expected batting average of a baseball player based upon the performance of the prior year. The noise in gravity wave detectors include the seismic effects of cows wandering in fields near the detector as well as passing ice cream trucks. The noise in batting averages include the effect of opposing pitchers, variations in wind direction, and the effects of jet lag and sleep deprivation. We can enhance the signal/noise ratio by either increasing the signal or reducing the noise. Unfortunately, this classic statement of reliability ignores the need for unidimensionality of our measures and equates expected scores with construct scores, a relationship that needs to be tested rather than

assumed (Borsboom & Mellenbergh, 2002).

We can credit Spearman (1904b) for the original formalization of reliability. In the first of two landmark papers (the other, Spearman, 1904a, laid the basis for factor analysis and measurement of cognitive ability) he developed the ordinal correlation coefficient and the basic principles of reliability theory. Spearman's fundamental insight was that an observed test score could be decomposed into two unobservable constructs: a *latent* score of interest and a residual but latent *error* score:

$$X = \chi + \epsilon. \tag{1}$$

Reliability was defined as the fraction of an observed score variance that was not error:

$$r_{xx} = \frac{V_X - \sigma_\epsilon^2}{V_X} = 1 - \frac{\sigma^2\epsilon}{V_X}. \tag{2}$$

If the test is unidimensional the product of the observed score variance and the reliability is an estimate of the variance of the *latent construct* (sometimes called the "true score")[1] in a test which we will symbolize[2] as $\sigma_\chi^2 = V_X - \sigma_\epsilon^2 = r_{xx}V_X$. Spearman (1904b) developed reliability theory because he was interested in correcting the observed correlation between two ability tests for their lack of reliability. In modern terminology, this disattenuated correlation ($\rho_{\chi\eta}$) represents the correlation between two latent variables (χ and η) estimated by the correlation of two observed tests (r_{xy}) corrected for the reliability of the observed tests (r_{xx} and r_{yy}) (see Figure 1).

$$\rho_{\chi\eta} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}. \tag{3}$$

Furthermore, given an observed score, the variance of the error of that score (σ_ϵ^2) is just the observed test variance times one minus the reliability and thus the standard deviation of the error associated with that score (the *standard error of measurement*) is:

$$\sigma_\epsilon = \sigma_x\sqrt{1 - r_{xx}}. \tag{4}$$

Although expressed as a correlation between observed scores, reliability is a ratio of reliable variance to total variance. In addition, because the covariance of the latent score with the observed score is just the reliable variance, the predicted latent score $\hat{\chi}$ is

$$\hat{\chi} = r_{xx}x \tag{5}$$

where x is the raw deviation score ($x = X - \bar{X}$). From Equation 4, we know the standard error of measurement and can give a confidence interval for our observed score and even for the estimated latent score (Charter & Feldt, 2001):

$$\hat{\chi}_i = r_{xx}x_i \pm t_{\alpha/2,df}\sigma_x\sqrt{r_{xx}(1 - r_{xx})} \tag{6}$$

where $t_{\alpha/2,df}$ represents Student's t with an appropriate probability level (e.g., $\alpha = .05$).

Increasing reliability reduces the standard error of measurement (Equation 4) and increases the observed correlation with external variables (Equation 3). That is, if we knew the reliabilities, we could correct the observed correlation to find the latent correlation and estimate the precision of our measurement. The problem for Spearman was, and remains for us today, how to find reliability?

[1] As discussed by Lord & Novick (1968) and Borsboom & Mellenbergh (2002), classical test theories "true score" is just the expected value and should not be confused with Platonic Truth.

[2] Observed variables and correlations are shown in conventional Roman fonts, latent variables and latent paths in Greek fonts.

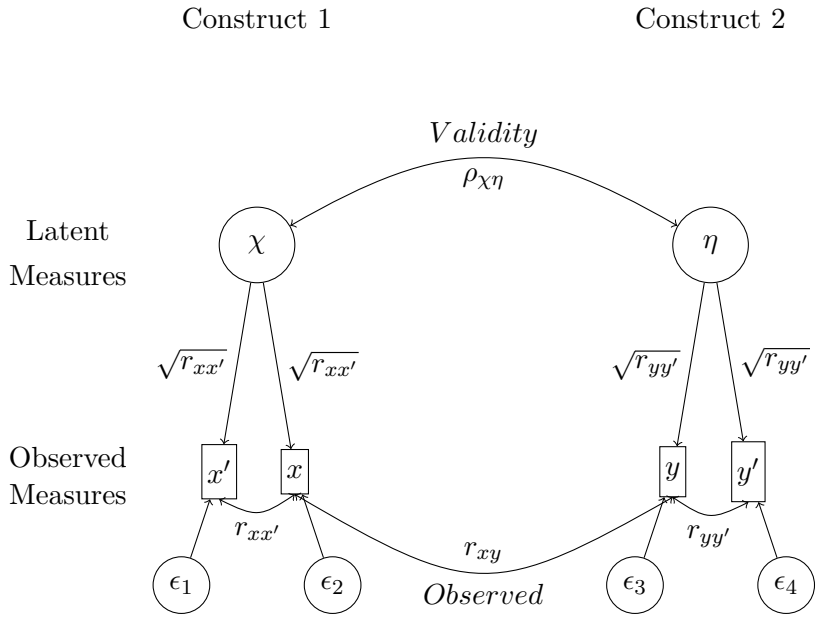

Figure 1. The basic concept of reliability and correcting for attenuation. Adjusting observed correlations (r_{xy}) by reliabilities ($r_{xx'}$, $r_{yy'}$) estimates underlying latent correlations ($\rho_{\chi\eta}$). (See Equation 3). Observed variables and correlations are shown in conventional Roman fonts, latent variables and latent paths in Greek fonts.

Equations 1 - 6 are intellectually interesting, but not very helpful, for they decompose an observed measure into the two unobservable variables of latent score and latent error. To make it even more complicated, all tests are assumed to measure something stable over time (denoted as T for *trait* like), something that varies over time (reflecting the current *state* and denoted as S), some specific variance (s) that is stable but does not measure our trait of interest, and some residual, random error (E) (Baltes, 1987; Cattell, 1966b; Hamaker, Schuurman, & Zijlmans, 2017; Kenny & Zautra, 1995).

Although ultimately interested in the *precision* of a score for each individual, reliability is expressed as a ratio of variances between individuals[3]: The reliability of X (r_{xx}) is just the percentage of total variability (V_X) that is not error. Unless we have repeated measures, error is an unknown mixture of variability due to differences in item means, the person x item interaction, and some un-modeled residual. The reliable between person variance is a mixture of Trait, State, and specific variance. (For instance, an item such as "I enjoy a lively party" is an unknown mixture of trait extraversion, state positive affect, and the specific wording of the item – how one interprets lively and party.) If we are interested in the stable trait scores, reliability is the ratio of (unobservable) trait variance (σ_T^2) to (observable) total test variance (V_x). (We use σ^2 to represent unobservable variances, V to represent observable variance.) That is,

$$r_{tt} = \frac{\sigma_T^2}{V_X} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_S^2 + \sigma_s^2 + \sigma_e^2}. \tag{7}$$

[3]We can also find *within* subject reliability across time. This will be discussed later and in the online supplement.

However, if we are interested in how well we are measuring a particular fluctuating state (e.g. an emotion) we want to know

$$r_{ss} = \frac{\sigma_S^2}{V_X} = \frac{\sigma_S^2}{\sigma_T^2 + \sigma_S^2 + \sigma_s^2 + \sigma_e^2}. \quad (8)$$

The problem becomes how to find σ_T^2 or σ_S^2 and how to separate their effects. Although Trait scores are thought to be stable over time, State scores, while fluctuating, show some (unknown) short term temporal stability. Consider a measure of depression. Part of an individual's depression score will reflect long term trait neuroticism and some of it reflects current negative emotional state. Two measures taken a few hours apart should produce similar trait and state values, although measures taken a year apart should reflect just the trait.

In all cases, we are interested in the scores for the individuals being measured. To make the problem even more complicated, it is likely that our Trait or State scores reflect some aggregation of item responses or of the ratings of judges. Thus, we want to assess the variance due to Traits or States that is independent of the effects of items or judges, how much variance is due to the items or judges, and finally how much variance is due to the interactions of items/judges with the Trait/State measures[4]. To be consistent with much of the literature, we will treat Trait and State as both latent sources of variance for the observed score X and refer to Trait as a stable across time and State as varying across time. We recognize, of course that Traits do change over longer periods of time but will use this stable/unstable distinction for relatively short temporal durations. Although some prefer to think of specific variance (σ_s^2) and error variance (σ_e^2) as hopelessly confounded, we prefer to separate them for there are some designs (e.g., test-retest vs. parallel forms) that allow us to distinguish them.

Reliability as defined in equations 7 and 8 is not just a function of the test, but also of who is being tested, where they are tested and when they are tested. Because it is a variance ratio, increasing between person variance without increasing the error variance will increase the reliability. Similarly, decreasing between person variance will decrease reliability. Generalizabilty theory (Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach, & Rajaratnam, 1965) is one way to estimate the individual variance components rather than their ratio. Another approach is Item Response Theory (e.g., Embretson, 1996; Lord & Novick, 1968; Lumsden, 1976; Mellenbergh, 1996; Rasch, 1966; Reise & Waller, 2009) which addresses this problem by attempting to get a measure of precision for a person's estimate that is independent of the variance of the population and depends upon just the probability of a particular person answering particular items.

Consistency, reliability and the data box

When Cattell (1946) introduced the *data box* it was a three way organization of measures taken over people, tests, and time. In typical Cattellian fashion, over the years this simple idea grew to as many as 10 dimensions (Cattell, 1966a; Cattell & Tsujioka, 1964). However, the three primary distinctions are still useful today (Nesselroade & Molenaar, 2016; Revelle & Wilt, 2016). Using these dimensions, Cattell (1964) distinguished between three ways that tests can be consistent: across occasions (reliability), across items (homogeneity), and across people (transferability or hardiness). We consider the first two of these concepts and leave the latter to a discussion of validity. These various types of reliability may be summarized graphically in terms of latent traits, paths, observed variables and correlations (Figure 2).

[4]Unfortunately, some prefer to use State to reflect the measure at a particular time point and to decompose this "State" into Trait and Occasion components (Cole, Martin, & Steiger, 2005).

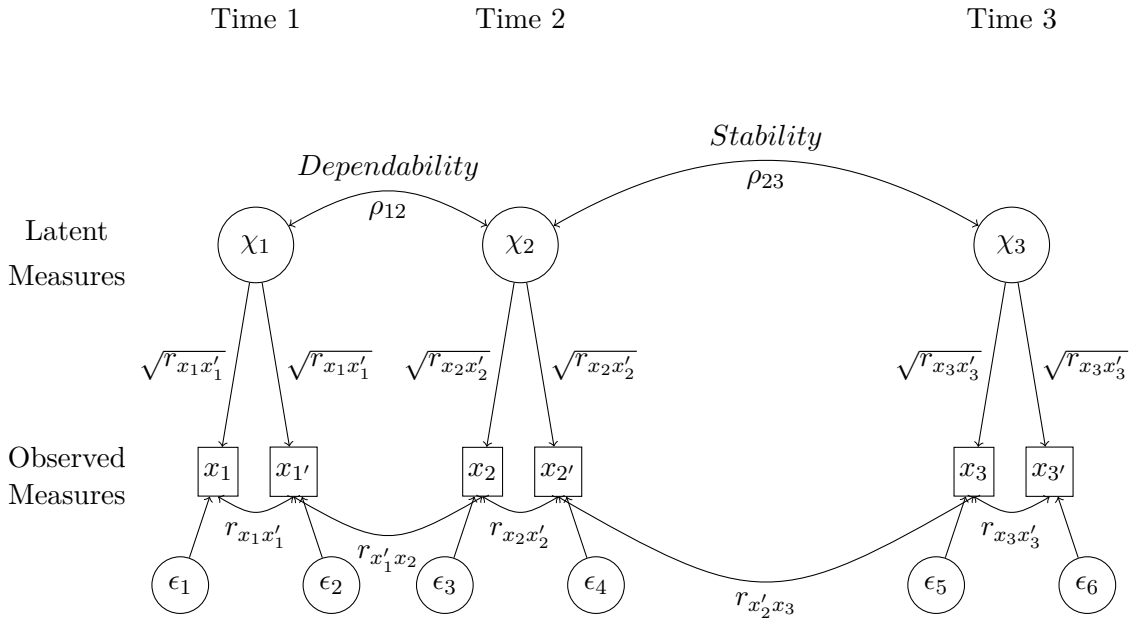

Figure 2. Reliability has many forms. Items within the same test provide estimates of *internal consistency*. Observed correlation between *alternate forms* or parallel tests given at the same time ($r_{x_1x'_1}$, $r_{x_2x'_2}$) estimate parallel test reliability. Tests given at (almost) the same time (times 1 and 2 e.g., $r_{x_1x_2}$, $r_{x'_1x'_2}$) provide measures of *dependabilty*, while measures taken over a longer period (times 2 and 3, e.g., $r_{x_1x_3}$, $r_{x_2x_3}$) are measures of *stability*. These measures differ in the amount of Trait and State and specific variance in the measures. Observed variables and correlations are shown in conventional Roman fonts, latent variables and latent paths in Greek fonts.

Alternative estimates of reliability

Test-retest of total scores

Perhaps the most obvious measure of reliability is the correlation of test with the same test some time later. For Guttman (1945), this was reliability. If we have only two time points (t_1 and t_2), this correlation is an unknown mixture of Trait, State and specific variance and, additionally, a function of the length of time between the two measures:

$$r_{t_1t_2} = \frac{\sigma_T^2 + \tau^{t_2-t_1}\sigma_S^2 + \sigma_s^2}{\sigma_T^2 + \sigma_S^2 + \sigma_s^2 + \sigma_e^2} \tag{9}$$

where τ (the auto-correlation due to short term state consistency) is less than 1 and thus the state effect ($\tau^{t_2-t_1}\sigma_S^2$) will become smaller the greater the time lag. If the intervening time is long enough that the State effect is minimal, we will still have specific variance, and the correlation of a test with the same test later is

$$r_{xx} = \frac{\sigma_T^2 + \sigma_s^2}{V_x} = \frac{\sigma_T^2 + \sigma_s^2}{\sigma_T^2 + \sigma_S^2 + \sigma_s^2 + \sigma_e^2}. \tag{10}$$

Table 1
Steps toward reliability analysis: choosing the appropriate R function to find reliability. All functions except for the `cfa` *and* `cor` *function are in the psych package.*

Steps	Statistic	R function
Preliminaries		
Hypothesis development		
Data collection		
Data input		`read.file`
Data screening		
Descriptive statistics	μ, σ, range	`describe`
Analysis of internal structure		
Exploratory Factor Analysis	$\mathbf{R} = \mathbf{F}\phi\mathbf{F}' + \mathbf{U}^2$	`fa`
Hierarchical structure	ω_h, ω_t	`omega`, `omegaSem`
Confirmatory Factor Analysis		*lavaan*::`cfa`
Estimation of various reliabilities		
Items (dichotomous, polytomous or continuous)		
One occasion		
general factor saturation	ω_h	`omega`
total common variance	ω_t	`omega`
average interitem r	$\overline{r_{ij}}$	`omega`, `alpha`
median interitem r		`omega`, `alpha`
mean test retest (tau equivalent)	α, λ_3	`omega`, `alpha`
smallest split half reliability	β	`splitHalf`, `iclust`
greatest split half reliability	λ_4	`splitHalf`, `guttman`
Two occasions		
test-retest correlation	r	`cor`
variance components	$\sigma_p^2, \sigma_i^2, \sigma_t^2$	`testRetest`
Multiple occasions		
within subject reliability	α	`multilevel.reliability`
variance components	$\sigma_p^2, \sigma_i^2, \sigma_t^2$	`multilevel.reliability`
Ratings (Ordinal or Interval)		
Single rater reliability	$ICC_{1..3}1$	`ICC`
Multiple rater reliability	$ICC_{1..3}k$	`ICC`
Ratings (Categorical)		
Two raters	κ	`cohen.kappa`

An impressive example of a correlation of the same measure over time is the correlation of .66 of ability as measured by the Moray House Exam at age 11 with the same test given to the same participants 69 years later when they were 80 years of age (Deary, Whiteman, Starr, Whalley, & Fox, 2004). This correlation was partially attenuated due to restriction of range for the 80 year old participants. (The less able 11 year olds were less likely to appear in the 80 year old sample.) When correcting for this restriction (Sackett & Yang, 2000), the correlation was found to be .73.

But the Scottish Longitudinal Study is unusually long, and is it more common to take test-retests over much shorter periods of time. In most cases it is important that we do not assume that the State effect is 0 (Chmielewski & Watson, 2009). It is more typical to find a pattern of correlations diminishing as a function of the time lag but not asymptotically approaching zero (Cole et al., 2005; Damian, Spengler, Sutu, & Roberts, 2018). This pattern is taken to represent a mixture of stable Trait variance and diminishing State effects such that the test-retest reliability across two time periods as shown in Equation 9 will become smaller the greater the time lag. Unfortunately, with only two time points we can not distinguish between the Trait and State effects. However, with three or more time points ($t_1, t_2, t_3, ..., t_n$), we can decompose the resulting correlations ($r_{x_1x_2}$,$r_{x_1x_3}$, $r_{x_2x_3}$, ...), into Trait and State components using Structural Equation Modeling (SEM) procedures (Hamaker et al., 2017) or simple path tracing rules (Chmielewski & Watson, 2009) and the resolution continues to improve with four or more time points (Cole et al., 2005; Kenny & Zautra, 1995). (See Figure 1 in the online supplement).

A large test-retest correlation over a long period of time indicates temporal *stability* (Boyle, Stankov, & Cattell, 1995; Cattell, 1964; Chmielewski & Watson, 2009). This should be expected if we are assessing something trait like (such as cognitive ability or perhaps emotional stability or extraversion) but not if we are assessing something thought to represent an emotional state (e.g., alertness or arousal). Because we are talking about correlations, mean levels can increase or decrease over time with no change in the correlation[5]. Measures of trait stability are a mixture of immediate test-retest *dependability* and longer term trait effects (Cattell, 1964; Chmielewski & Watson, 2009). For Boyle et al. (1995) and Cattell (1964), dependability was the immediate test-retest correlation, for Chmielewski & Watson (2009) the time lag of two weeks is considered an index of dependability. To Wood et al. (2017), dependability is assessed by repeating the same items later in one testing session.

All of these indicators of dependability and stability are in contradiction to the long held belief that a problem with test-retest reliability is that it introduces memory effects of learning and practice (Kuder & Richardson, 1937). As evidence for the memory effect, Wood et al. (2017) reports that the average response times to the second administration of identical items in the same session is about 80% of the time of the first administration.

In the online supplement (Table 1) we compare multiple estimates of reliability for three different example data sets available in the *psychTools* package (Revelle, 2019b) in the R open source statistical system (R Core Team, 2019). We describe these data sets in some detail as they are useful demonstrations of trait and state variations. We compare immediate retest to short (45 minutes) and then longer delay (one to seven days) on 10 mood items and 9-24 item trait scales for one to four weeks.

To compare the effects of an immediate retest versus a short delay versus a somewhat longer delay, consider the `msqR`, `sai` and `tai` example data sets[6] the analyses discussed below are demon-

[5]For example, participants in the Scottish Longitudinal Study performed better in adulthood than they did as 11 year olds but the correlations showed remarkable stability.

[6]The data 4, ($N > 4,000$) were collected as part of a long term series of studies of the interrelationships between

strated in the supplementary online material. We will use some of these items in the subsequent examples evaluating and comparing the immediate dependability and the 45 minute and multi-day stability coefficients of these measures. 10 items were given as part of the STAI and then immediately given again (with a slightly different wording) as part of the MSQ. Five of the items were scored in a positive (anxious) direction, five in the negative (non-anxious) direction.

Components of variance estimated by test-retest measures

A powerful advantage of repeating items is that it allows for an assessment of subject consistency across time (the correlation for each subject of their pattern of responses across the two administrations) as well as the consistency of the items (the correlation across subjects of responses to each item) (DeSimone, 2015; Wood et al., 2017). This allows for identification of unstable items and inconsistent responders. In addition, by using multi-level analyses[7] it is possible to estimate the variance components due to people, items, the person x item interaction, time, the person x time interaction, and the residual (error) variance (DeSimone, 2015; Revelle & Wilt, 2019; Shrout & Lane, 2012). This is implemented for example as the `testRetest` function in the *psych* package. The responses to any particular item can be thought to represent multiple sources of variance, and the reliability of a test made up of items is thus a function of those individual sources of variance. If we let P_i represent the i_{th} person, I_j the j_{th} item, T_k the first or second administration of the item, then the response to any item (e.g., of an anxiety inventory) is a function of the mean level of all of the items in the test, the trait level of the person taking the item, the difficulty or average endorsement frequency of the item, any temporal effects of a first or second administration, and all the possible interactions between these terms:

$$X_{ijk} = \mu + P_i + I_j + T_k + P_iI_j + P_iT_k + I_jT_k + P_iI_jT_k + \epsilon. \tag{11}$$

With complete data, we can find these components using conventional repeated measures analysis of variance of the data (i.e., `aov` in core R) or using multi-level functions such as `lmer` in the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) for R. As an example of such a variance decomposition consider the 10 overlapping mood items discussed in the online supplement (Table 1). 19% of the variance of the anxiety scores was due to between person variability, 25% to the very short period of time, 19% to the interaction of person by time, etc. and 13% was residual (unexplained) variance.

Multi-level modeling approaches are particularly appropriate if repeating the same measure multiple times (e.g., in an experience sampling study: Bolger & Laurenceau, 2013; Fisher, 2015; Mehl & Conner, 2012; Mehl & Robbins, 2012; Wilt, Funkhouser, & Revelle, 2011; Wilt, Bleidorn, & Revelle, 2016, 2017). We can derive multiple measures of reliability, across subjects, across time, across items and the various person x time, person x items, time x item interactions (Cranford et al.,

---

the stable personality dimensions of extraversion, neuroticism, and impulsivity (e.g., Eysenck & Eysenck, 1964; Revelle, Humphreys, Simon, & Gilliland, 1980), situational stressors (caffeine, time of day, movie induced affect, e.g., Anderson & Revelle, 1983, 1994; Rafaeli, Rogers, & Revelle, 2007; Rafaeli & Revelle, 2006) momentary affective and motivational state (e.g. energetic and tense arousal (Thayer, 1978, 1989), state anxiety (Spielberger, Gorsuch, & Lushene, 1970)), and cognitive performance (Humphreys & Revelle, 1984). The Motivational State Questionnaire (MSQ, Revelle & Anderson, 1998) included 75 items taken from a number of mood questionnaires (e.g., Thayer, 1978; Larsen & Diener, 1992; Watson, Clark, & Tellegen, 1988) and had 10 anxiety items that overlapped with the state version of the State Trait Anxiety Inventory (Spielberger et al., 1970). See the online supplement for more details.

[7] Analytic strategies for analyzing such multi-level data have been given different names in a variety of fields and are known by a number of different terms such as the random effects or random coefficient models of economics, multi-level models of sociology and psychology, hierarchical linear models of education or more generally, mixed effects models (Fox, 2016).

2006; Shrout & Lane, 2012). This is implemented in the `multilevel.reliability` function and discussed in more detail in a tutorial for analyzing dynamic data (Revelle & Wilt, 2019) as well as in the online supplement. Although these variance components can be found using traditional repeated measures analysis of variance, it is more appropriate to use multi-level techniques, particularly in the case of missing data.

Stability needs to be adjusted for dependability and thus the .36 stability over two days of the SAI (online supplement Table 1) should be adjusted for the immediate dependability of .85 to suggest a two day stability of anxious mood of .42 which is notably similar to that of the state-trait correlation of .43. When measuring mood, we need to disentangle the episodic memory components of the state measure from the semantic memory involved when answering trait like questions (Cattell, 1964; Chmielewski & Watson, 2009). State measures of affectivity probably involve episodic memory whereas trait measures of similar constructs (e.g., trait anxiety or neuroticism) likely tap semantic memory (Klein, Cosmides, Tooby, & Chance, 2002). With only two measures of state anxiety and one of trait anxiety, we can not disentangle how much of the trait measure is state (Equation 9) but if we had more measures over longer periods of time we would be able to do so.

Alternate Forms

If we do not want to wait for a long time and we do not want to exactly repeat the same items, we can estimate reliability by creating another test (an alternate form) that has conceptually similar but semantically different items. If measuring the same construct (e.g. arithmetic performance) we can subtly duplicate items on each form and even match for possible difficulty of order effects (a1: what is 6+3?, a2: what is 4 + 5? versus b1: what is 3+6? and b2: what is 5 + 4 ?). Cattell (1964) discusses "Herringbone" consistency, which are essentially parallel forms: Each half of the test is made up of half of the items of multiple constructs, and each is duplicated in the other half (math, english, social studies). Although creating alternate forms by hand is tedious, it has become possible for ability items to generate alternate forms using computer Automatic Item Generation techniques (Embretson, 1999; Leon & Revelle, 1985; Loe & Rust, 2017).

Alternate forms given at the same time eliminate the effect of the specific item variance but do not remove any motivational state effect: Sometimes alternate forms can be developed when a longer test is split into multiple shorter tests. As an example of this, consider the `sai` data set discussed in the online supplement which includes 20 items, 10 of which overlapped with the `msqR` data set and were used for our examples of test-retest and repeated measure reliability. The other 10 can be thought of as an alternate form of the anxiety measure and indeed correlate .74 with the target items from the `sai` and `msqR`. These correlations are less than when we actually repeat the same items by correlating the overlapping items of the `sai` and `msqR` (.85); are almost identical when we consider their short term dependability (.76); but less than estimates of internal consistency such as α or the average split half reliability (.83 - .87).

Split half (adjusted for test length)

If we have gone to the trouble of developing two alternate forms for a concept, and then administered both forms to a sample of participants, it is logical to ask what is the reliability of the composite formed from both of these tests. That is, if we have the correlation between two five item tests, what would be the reliability of the composite 10 item test? With a bit of algebra, we can predict it using a formula developed by Spearman (1910) and Brown (1910):

$$r_{xx} = \frac{2 * r_{x_1x_2}}{1 + r_{x_1x_2}}. \tag{12}$$

It is important to note the correlation between the two parts ($r_{x_1x_2}$) is not the split half reliability, but is used to find the split half reliability (r_{xx}) found by the "Spearman-Brown prophecy formula" (Equation 12)

Given that we have written n items and formed them into two splits of length $n/2$, what if we formed a different split? How should we split the items into two groups? Odd/even, first half/last half, randomly? This is a combinatorially difficult problem, in that there are $\frac{n!}{2(n/2)!(n/2)!}$ unique ways to split a test into two equal parts. While there are only 126 possible splits for the 10 anxiety items discussed above, this becomes 6,435 for a 16 item ability test, 1,352,078 for the 24 item EPI Extraversion scale (Eysenck & Eysenck, 1964) and over 4.5 billion for a 36 item test. The `splitHalf` function will try all possible splits for tests of 16 items or less, and then sample 10,000 splits for tests longer than that. The distribution of all possible splits for the 10 state anxiety items discussed earlier show that greatest split-half reliability is .92, the average is .87, and the lowest is .66 (Figure 3 panel A). This is in contrast to all the possible splits of 16 ability items taken from the International Cognitive Ability Resource (ICAR, Condon & Revelle, 2014) collected as part of the SAPA project (Revelle, Wilt, & Rosenthal, 2010), (Revelle et al., 2016) where the greatest split half reliability was .87, the average is .83, and the lowest is .73 (Figure 3 panel B). The 24 items of the EPI show strong evidence for non-homogeneity, with a maximum split half reliability of .81, an average of .73, and a minimum of .42 (Figure 3 part C). This supports the criticism that the EPI E scale tends to measure the two barely related constructs of sociability and impulsivity (Rocklin & Revelle, 1981). The EPI-N scale, on the other hand, shows a maximum split half of .85, a mean of .80, and a minimum of .65, providing strong evidence for a relatively homogeneous scale (Figure 3 part D) Examining these various splits is one way to understand the homogeneity of the test. For if the various spits differ a great deal (e.g., the EPI E scale) this can be taken as a warning that the test is not unidimensional.

Internal consistency and domain sampling

All of the above procedures are finding the correlation between two forms or occasions of a test. But what if there is just one form and one occasion? The approaches that consider just one test are collectively known as internal consistency procedures but also borrow from the concepts of domain sampling and can use the variance decomposition techniques discussed earlier. Some of these techniques, e.g., Cronbach (1951); Guttman (1945); Kuder & Richardson (1937) were developed before advances in computational speed made it trivial to find the factor structure of tests, and were based upon test and item variances. These procedures (α, λ_3, KR20) were essentially short cuts for estimating reliability. The variance decomposition procedures continued this approach but expanded to be known as *generalizability theory* (Cronbach et al., 1963; Gleser et al., 1965; Vispoel, Morris, & Kilinc, 2018) and allow for the many reliability estimates discussed before.

There are a number of different approaches for estimating reliability when there is just one test and one time. The earliest was to split the test into two random split halves and then adjust the resulting correlation between these two splits using the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910).

Unfortunately, as we showed in Figure 3 not all random splits produce equal estimates. If we consider all of the items in the test to be randomly sampled from some larger domain (e.g., trait-descriptive adjectives sampled from all words in the Oxford Unabridged Dictionary or sociability items sampled from a potentially infinite number of ways of being sociable) then we can think of the test as a sample of that domain. Because the item covariances should reflect just shared domain variance, but item variance will be an unknown mixture of domain and specific and error variance, the amount of domain variance in a test would vary as the square of the number of items in the

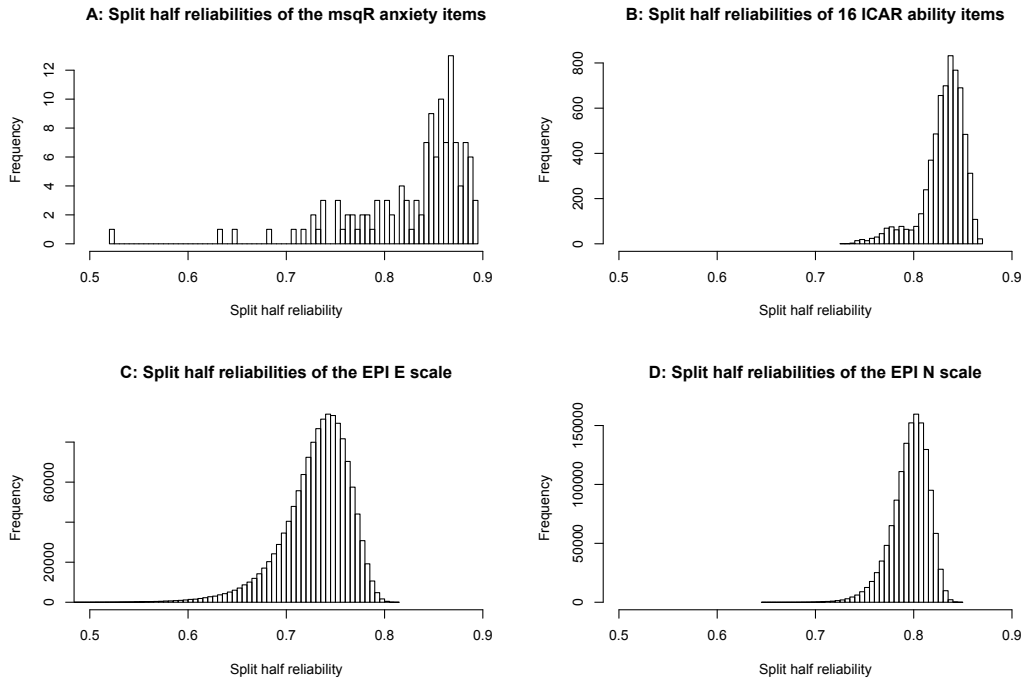

Figure 3. The distribution of 126 split half reliabilities for the 10 state anxiety items (panel A) and the 1,352,078 splits of the 24 EPI Extraversion items (panel C) suggests that the tests are not univocal while that of the 6,435 splits of the ICAR ability items (panel B) and the 1,352,078 splits of the EPI N scale (panel D) suggests greater homogeneity .

test times the average covariance of the items in the test. Considering items as measuring ability, Kuder & Richardson (1937) proposed several estimates of the reliability of the average split half, with their most well known being their 20th equation (and thus known as KR20).

A more general form of KR20 allows items to be not just right or wrong and thus corrects for the sum of the individual item variances. This is known as coefficient α (Cronbach, 1951) as well as λ_3 (Guttman, 1945)

$$\alpha = \lambda_3 = \frac{n}{n-1}\frac{V_t - \sum v_i}{V_t} \tag{13}$$

where V_t is the total test variance, v_i is the variance for a particular item, and there are n items in the test.

α and λ_3 may be thought of as the correlation of a test with a non-existent test just like it. That is, they are estimates of reliability based upon a fictitious parallel test. α estimates the correlation between the observed test and its hypothetical twin by assuming that the average covariance within the observed test is the same as the average covariance between items of the observed and with the hypothetical test. It is correct to the extent that the average inter-item correlation correctly estimates the amount of domain score variance (an unknown mixture of trait and state variance) in each item. But this is only correct if all the items have equal covariances and differ only in their observed variances. In this case they are said to be τ equivalent (Lord & Novick, 1968), which is a fancy way of saying that they all have equal covariances with the latent score represented by the test and have equal factor loadings on the single factor of the test. This is very unlikely in practice and deviations from this assumption will lead to α underestimating reliability

(Teo & Fan, 2013).

In addition to λ_3, Guttman (1945) considered five alternative ways of estimating reliability by correcting for the error variance of each item. All of these equations recognize that some of the item is reliable variance, the problem is how much? λ_3 and α assume that the average item covariance is a good estimate, λ_6 uses the Squared Multiple Correlation (smc) for each item as an estimate of its reliable variance while λ_2 uses a function of the average squared item covariance. λ_4 is just the maximum split half reliability.

One advantage of using the mean item covariance is that it can be identified from an analysis of variance perspective rather than actually finding all the inter-covariances. That is, just decompose the total test variance into three components: the between person variance σ_P^2, the between item variance, σ_I^2, and the interaction of person x item, σ_e^2. Then reliability is just $1 - \frac{\sigma_e^2}{\sigma_P^2}$ (Feldt, Woodruff, & Salih, 1987; Hoyt, 1941). By expressing it in this manner, Feldt et al. (1987) were able to derive an F distribution for α, and thus a means for finding confidence intervals. This is implemented as the `alpha.ci` function in the *psych* package. Alternative procedures for the confidence interval for α have been developed by Duhachek & Iacobucci (2004). Perhaps the biggest advantage to the variance approach to KR20, α, or λ_3 was that in the 1930s-1950s calculations were done with desk calculators rather than computers and it was far simpler to find the n item variances and one total test variance than it was to find the n*(n-1)/2 item covariances. In the modern era, such short cuts are no longer necessary.

Two problems with α. Although easy to calculate from just the item statistics and the total score, α and λ_3 are routinely criticized as poor estimates of reliability because they do not reflect the structure of the test (Bentler, 2009; Cronbach & Shavelson, 2004; S. Green & Yang, 2009; Revelle & Zinbarg, 2009; Sijtsma, 2009). Perhaps because the ability to find α is available in easy to use software packages, it is routinely used. This is unfortunate; except for very rare conditions, α is both an underestimate of the reliability of a test (because of the lack of τ equivalancy, Bentler, 2009),(Bentler, 2017; Sijtsma, 2009) and an overestimate of the fraction of test variance that is associated with the general variance in the test (Revelle, 1979; Revelle & Zinbarg, 2009; Zinbarg, Revelle, Yovel, & Li, 2005). As we show in the online supplement (Table 2), α provides no information about the constancy or stability of the test. For our mood items, α (.83 - .87) exceeded the short term constancy estimates (.42 - .76) and greatly exceeded the two day stability coefficients (.36 - .39). For the trait measures (particularly of impulsivity), the low α (.51) did not reflect the relatively high (.70) two-four week stability of the measures. That is to say, knowing α told us nothing about test-retest constancy or stability.

If not an estimate of reliability, does α measure internal consistency? No. For it is just a function of the number of items and the average correlation between the items. It is not a function of the uni-dimensionality of the test. It is easy to construct example tests with equal α values that reflect one test with homogenous items, two slightly related subtests or even two unrelated subtests each with homogeneous items (see, e.g., Revelle, 1979; Revelle & Wilt, 2013).

Model based estimates

That “internal consistency" estimates do not reflect the internal structure of the test becomes apparent when we apply “model based" techniques to examine the factor structure of the test. These procedures actually examine the correlations or covariances of the items in the test. Thanks to improvements in computational power, the task of finding correlations and the factor structure of a 10 item test has been transformed over the past two generations from being a summer research project for an advanced graduate student to an afternoon homework assignment for undergraduates. Using the latent variable modeling approach of factor analysis, these procedures decompose the test

variance into that which is common to all items (**g**, a general factor), that which is specific to some items (orthogonal group factors, **f**) and that which is unique to each item (typically confounding specific, **s**, and error variance, **e**). Many researchers have discussed this approach in great detail (e.g., Bentler, 2017; McDonald, 1999; Revelle & Zinbarg, 2009; Zinbarg et al., 2005) and we just summarize the main points here. Most importantly for applied researchers, as we show in the online supplement, model based techniques are just as easy to implement in modern software as are the more conventional approaches.

The observed score on a test item may be modeled in terms of the sum of the products of factor scores ($\mathbf{g}, \mathbf{f}, \mathbf{s}, \mathbf{e}$) and loadings ($\mathbf{c}, \mathbf{A}, \mathbf{D}$) on these factors:

$$\mathbf{x} = \mathbf{cg} + \mathbf{Af} + \mathbf{Ds} + \mathbf{e} \tag{14}$$

Because the reliable variance of the test is that which is not error, the reliability of a test with standardized items should be

$$\omega_t = \frac{\mathbf{1'cc'1} + \mathbf{1'AA'1}}{V_x} = 1 - \frac{\Sigma(1 - h_i^2)}{V_x} = 1 - \frac{\Sigma u_i^2}{V_x} \tag{15}$$

where h_i^2 is the item communality and u_i^2 is the item uniqueness. The percentage of the total variance that is due to the general factor (ω_g, McDonald, 1999) is

$$\omega_g = \frac{\mathbf{1'cc'1}}{V_X} = \frac{\mathbf{1'cc'1}}{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1} + \mathbf{1'ee'1}} = 1 - \frac{(\Sigma c_i)^2}{V_x}, \tag{16}$$

where the total test variance (V_x) is the sum of the elements of all the item variances and covariances and $(\Sigma c_i)^2$ is the squared sum of the loadings on the general factor.

Normally, the specific item variance is confounded with the residual item (error) variance, but if we have a way of estimating the specific variance by examining the correlations with items not in the test, (e.g., repeated items, Wood et al., 2017) then we can include it as part of the reliable variance (Bentler, 2017):

$$\omega_t = \frac{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1}}{V_X} = \frac{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1}}{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1} + \mathbf{1'ee'1}}. \tag{17}$$

Unfortunately, in his development of ω, McDonald (1999) refers to two formulae (6.20a and 6.20b) one for ω_t and one for ω_g and calls them both ω (Zinbarg et al., 2005). These two coefficients are very different, for one is an estimate of the total reliability of the test (ω_t), the second is an estimate of the amount of variance in the test due to single, general factor (ω_g). Then to make it even more complicated, there are two ways to find the general factor. One method uses a bifactor solution (Holzinger & Swineford, 1937; Reise, 2012; Rodriguez, Reise, & Haviland, 2016) using structural equation modeling software (e.g., *lavaan*, Rosseel, 2012), the other extracts a higher order factor from the correlation matrix of lower level factors and then applies a transformation developed by Schmid & Leiman (1957) to find the general loadings on the original items. The bifactor solution (ω_g) tends to produce slightly larger estimates than the Schmid-Leiman procedure (ω_h) because it forces all the cross loadings of the lower level factors to be 0. Following Zinbarg et al. (2005) we designate the Schmid-Leiman solution as ω_h recognizing the hierarchical nature of the solution. Both approaches are implemented in the *psych* package.

An important question when examining a hierarchical structure is how many group factors to specify when calculating ω_h? The Schmid-Leiman procedure is defined if there are three or more group factors, and with only two group factors the default is to assume that they are both equally

important (Zinbarg, Revelle, & Yovel, 2007). While the Schmid-Leiman approach is exploratory, the bifactor approach is a confirmatory model that requires specifying which variables load on each group factor.

How do these various approaches differ and what difference does it make? If we want to correct observed correlations for attenuation by using Equation 3 then underestimating reliability will lead to serious overestimation of the true validity of a measure. This is why there has been so much work on trying to estimate the greatest lower bound of reliability (e.g., Bentler, 2017). In this case if α underestimates reliability it is a poor measure to use when correcting for attenuation. In addition, many of the conventional measures do not reflect the percentage of total variance that is actually common to all of the items in the test. For factor analytic approaches, this is only done by ω_g and ω_h; for non-model based procedures this is the worst split half reliability (β).

In order to show how these various approaches can give very different values, we consider a real life data set consisting of the 10 anxiety items discussed in the online supplement. We show the correlation matrix as well as different reliability estimates in Table 2. Even though the greatest reliability estimates exceed .90, it is important to remember that this does not imply anything about the stability of the measure which is just .30 after two days. (See Table 1 in the online supplement.)

The ω_t based value of .88 agrees closely with the greatest split half of .89 or the duplicate item estimate of .92. These are all estimates of the total reliable variance. The worst split half .56 and ω_g values of .55 suggest that slightly less than 60% of the test reflects one general factor of anxiety. The difference between the .9 and the .6 values suggest that roughly 30% of the total test variance is due to the positively worded versus negatively worded group variance. That is, roughly 2/3 of the reliable test variance represents one construct, and about 1/3 represents something not shared with the total test. Note that the α of .83 does not provide as much information.

Tetrachoric, polychoric, and Pearson correlations

Test scores are typically the sum or average of a set of individual items. Each item is thought to reflect some underlying latent trait. Because the items are not continuous but rather are dichotomous or polytomous, the normal Pearson inter-item correlation will be attenuated from what would be observed if it were possible to correlate the latent scores associated with each item. The latent correlation can be estimated using tetrachoric or polychoric correlations which find what a (latent) continuous bivariate normal correlation would be given the observed pair-wise cell frequencies. The use of such correlations is recommended when examining the structure of a set of items using factor analysis for a clearer structure will appear and artificial difficulty factors will not be found. However, the temptation to use tetrachoric or polychoric correlations when finding the reliability of a test using any of the formulas in Table 2 (e.g., McNeish, 2017) should be resisted, for this will lead to overestimates of the amount of variance in the observed test made up of the observed items (Revelle & Condon, 2018).

Reliability and test length

With the exception of the worst split half reliability (β) and hierarchical ω (estimated either by a bi-factor approach, ω_g or the Schmid-Leiman procedure ω_h) all of the reliability estimates in Table 2 are functions of test length and will tend asymptotically towards 1 as the number of items increases. Examining the equations in Table 2 makes this clear: each method replaces the diagonal of the test, $tr(\mathbf{V_x})$, with the sum of some estimate based on the item reliability (r_{ii}, h^2, the SMC, $\sqrt{\bar{r}_{ij}^2}$, or $\bar{r}_{ij}$) and then compares this adjusted test variance to the total test variance. But as the number of items in the test increases, the effect of the diagonal elements

Table 2

Calculating multiple measures of internal consistency reliability demonstrated on 10 items from the Motivational State Questionnaire (`msqR` data set, N = 3032.) The ten items may be thought of as measures of state anxiety. Five are positively scored, five negatively. General factor loadings (g) and group factor loadings were found from the `omegaSem` function which applies a bi-factor solution. The hierarchical solution from `omega` applies the Schmid-Leiman transformation and has slightly lower general factor loadings. Split half calculations were done by finding all possible splits of the test. Although the statistics shown are done by hand, they are all done automatically in various psych functions (see Table 1).

10 anxiety items from the `msqR` data set

Variable	anxis	jttry	nervs	tense	upset	at.s-	calm-	cnfd-	cntn-	rlxd-	g	F1*	F2*	h2
anxious	1.00										0.41	0.60	0.00	0.53
jittery	0.47	1.00									0.47	0.41	0.00	0.39
nervous	0.54	0.48	1.00								0.43	0.60	0.00	0.55
tense	0.57	0.48	0.57	1.00							0.54	0.58	0.00	0.63
upset	0.29	0.16	0.35	0.45	1.00						0.34	0.32	0.00	0.22
at.ease-	0.23	0.24	0.29	0.35	0.30	1.00					0.66	0.00	0.47	0.66
calm-	0.28	0.32	0.31	0.36	0.23	0.62	1.00				0.69	0.00	0.31	0.57
confident-	-0.01	-0.02	0.08	0.07	0.17	0.44	0.31	1.00			0.11	0.00	0.77	0.61
content-	0.07	0.04	0.13	0.19	0.29	0.56	0.45	0.61	1.00		0.31	0.00	0.74	0.65
relaxed-	0.27	0.34	0.30	0.40	0.29	0.60	0.56	0.34	0.45	1.00	0.68	0.00	0.33	0.57
SMC	0.42	0.37	0.44	0.52	0.27	0.55	0.47	0.40	0.51	0.47				
r_{ii}	0.73	0.69	0.70	0.77	0.76	0.63	0.65	0.70	0.73	0.60				

	Formula	Calculation	Reliability measure
Total variance = $V_X = \Sigma(R_{ij}) = 39.60$			
Total reliable item variance = $\Sigma r_{ii} = 6.97$	$glb = \frac{V_x - tr(R) + \Sigma(r_{ii})}{V_x}$	$\frac{39.60-10+6.97}{39.60}$	= .923
r best split (A= 1, 2, 5, 6, 8 vs B = 3, 4, 7, 9, 10) = .81	$\lambda_4 =$ best split half $= \frac{2r_{ab}}{1+r_{ab}}$	$\frac{2*.81}{1+.81}$	= .895
Total common variance = $\Sigma h_i^2 = 5.37$	$\omega_t = \frac{V_x - tr(R) + \Sigma h_i^2}{V_x}$	$\frac{39.60-10+5.37}{39.60}$	= .883
Total squared multiple correlation $\Sigma(SMC) = 4.43$	$\lambda_6 = \frac{V_x - tr(R) + \Sigma(SMC)}{V_x}$	$\frac{39.60-10+4.43}{39.60}$	= .859
Average squared correlation= $\bar{r}^2 = \frac{\Sigma R_{ij}^2 - tr(R^2)}{n*(n-1)} = .137$	$\lambda_2 = \frac{V_x - tr(R) + \sqrt{\bar{r}^2 * n/(n-1)}}{V_x}$	$\frac{39.6-10+.\sqrt{.137*10/9}}{39.60}$	= .841
	$\alpha = \frac{n}{n-1}\frac{V_x - tr(R)}{V_x}$	$\frac{10}{9}\frac{39.60-10}{39.60}$	= .831
Average correlation = $\bar{r} = \frac{V_X - tr(V_X)}{n*(n-1)} = 0.329$	$\alpha = \frac{n\bar{r}}{1+(n-1)\bar{r}}$	$\frac{10*.329}{1+9*.329}$	= .831
r worst split (A = 1-5 vs. B= 6-10) = .385	$\beta =$ worst split half $= \frac{2r_{ab}}{1+r_{ab}}$	$\frac{2*.385}{1+.385}$	= .556
Sum of g loadings = 4.65 (bi-factor)	$\omega_g = \frac{(\Sigma g_i)^2}{V_X}$	$\frac{4.65^2}{39.60}$	= .545
Sum of g loadings = 4.09 (Schmid-Leiman)	$\omega_h = \frac{(\Sigma g_i)^2}{V_X}$	$\frac{4.09^2}{39.60}$	= .422

becomes less as a fraction of the total test variance. Thus, the limit of the glb, $\lambda_4, \omega_t, \lambda_6, \lambda_2, \alpha$ as n increases to infinity is 1. ω_h does not have this problem as it will increase towards the limit of $\omega_{g_\infty} = \frac{\mathbf{1'cc'1}}{V_X} = \frac{\mathbf{1'cc'1}}{\mathbf{1'cc'1+1'AA'1+1'DD'1}}$. When comparing reliabilities between tests of different lengths, it is useful to include the reliability of each test as if they were just one item each. In the case of α, $\alpha_1 = \bar{r}_{ij}$, Other single item reliability measures are the average item test retest ($glb_1 = \bar{r}_{ii}$), the average communality ($\omega_{t_1} = \bar{h}_i^2$), the average SMC ($\lambda_{6_1} = \overline{SMC_i}$), or the square root of the average squared correlation ($\lambda_{2_1} = \sqrt{\bar{r}_{ij}^2}$).

Generalizability Theory

Most discussions of reliability consider reliability as the correlation of a test with a test just like it. Test-retest and alternate form reliabilities are the most obvious examples. Internal consistency measures are functionally estimating the correlation of a test with an imaginary test just like it. These estimates are based upon the patterns of correlations of the items within the test. An alternative approach makes use of Analysis of Variance procedures to decompose the total test variance into that due to individuals, to items, to time, relevant interactions, and to residual (Cronbach et al., 1963; Gleser et al., 1965; Shavelson, Webb, & Rowley, 1989; Vispoel et al., 2018). We have already discussed this in the context of test-retest reliability. This technique is most frequently applied to the question of the reliability of judges who are making ratings of targets, but the logic can be applied equally easily to item analysis.

Reliability of raters

Consider the case where we are rating numerous subjects with only a few judges. We might do a small study first to determine how much our judges agree with each other, and depending upon this result, decide upon how many judges to use going forward. As an example, examine the data from 5 judges (raters) who are rating the anxiety of 10 subjects (Table 5 in the online supplement). If raters are expensive, we might want to use the ratings of just one judge rather than all five. In this case, we will want to know how ratings of any single judge will agree with those from the other judges. In this case, differences in leniency (the judges' means) between judges will make a difference in their judgements. In addition, different judges might use the scale differently, with some having more variance than others. We also need to think about how we will use the judges. Will we use their ratings as given, will we use their ratings as deviations from their mean, or will we pool the judges? All of these choices lead to different estimates of generalizability. Shrout & Fleiss (1979) provide a very clear exposition of three different cases and the resulting equations for reliability. (See Equations 17-19 in the supplement.) Although they express their treatment in terms of Mean Squares derived from an analysis of variance (e.g., the `aov` function in R), it is equally easy to do this with variance components estimated using a mixed effects linear model (e.g., `lmer` from the *lme4* package (Bates et al., 2015) in R). Both of these procedures are implemented in the `ICC` function in the *psych* package. This is discussed in more detail in the online supplement.

The intraclass correlation is appropriate when ratings are numerical, but sometimes ratings are categorical (particularly in clinical diagnosis or in evaluating themes in stories). This then leads to measures of agreement of nominal ratings. Rediscovered multiple times and given different names (Conger, 1980; Scott, 1955; Hubert, 1977; Zapf, Castell, Morawietz, & Karch, 2016) perhaps the most standard coefficient is known as Cohen's Kappa (Cohen, 1960, 1968) which adjusts observed proportions of agreement by the expected proportion:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{f_o - f_e}{N - f_e} \tag{18}$$

where $p_o = \frac{f_o}{N}$ is the observed proportion (p_o) or frequency of agreement (f_o) between two observers, and $p_e = \frac{f_e}{N}$ is the expected proportion or frequency of agreement (f_e) (Cohen, 1960) (See Table 7 in the supplement). Because raw agreements will reflect the base rates of judgements, κ corrects for the expected agreement on the assumption of independence of the raters. Thus, if two raters each use one category 60% of the time, we would expect them to agree by chance 36% of the time in their positive judgements and 16% in their negative judgements. Various estimates of correlations of nominal data have been proposed and differ primarily in the treatment of the correction for chance agreement (Feng, 2015). Thus, κ adjusts for differences in the marginal likelihood of judges, while Krippendorf's α_k does not (Krippendorff, 1970, 2004). To Krippendorff (2004) this is a strength of α_k, but to Fleiss it is not (Krippendorff & Fleiss, 1978).

If some disagreements are more important than others, we have weighted κ which with appropriate weights is equal to the intraclass correlation between the raters (Cohen, 1968; Fleiss & Cohen, 1973). For multiple raters, the average κ is known as Light's κ (Conger, 1980; Light, 1971).

Real life examples of a range of κ values are given by Freedman et al. (2013) in a discussion of the revised DSM where the κ values for clinical diagnoses range from "very good agreement" ($> .60$) for major neurocognitive disorders or post-traumatic stress disorder, to "good" (.40-.60) for bipolar II, or schizophrena, to "questionable agreement" (.2-4) for generalized anxiety or obsessive compulsive disorder, to values which did not exceed the confidence values of 0. When comparing the presence or absence of each of five narrative themes in a life story interview, Guo, Klevan, & McAdams (2016) report how two independent raters of each of 12 different interview segments showed high reliability of judgements with κ values ranging from .61 (did the story report early advantage) to .83 (did the story discuss prosocial goals).

Multilevel reliability

With the introduction of cell phones and various apps, it has become much easier to collect data within subjects over multiple occasions (e.g., Bolger & Laurenceau, 2013; A. S. Green, Rafaeli, Bolger, Shrout, & Reis, 2006; Mehl & Conner, 2012; Wilt et al., 2011, 2017). This has taken us from the daily diary to multiple mood measures taken multiple times per day. These techniques lead to fascinating data, in that we can examine patterns of stability and change within individuals over time. These intensive longitudinal methods (Walls & Schafer, 2006) "captures life as it is lived" (Bolger, Davis, & Rafaeli, 2003). They also lead to important questions about reliability. How consistent is one person over time? How stable are the differences between people over time? The same decomposition of variance techniques discussed for raters and for generalizability theory can be applied to an analysis of temporal patterns of reliability (Shrout & Lane, 2012; Revelle & Wilt, 2019). That is to say, we decompose the responses into variance components due to stable individual differences (σ_p^2), to differences due to time (σ_t^2), to the interaction of person by time effects (σ_{p*t}^2, and to residual error σ_e^2). Shrout & Lane (2012) give the SPSS and SAS syntax to do these calculations. In R this merely requires calling the `multilevel.reliability` function in *psych*. In the interest of space, we refer the interested reader to Shrout & Lane (2012), Revelle & Wilt (2019) and to the discussion, equations (11-16), and the examples in the online supplement.

When doing multilevel reliability, it is straightforward to find the reliability of each individual subject over items and over time. People are not the same and the overall indices do not reflect how some subjects show a very different pattern of response. The `multilevel.reliability` function returns reliability estimates for each subject over time, as well as the six estimates shown in the online supplement for 77 subjects on our ten anxiety items across four time points.

Composite Scores

The typical use of reliability coefficients is to estimate the reliability of relatively homogeneous tests. Indeed, the distinctions made between ω_h, α, and ω_t are minimized if the test is completely homogeneous. But if the test is intentionally made up of unrelated or partly unrelated content, then we need to consider the reliability of a composite score. A composite is sometimes referred to as a stratified test, where the strata may be difficulty or content based (Cronbach, Schönemann, & McKie, 1965). The stratified reliability (ρ_{xx_s}) of a composite test is found by replacing the variance of each subtest in the total test with its reliable variance and then dividing the resulting sum by the total test variance:

$$\rho_{xx_s} = \frac{V_t - \Sigma v_i + \Sigma p_{xx_i} v_i}{V_t} \tag{19}$$

where ρ_{xx_i} is reliability of the subtest and v_i is the variance of the subtest (Rae, 2007). Conceptually, this approach is very similar to ω_t (McDonald, 1999).

A procedure for weighting the elements of the composite to maximize the reliability of composite scores is discussed by Cliff & Caruso (1998) who suggest this as a procedure for Reliable Components Analysis (RCA) which they see as an alternative to a EFA or PCA.

Reliability of a difference score

Logically similar to the reliability of a composite is the reliability of a difference score (equation 20). Sometimes researchers want to find the difference between two scores (e.g., verbal and spatial ability or anxiety and depression). Even though the two tests themselves are highly reliable (ρ_{xx}, ρ_{yy}) , if they also have a high correlation, (r_{xy}) the reliability of the difference will be be substantially lower. Indeed, if the correlation between the two scales matches their reliability, the reliability of the difference will be 0. Given this reduction in reliability, individual differences in change or in pattern should be interpreted cautiously. We give an example of this problem when comparing the difference of two cognitive tests (i.e., verbal vs. spatial reasoning) in the online supplement.

$$\rho_{x-y} = \frac{\rho_{xx} + \rho_{yy} - 2r_{xy}}{2 * (1 - r_{xy})} \tag{20}$$

Beyond Classical Test Theory

Reliability is a joint property of the test and the people being measured by the test (refer back to Equation 2). For fixed amount of error, reliability is a function of the variance of the people being assessed. Scores from a test of ability will be reliable if given to a random sample of 18-20 year olds, but much less reliable if given to students at a particularly selective college because there will be less between person variance. The reliability of scores of emotional stability will be higher if given to a mixture of psychiatric patients and their spouses than it will be if given just to the patients. That is, reliability is not a property of test scores independent of the people taking it. This is the basic concept of Item Response Theory (IRT), called by some the "new psychometrics" (Embretson, 1996, 1999; Embretson & Reise, 2000) and which models the individual's patterns of response as a function of parameters (discrimination, difficulty) of the item.

By focusing on item difficulty (endorsement frequency) it is possible to consider the range of application of our scores. Items are most informative if they are equally likely to be passed or failed (endorsed or not endorsed). But this can only be the case for a particular person taking the test and can not be the case for a person with a higher or lower latent score. Although test scores are maximally reliable if all of the items are equally difficult, such scores will not be very discriminating

at any other than at that level (Loevinger, 1954). Thus, we need to focus on spreading out the items across the range to be measured.

The essential assumptions of IRT is that items can differ in how hard they are, as well as how well they measure the latent trait. Although seemingly quite different from classical approaches, there is a one-to-one mapping between the difficulty and discrimination parameters of IRT and the factor loadings and item response thresholds found by factor analysis of the polychoric correlations of the items (Kamata & Bauer, 2008; Markon, 2013; McDonald, 1999) . The relationship of the IRT approach to classical reliability theory is given a very clear explication by Markon (2013) who examines how test information (and thus the reliability) varies by subject variance as well as trait level. A test can be developed to be reliable for certain discriminations (e.g. between psychiatric patients) and less reliable for discriminating between members of a control group. The particular strength of IRT approaches is the use in tailored or adaptive testing where the focus is on the reliability for a particular person at a particular level of the latent trait. (See the discussion of IRT in the supplement, particularly Figure 3 which shows how reliability differs as a function of latent score.)

The several uses of reliability

Reliability is measured for at least three different purposes: correcting for attenuation, estimating expected scores, and providing confidence intervals around these estimates. When comparing test reliabilities, it is useful to remember that reliability has non-linear relations with the standard error as well as with the signal/noise ratio (Cronbach et al., 1965). That is, seemingly small differences in reliability between tests can reflect large differences in the ratio of reliable signal to unreliable noise or the size of the standard error of measurement. Consider the signal to noise ratio of tests with reliability of .7, .8., .9, and .95.

$$\frac{Signal}{Noise} = \frac{\rho_{xx}}{1-\rho_{xx}}.$$

Thus an improvement in reliability from .7 ($\frac{.7}{.3} = 2.33$) to .8 ($\frac{.8}{.2} = 4$) is a much smaller change in signal to noise than that from .8 to .9 ($\frac{.9}{.1} = 9$) which in turn is much less than from .9 to .95 ($\frac{.95}{.05} = 19$).

Corrections for attenuation

Reliability theory was originally developed to adjust observed correlations between related constructs for the error of the measurement in each construct (Spearman, 1904b). Such corrections for attenuation were perhaps the primary purpose behind reliability and are the reason that some recommend routinely correcting for reliability when doing meta analyses (Schmidt & Hunter, 1999). However such a correction is appropriate only if the measure is seen as the expected value of a single underlying construct. Examples of when the expected score of a test is not the same as the theoretical construct that accounts for the correlations between the observed variables include chicken sexing (Lord & Novick, 1968) or the diagnosis of Alzheimers (Borsboom & Mellenbergh, 2002). Modern software for Structural Equation Modeling (e.g., Rosseel, 2012) models the pattern of observed correlations in terms of a measurement (reliability) model as well as a structural (validity) model.

Reversion to mediocrity

Given a particular observed score, what do we expect that score to be if the measure is given again? That high scores decrease and low scores increase is just a function of the reliability of the

test (Equation 5) with larger drops and gains for extreme scores than for moderate scores. Although expected, these regression effects can mislead those who do not understand reliability and lead to surprise when successful baseball players are less successful the next year (Schall & Smith, 2000) or when poorly performing pilots improve but better performing pilots get worse (Kahneman & Tversky, 1973). That superior performance is partly due to good luck is hard for high performers to accept and that poor performance is partly due to bad luck leads to false beliefs about the lack of effect for rewards and the strong effect of punishment (Kahneman & Tversky, 1973).

Confidence intervals, expected scores, and the standard error

Not only does reliability affect the regression towards the mean, it also affects the precision of measurement. The standard error of measurement is a function of sample variability as well as the reliability (Equation 4). Confidence intervals for expected scores are symmetric around the expected score (Equation 5), and therefore are not symmetric around the observed score. Combining these two equations and taking into account the variance of the expected score we see that the confidence interval for an expected score based upon an observed score, X, with a sample variance of Vx, mean of $\bar{X}$ (and thus deviation score, x and estimated reliability of ρ_{xx} is

$$\rho_{xx}(x) - \sqrt{\rho_{xx} V_x (1 - \rho_{xx})} < \rho_{xx}(x) < \rho_{xx}(x) + \sqrt{\rho_{xx} V_x (1 - \rho_{xx})}. \tag{21}$$

Estimating and reporting reliability

We have included many equations and referred to many separate R functions. What follows is a brief summary with an accompanying flow chart that we presented earlier (Table 1).

Preliminary steps

The most important question to ask should be done before collecting the data: what are we trying to measure and how are we trying to measure it? Does the measure to be analyzed represent a single construct or is the factor structure more complicated? The next question is who are the subjects of interest? The reliability of test scores is not a property of the test, but a joint function of the people taking the test and of the test itself. Thus specifying the latent construct and the population of interest is essential before collecting and analyzing data.

Once one has decided what to measure, the test items must be given to willing (and one can hope interested) participants. Steps should be taken to ensure participant involvement. Measures to take include the classic issues of data screening. Subjects who respond too rapidly or carelessly will not provide reliable information (Wood et al., 2017). If response times are available, it is possible to screen for implausibly fast responses. If items are repeated in the same session, it is also possible to screen for temporal consistency (DeSimone, 2015; Wood et al., 2017).

Type of measurement and tests for unidimensionality.

Is the test given more than once? Is it given many times? Are the data based upon item responses or ratings? Are the data categorical, dichotomous, polytomous, or continuous? For the latter three, examining the structure of the correlations should be done to confirm the factor structure is as expected. A unidimensional scale would be expected to have just one large factor. More typical scales will probably have some sub-groups which can be explored using hierarchical or bifactor models. If the data are dichotomous or polytomous, using the latent variable correlations estimated by tetrachoric or polychoric correlations will show the structure more clearly. However, when estimating the reliability of the resulting scores, the statistics should be based upon the Pearson correlations.

Which reliability to estimate

As we have discussed before, there is no one reliability estimate. If giving just one test on one occasion we need to rely on internal consistency measures: ω_h, β and the worst split half reliability are estimates of the amount of general factor variance in the test scores. Simulations suggest that for very low levels of general factor saturation that the EFA based ω_h is positively biased and that a CFA based estimate (ω_g) is more accurate. ω_t is a model based estimate of the Greatest Lower Bound of the total reliability of a test as is the best split half reliability (λ_4). If the items are repeated within one form, the *glb* can be found based upon the item test-retest values.

If tests are given twice, then test-retest measures *dependability* over the short term or *stability* over a longer term. Variance decomposition techniques can be used to estimate how much variance is due to individuals, to the items, and to changes over time.

If tests are given many times, then multiple measures of reliability are relevant, each implying a different generalization: is time treated as fixed or random effect, are items seen as fixed or random. A powerful addition to this design is that reliability over time can be found for each subject as well as all of the subjects. The scores of some subjects may be much more reliable than others.

If the measures are not items, but rather raters, and we want to know the limits of generalizability of the raters to different raters, or for pooled raters, we can find estimates of the intra-class correlations. There are several of these, all can be estimated the same way.

These many forms of reliability coefficients (Table 1) may all be found in the *psych* package (Revelle, 2019a) for the open source statistics environment, R (R Core Team, 2019). *psych* was specially developed for personality oriented psychologists to be both thorough and easy to use. Although some of these statistics are available in commercial software packages, the *psych* package provides them all in one integrated set of functions. See Rstudio (RStudio Team, 2016) for a convenient interface. We show the specific commands to use to find all of these coefficients in the online supplement to this article.

Conclusions

Although we have used many equations to discuss it and many ways to estimate it, at its essence, reliability is a very simple concept: Reliability is a property of the test scores and is the correlation of a test with a test just like it, or alternatively, the fraction of the test score variance which is not due to error. Unfortunately, there is not just one reliability that needs to be reported, but rather a variety of coefficients, each of which is most appropriate for certain purposes. Are we trying to generalize over items, over time, over raters? Are we estimating unidimensionality, general factor saturation, or total reliable variance? Each of these questions leads to a different estimate (Table 1). So rather than ask what is the reliability, we should ask which reliability and reliability for what?

The initial appeal of α or KR20 reliability estimates were that they were simple to calculate in the pre-computer era. But this has not been the case for the past 60 years. The continued overuse of α is probably due to the ease of calculation in common commercial software. But with modern, open source software such as R, this is no longer necessary. α, ω_h, ω_t, minimum (β) and maximum (λ_4) split halfs, six ICCs, and six repeated measure reliabilities are all available with one or two simple commands. (See the online supplement for a guided tour.) It should no longer be acceptable to report one coefficient that is only correct if all items are exactly equally good measures of a construct. Readers are encouraged to report at least two coefficients (e.g., ω_h and ω_t) and then discuss why each is appropriate for the inference that is being made. They are discouraged from reporting just α unless they can justify the assumptions implicit in using it (i.e., τ equivalence

and unidimensionality). When reporting the reliability of raters, it is useful to report all six ICCs and then explain why one is most appropriate. Similarly, when reporting multilevel reliabilities, an awareness of what generalizations one wants to make is required before choosing between the six possible indices.

References

Anastasi, A. (1967). Psychology, psychologists and psychological testing. *American Psychologist*, *22*(4), 297-306.

Anderson, K. J., & Revelle, W. (1983). The interactive effects of caffeine, impulsivity and task demands on a visual search task. *Personality and Individual Differences*, *4*(2), 127-134. doi: 10.1016/0191-8869(83)90011-9

Anderson, K. J., & Revelle, W. (1994). Impulsivity and time of day: Is rate of change in arousal a function of impulsivity? *Journal of Personality and Social Psychology*, *67*(2), 334-344. doi: 10.1037/0022-3514.67.2.334

Baltes, P. B. (1987). Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology*, *23*(5), 611-626. doi: 10.1037/0012-1649.23.5.611

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1), 1-48*, *67*(1), 1-48. (R package version 1.1-8) doi: 10.18637/jss.v067.i01.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137-143. doi: 10.1007/s11336-008-9100-1

Bentler, P. M. (2017). Specificity-enhanced reliability coefficients. *Psychological Methods*, *22*(3), 527 - 540. doi: 10.1037/met0000092

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579-616. doi: 10.1146/annurev.psych.54.101601.145030

Bolger, N., & Laurenceau, J. (2013). *Intensive longitudinal methods*. New York, N.Y.: Guilford.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505-514. doi: 10.1016/S0160-2896(02)00082-X

Boyle, G. J., Stankov, L., & Cattell, R. B. (1995). Measurement and statistical models in the study of personality and intelligence. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (p. 417-446). Boston, MA: Springer US. doi: 10.1007/978-1-4757-5571-8_20

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

Cattell, R. B. (1946). Personality structure and measurement. I. The operational determination of trait unities. *British Journal of Psychology*, *36*, 88-102. doi: 10.1111/j.2044-8295.1946.tb01110.x

Cattell, R. B. (1964). Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology*, *55*(1), 1 - 22. doi: 10.1037/h0046462

Cattell, R. B. (1966a). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (p. 67-128). Chicago: Rand-McNally.

Cattell, R. B. (1966b). Patterns of change: Measurement in relation to state dimension, trait change, lability, and process concepts. *Handbook of multivariate experimental psychology*, 355–402.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, *24*(1), 3-30. doi: 10.1177/001316446402400101

Charter, R. A., & Feldt, L. S. (2001). Confidence intervals for true scores: is there a correct approach? *Journal of Psychoeducational Assessment*, *19*, 350-364. doi: 10.1177/073428290101900404

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*(1), 186 - 202. doi: 10.1037/a0015618

Cliff, N., & Caruso, J. C. (1998). Reliable component analysis through maximizing composite reliability. *Psychological Methods*, *3*(3), 291 - 308. doi: 10.1037/1082-989X.3.3.291

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(37-46). doi: 10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. doi: 10.1037/h0026256

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, *10*(1), 3–20. doi: 10.1037/1082-989X.10.1.3

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52-64. doi: 10.1016/j.intell.2014.01.004

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322 - 328. doi: 10.1037/0033-2909.88.2.322

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917-929. doi: 10.1177/0146167206287721

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *41*, 137-163. doi: 10.1111/j.2044-8317.1963.tb00206.x

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, *25*(2), 291-312. doi: 10.1177/001316446502500201

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391-418. doi: 10.1177/0013164404266386

Damian, R. I., Spengler, M., Sutu, A., & Roberts, B. W. (2018). Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *Journal of Personality and Social Psychology*. doi: 10.1037/pspp0000210

Deary, I. J., Whiteman, M., Starr, J., Whalley, L., & Fox, H. (2004). The impact of childhood intelligence on later life: Following up the Scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, *86*, 130–147. doi: 10.1037/0022-3514.86.1.130

DeSimone, J. A. (2015). New techniques for evaluating temporal consistency. *Organizational Research Methods*, *18*(1), 133-152. doi: 10.1177/1094428114553061

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ase): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*(5), 792-808.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341-349. doi: 10.1037/1040-3590.8.4.341

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433. doi: 10.1007/BF02294564

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates.

Eysenck, H. J., & Eysenck, S. B. G. (1964). *Eysenck Personality Inventory*. San Diego, California: Educational and Industrial Testing Service.

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

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93-103. doi: 10.1177/014662168701100107

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology*. doi: 10.1027/1614-2241/a000086

Fisher, A. J. (2015). Toward a dynamic model of psychological assessment: Implications for personalized care. *Journal of Consulting and Clinical Psychology*, *83*(4), 825 - 836. doi: 10.1037/ccp0000026

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613-619. doi: 10.1177/001316447303300309

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Sage.

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., ... Yager, J. (2013). The initial field trials of DSM-5: New blooms and old thorns. *American Journal of Psychiatry*, *170*(1), 1-5. doi: 10.1176/appi.ajp.2012.12091189

Glass, G. V. (1986). Testing old, testing new: schoolboy psychology and the allocation of intellectual resources. In B. S. Plake & J. C. Witt (Eds.), *The future of testing* (p. 9-27). Hillsdale. N.J.: Lawrence Erlbaum Associates.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, *30*(4), 395-418. doi: 10.1007/BF02289531

Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods*, *11*(1), 87-105. doi: 10.1037/1082-989X.11.1.87

Green, S., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121-135. doi: 10.1007/s11336-008-9098-4

Guo, J., Klevan, M., & McAdams, D. P. (2016). Personality traits, ego development, and the redemptive self. *Personality and Social Psychology Bulletin*, *42*(11), 1551-1563. doi: 10.1177/0146167216665093

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. doi: 10.1007/BF02288892

Hamaker, E. L., Schuurman, N. K., & Zijlmans, E. A. O. (2017). Using a few snapshots to distinguish mountains from waves: Weak factorial invariance in the context of trait-state research. *Multivariate Behavioral Research*, *52*(1), 47-60. doi: 10.1080/00273171.2016.1251299

Holzinger, K., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. doi: 10.1007/BF02287965

Hoyt, C. (1941, Jun). Test reliability estimated by analysis of variance. *Psychometrika*, *6*(3), 153–160. doi: 10.1007/BF02289270

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, *84*(2), 289 - 297. doi: 10.1037/0033-2909.84.2.289

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review*, *91*(2), 153-184. doi: 10.1037/0033-295X.91.2.153

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237-251. doi: 10.1037/h0034747

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136-153. doi: 10.1080/10705510701758406

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of consulting and clinical psychology*, *63*(1), 52-59. doi: /10.1037/0022-006X.63.1.52

Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: multiple systems, multiple functions. *Psychological review*, *109*(2), 306-329. doi: 10.1037/0033-295X.109.2.306

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, *2*, 139–150. doi: 10.2307/270787

Krippendorff, K. (2004, 7). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433. doi: 10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K., & Fleiss, J. L. (1978). Reliability of binary attribute data. *Biometrics*, *34*(1), 142–144.

Kuder, G., & Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160. doi: 10.1007/BF02288391

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Emotion* (p. 25-59). Thousand Oaks, CA: Sage Publications, Inc.

Leon, M. R., & Revelle, W. (1985). Effects of anxiety on analogical reasoning: A test of three theoretical models. *Journal of Personality and Social Psychology*, *49*(5), 1302-1315. doi: 10 .1037//0022-3514.49.5.1302

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365 - 377. doi: 10.1037/h0031643

Loe, B. S., & Rust, J. (2017). The perceptual maze test revisited: Evaluating the difficulty of automatically generated mazes. *Assessment*, *0*(0). doi: 10.1177/1073191117746501

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493 - 504. doi: 10.1037/h0058543

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325-336. doi: 10.1177/001316445501500401

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley Pub. Co.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, *27*, 251-280. doi: 10.1146/ annurev.ps.27.020176.001343

Markon, K. E. (2013). Information utility: Quantifying the total psychometric information provided by a measure. *Psychological Methods*, *18*(1), 15-35. doi: 10.1037/a0030638

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J.: L. Erlbaum Associates.

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*. doi: 10.1037/met0000144

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*(4), 289-374. doi: 10.1037/h0060985

Mehl, M. R., & Conner, T. S. (2012). *Handbook of research methods for studying daily life*. New York: Guilford Press.

Mehl, M. R., & Robbins, M. L. (2012). Naturalistic observation sampling: The electronically activated recorder (ear). In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life.* New York, NY: Guilford Press.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293-299. doi: 10.1037/1082-989X.1.3.293

Nesselroade, J. R., & Molenaar, P. C. M. (2016, May). Some behaviorial science measurement concerns and proposals. *Multivariate Behavioral Research*, *51*(2-3), 396–412. doi: 10.1080/ 00273171.2015.1050481

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rae, G. (2007). A note on using stratified alpha to estimate the composite reliability of a test composed of interrelated nonhomogeneous items. *Psychological Methods*, *12*(2), 177 - 184. doi: 10.1037/1082-989X.12.2.177

Rafaeli, E., & Revelle, W. (2006). A premature consensus: Are happiness and sadness truly opposite affects? *Motivation and Emotion*, *30*(1), 1-12. doi: 10.1007/s11031-006-9004-2

Rafaeli, E., Rogers, G. M., & Revelle, W. (2007). Affective synchrony: Individual differences in mixed emotions. *Personality and Social Psychology Bulletin*, *33*(7), 915-932. doi: 10.1177/0146167207301009

Rasch, G. (1966). An item analysis which takes individual differneces in to account. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 49–57. doi: 10.1111/j.2044-8317.1966.tb00354.x

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696. doi: 10.1080/00273171.2012.715555

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology*, *5*, 27–48.

Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57-74. doi: 10.1207/s15327906mbr1401_4

Revelle, W. (2019a, June). psych: Procedures for personality and psychological research [Computer software manual]. https://CRAN.r-project.org/package=psych. Retrieved from https://CRAN.R-project.org/package=psych (R package version 1.9.6)

Revelle, W. (2019b, June). psychtools: Tools to accompany the psych package for psychological research [Computer software manual]. https://CRAN.r-project.org/package=psychTools. Retrieved from https://CRAN.R-project.org/package=psychTools (R package version 1.9.6)

Revelle, W., & Anderson, K. J. (1998). *Personality, motivation and cognitive performance: Final report to the army research institute on contract MDA 903-93-K-0008* (Tech. Rep.). Evanston, Illinois, USA.: Northwestern University.

Revelle, W., & Condon, D. M. (2018). Reliability. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: a multidisciplinary reference on survey, scale and test development.* London: John Wily & Sons.

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2016). Web and phone based data collection using planned missing designs. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *Sage handbook of online research methods* (2nd ed., p. 578-595). Sage Publications, Inc.

Revelle, W., Humphreys, M. S., Simon, L., & Gilliland, K. (1980). Interactive effect of personality, time of day, and caffeine: A test of the arousal model. *Journal of Experimental Psychology General*, *109*(1), 1-31. doi: 10.1037/0096-3445.109.1.1

Revelle, W., & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of Research in Personality*, *47*(5), 493-504. doi: 10.1016/j.jrp.2013.04.012

Revelle, W., & Wilt, J. (2016). The data box and within subject analyses: A comment on Nesselroade and Molenaar. *Multivariate Behavioral Research*, *51*(2-3), 419-421. doi: 10.1080/00273171.2015.1086955

Revelle, W., Wilt, J., & Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory and executive control* (p. 27-49). New York, N.Y.: Springer.

Revelle, W., & Wilt, J. A. (2019). Analyzing dynamic data: a tutorial. *Personality and Individual Differences*, *136*(1), 38-51. doi: /10.1016/j.paid.2017.08.020

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika*, *74*(1), 145-154. doi: 10.1007/s11336-008-9102-z

Rocklin, T., & Revelle, W. (1981). The measurement of extraversion: A comparison of the Eysenck Personality Inventory and the Eysenck Personality Questionnaire. *British Journal of Social Psychology*, *20*(4), 279-284. doi: 10.1111/j.2044-8309.1981.tb00498.x

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological methods*, *21*(2), 137-150. doi: 10.1037/met0000045

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

RStudio Team. (2016). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA. Retrieved from http://www.rstudio.com/

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112 - 118. doi: 10.1037/0021-9010.85.1.112

Schall, T., & Smith, G. (2000). Do baseball players regress toward the mean? *The American Statistician*, *54*(4), 231-235. doi: 10.1080/00031305.2000.10474553

Schmid, J. J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 83-90. doi: 10.1007/BF02289209

Schmidt, F. L., & Hunter, J. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. , 1(2), 199-223. *Psychological Methods*, *1*(2), 199-223. doi: 10.1037/1082-989X.1.2.199

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, *27*(3), 183 - 198. doi: 10.1016/S0160-2896(99)00024-0

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, *19*(3), 321-325. doi: 10.1086/266577

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922 - 932. doi: 10.1037/0003-066X.44.6.922

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428. doi: 10.1037/0033-2909.86.2.420

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In *Handbook of research methods for studying daily life.* Guilford Press.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120. doi: 10.1007/s11336-008-9101-0

Spearman, C. (1904a). "General Intelligence," objectively determined and measured. *American Journal of Psychology*, *15*(2), 201-292. doi: 10.2307/1412107

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72-101. doi: 10.2307/1412159

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

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher*, *22*(2), 209-213. doi: 10.1007/s40299-013-0075 -z

Thayer, R. E. (1978). Toward a psychological theory of multidimensional activation (arousal). *Motivation and Emotion*, *2*(1), 1-34. doi: 10.1007/BF00992729

Thayer, R. E. (1989). *The biopsychology of mood and arousal.* The biopsychology of mood and arousal. xi, 234 pp. New York, NY: Oxford University Press.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1-26. doi: 10.1037/met0000107

Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data.* Oxford University Press.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Wilt, J., Bleidorn, W., & Revelle, W. (2016). Finding a life worth living: Meaning in life and graduation from college. *European Journal of Personality*, *30*, 158-167. doi: 10.1002/per.2046

Wilt, J., Bleidorn, W., & Revelle, W. (2017). Velocity explains the links between personality states and affect. *Journal of Research in Personality*, *69*(86-95). doi: 10.1016/j.jrp.2016.06.008

Wilt, J., Funkhouser, K., & Revelle, W. (2011). The dynamic relationships of affective synchrony to perceptions of situations. *Journal of Research in Personality*, *45*, 309–321. doi: 10.1016/ j.jrp.2011.03.005

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454-464. doi: 10.1177/1948550617703168

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16:93*. doi: 10.1186/s12874-016-0200-9

Zinbarg, R. E., Revelle, W., & Yovel, I. (2007). Estimating ω_h for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement*, *31*(2), 135-157. doi: 10.1177/0146621605278814

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133. doi: 10.1007/s11336-003-0974-7