Reliability from α to ω: A Tutorial

William Revelle and David Condon
Northwestern University

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 are straight forward and need to be understood by methodologists as well as practioners. 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. Easy to use, open source software is applied to examples of real data, and comparisons are made between the many types of reliability.

Contents

Reliability **2**
- Reliability as a variance ratio . . . 3
- Consistency, reliability and the data box . . . 7

Alternative estimates of an elusive concept **7**
- Test-retest of total scores . . . 7
- Components of variance estimated by test-retest measures . . . 13
- Alternate Forms . . . 15
- Split half (adjusted for test length) . . . 15
- Internal consistency and domain sampling . . . 16
- Model based estimates . . . 20
- Tetrachoric, polychoric, and Pearson correlations . . . 23

---

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 final version as submitted for review to *Psychological Assessment* on May 15, 2018.

Reliability and test length . . . 23
Generalizability Theory . . . 25
Reliability of raters . . . 25
Multilevel reliability . . . 29
Composite Scores . . . 32
Beyond Classical Test Theory . . . 33

The several uses of reliability **34**
Corrections for attenuation . . . 34
Reversion to mediocrity . . . 35
Confidence intervals, expected scores, and the standard error . . . 35

Estimating and reporting reliability **35**
Preliminary steps . . . 36
Type of measurement and tests for unidimensionality. . . . 36
Which reliability to estimate . . . 36

Conclusions **37**

Appendix **47**
First steps: installing *psych* and making it active . . . 47
Entering your data . . . 47
Specifying the items we want . . . 48
Consistency using the `testRetest` function . . . 48
Split reliability using the `splitHalf` function . . . 49
Internal consistency using the `alpha` and `omega` functions . . . 50
Parallel Forms . . . 53
Inter rater reliability using the `ICC` function . . . 54
Reliability over time: the `multilevelReliability` function . . . 55

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, 1951; McDonald, 1999) and it remains an active topic of research (Sijtsma, 2009; Revelle & Zinbarg, 2009; Bentler, 2009; McNeish, 2017; Wood et al., 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. 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 over 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.

The fundamental question in reliability is to what extent do scores measured at one time and one place with one instrument predict scores at another time and/or another 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 cognitive 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 the batting average of a baseball player. 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 Charles Spearman (1904b) for the first formalization of reliability. In the first of two landmark papers (the other, Spearman, 1904a, laid the basis for factor anal-

ysis 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:

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

The product of the observed score variance and the reliability is an estimate of the *latent construct* (sometimes called the "true score") variance in a test which we will symbolize 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 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 is

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

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 estimated latent score:

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

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

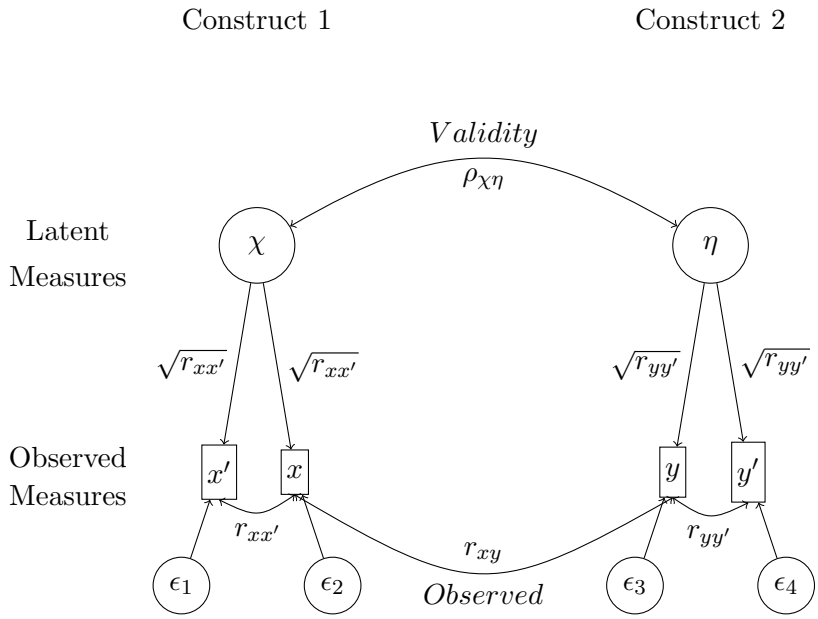

Figure 1. The basic concept of reliability and correcting for attenuation. Adjusting observed correlations (r_{xy}) by reliabilities ($r_{xx'}$, $r_{yy'}$) estimates underlying latent correlations ($\rho_{\chi\eta}$). (See Equation 3). Observed variables and correlations are shown in conventional Roman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

estimate the precision of our measurement. The problem for Spearman was, and remains for us today, how to find reliability?

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 et al., 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[1]: The reliability of a measure

[1]We can also find *within* subject reliability across time. This will be discussed later.

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 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,

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

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

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

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/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[2]. 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.

[2] 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 et al., 2005).

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 et al., 1963; Gleser et al., 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; 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 his *data box* it was a three way organization of measures taken over people, tests, and time. In typical Cattellian fashion, over the years this basic data relations matrix (BDRM or data box) 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).

> The generic concept of test consistency from which the above three parameters derive can be verbally defined as: the extent to which a test continues to measure the same psychological concept despite such changes as inevitably and normally occur in a test, its administration and the populations to which it is administered. ... Thus our estimate of the true reliability will be affected by a sampling of people and occasions; of the true homogeneity by sampling of items (or test elements) and people, and of the true transferability across populations by the sampling of people from various cultures and occasions. (Cattell, 1964, p 11)

These various types of reliability may be summarized graphically in terms of latent traits, paths, observed variables and correlations (Figure 2).

Alternative estimates of an elusive concept

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

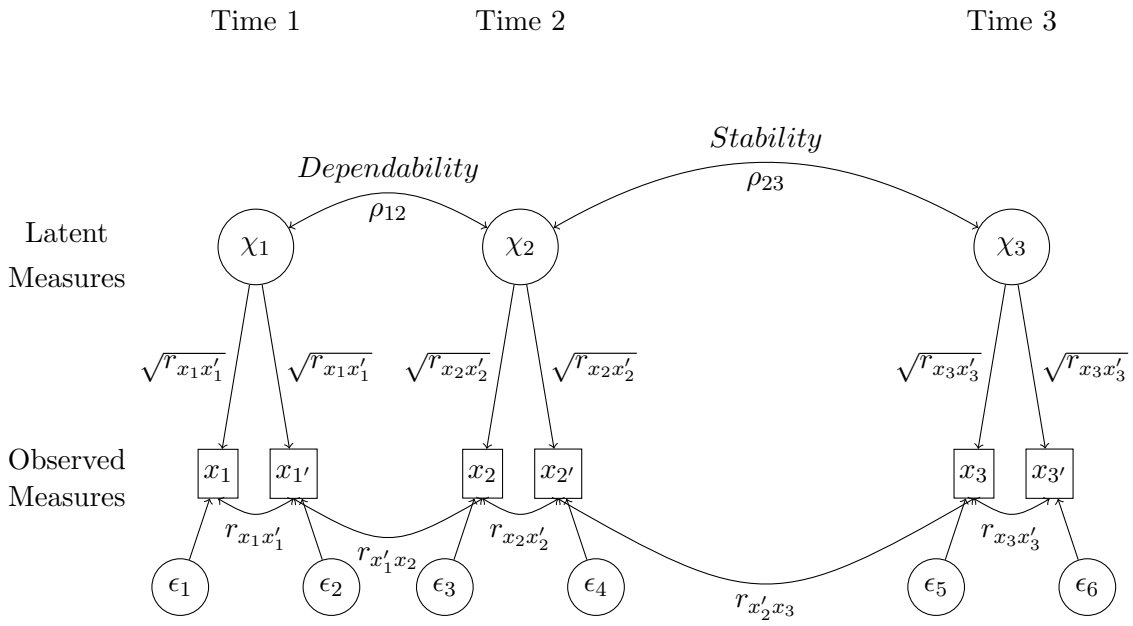

Figure 2. Reliability has many forms. Items within the same test provide estimates of *internal consistency*. Observed correlation between *alternate forms* or parallel tests given at the same time ($r_{x_1x_1'}$, $r_{x_2x_2'}$) estimate parallel test reliability. Tests given at (almost) the same time (times 1 and 2 e.g., $r_{x_1x_2}$, $r_{x_1'x_2'}$) 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.

variance and, additionally, a function of the length of time between the two measures:

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

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

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

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 et al., 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). 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).

A large test-retest correlation over a long period of time indicates temporal *stability* (Boyle et al., 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[3]. Measures of trait stability are a mixture of immediate test-retest *dependability* and longer term trait effects

[3]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.

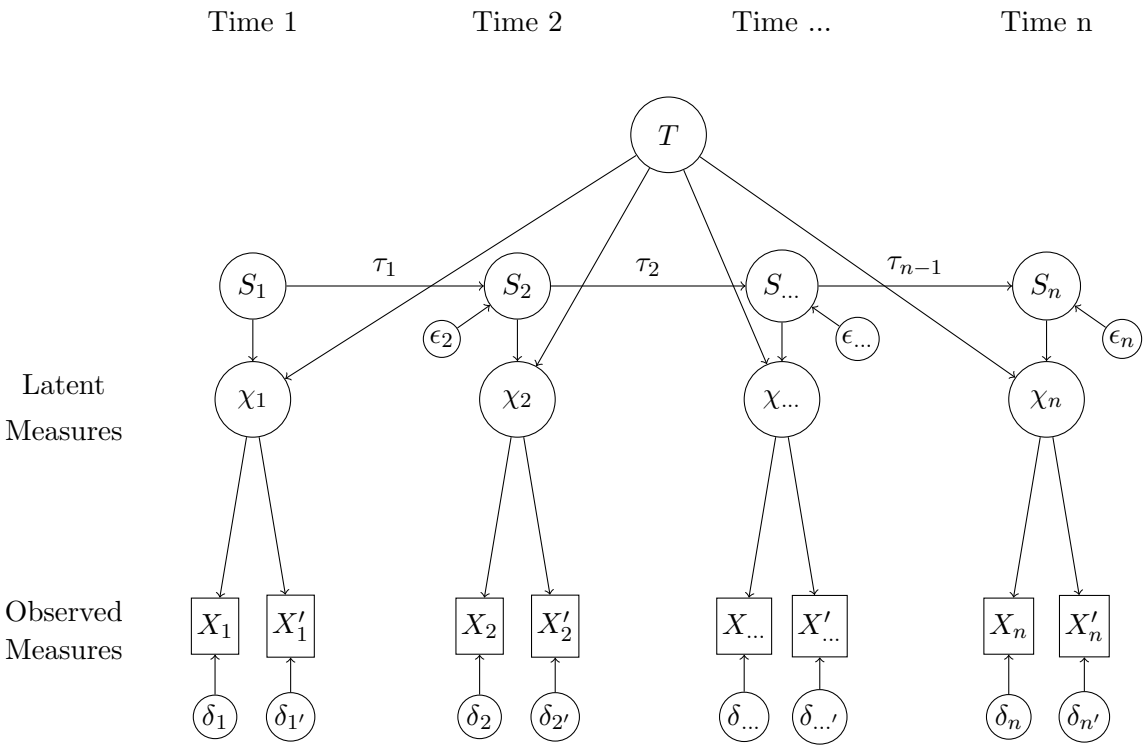

Figure 3. How traits and states affect short term and long term reliability. Observed scores (X_1 ... X_n) of the same trait measured over time reflect a latent Trait, time varying States (S_1 ... S_n) and measurement errors (δ_1 ... $\delta_{n'}$). The latent states are assumed to be auto-correlated (τ_i), such that over longer time periods the correlation will tend towards 0 ($\tau^n \rightarrow 0$) (see Equation 9). Adapted from (Cole et al., 2005).

(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 produces

> "estimates that are too high, because of material remembered on the second application of the test. This memory factor cannot be eliminated by increasing the length of time between the two applications, because of variable growth in the function tested within the population of individuals. These difficulties are

so serious that the method is rarely used" (Kuder & Richardson, 1937, p 151).[4]

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.

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 from the *psych* package (Revelle, 2018) in the `R` open statistical system (R Core Team, 2018); the analyses discussed below are demonstrated the Appendix. The data ($N > 4,000$) were collected as part of a long term series of studies of the interrelationships between the stable personality dimensions of extraversion, neuroticism, and impulsivity (e.g., Eysenck & Eysenck, 1964; Revelle et al., 1980), situational stressors (caffeine, time of day, movie induced affect, e.g., Anderson & Revelle, 1983, 1994; Rafaeli et al., 2007; Rafaeli & Revelle, 2006) momentary affective and motivational state (e.g. energetic and tense arousal (Thayer, 1978, 1989), state anxiety (Spielberger et al., 1970)), and cognitive performance (Humphreys & Revelle, 1984). The Motivational State Questionnaire (MSQ) included 75 items taken from a number of mood questionnaires (e.g., Thayer, 1978; Larsen & Diener, 1992; Watson et al., 1988) and had 10 anxiety items that overlapped with the state version of the State Trait Anxiety Inventory (Spielberger et al., 1970). The MSQ and STAI were given before any motivational manipulations were given, and then sometimes given at the end of the experimental session. We will use these 10 items in the subsequent examples evaluating and comparing the immediate dependability and the 45 minute and multi-day stability coefficients of these measures. These 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. An additional 20 items of the STAI were given with trait instructions and are reported as the `tai` dataset. The movie manipulation used minute film clips to induce fear (Halloween), depression (concentration camp), happiness (Parenthood), and a control movie (a nature film). The caffeine condition contrasted 4 mg/kg of caffeine to a placebo,

As is seen in Table 1, the immediate correlations for the 10 item state scales (test dependability) was .85. As expected, the 45 minute correlations were smaller (.42-.76) and those with one day delay were smaller yet (ranging from .36 to .39) with a mean of .38. This is in contrast to the immediate correlations of state with trait measures (.48) and after two days (.43) suggesting that the state measure has a trait component (or that the trait measure has a state component). The state retest measures were also much lower than the retest correlations of the EPI Impulsivity (.70) and Sociability (.81) subscales.

[4] In performance measures (e.g. cognitive ability assessment), this memory effect is described in terms of learning or practice effects.

Table 1: Comparing multiple estimates of reliability. SAI and MSQ contained 10 items measuring anxious vs. calm mood. Test-retest values include very short term (dependability) and longer term (stability) measures. Short term dependabilities are of mood measures pre and post various mood manipulations. One- to two-day delay stabilities are mood measures pre any mood manipulation. Dependability measures are based upon these 10 items given in the same session although using two different questionnaires. Trait anxiety was given with "how you normally feel", state anxiety asked "do you feel ". The two-four week delay compares personality trait measures (Impulsivity and Sociability) given as part of a group testing session and then part of various experimental sessions. Internal consistency estimates for α and ω do not require retesting. When giving the test twice, it is possible to find the consistency of each item (r_{ii}).The particular functions for estimating these coefficients are all part of the *psych* package.

	State		Trait			estimation
	SAI	MSQ	SAI	TAI	EPI		functions
Types of reliability			MSQ	(SAI)	Imp	Soc	
Test-retest							
Short term (test dependability)							
45 minutes (control)	.76	.74					`testRetest`
45 minutes (caffeine)	.73	.71					`testRetest`
45 minutes (films)	.42	.43					`testRetest`
Longer delay (stability)							
1-2 days	.36	.39		.43*			`testRetest`
1-4 weeks					.70	.81	
Average r_{ii} over time (item dependability)							
45 minutes (control)	.60	.57					`testRetest`
45 minutes (caffiene)	.61	.58					`testRetest`
45 minutes (film)	.39	.40					`testRetest`
1-2 days	.29	.30					`testRetest`
1-4 weeks					52	.56	`testRetest`
Parallel form approach							
Parallel tests	.74	.74		.48*			`scoreItems`
Duplicated tests (test dependability)			.85				`testRetest`
average r_{ii}(item dependability)			.67				`testRetest`
Internal consistency							
greatest split half (λ_4)	.91	.89		.94	.61	.83	`splitHalf`
ω_t	.90	.87		.92	.62	.80	`omega`
SMC adjusted (λ_6)	.89	.86		.92	.52	.78	`splitHalf`
α (λ_3)	.87	.83		.90	.51	.76	`alpha`
average split half	.86	.83		.90	.50	.76	`splitHalf`
ω_g	.56	.45		.67	.31	.62	`omega`
smallest split half	.66	.57		.79	.41	.66	`splitHalf`
worst split half (β)	.66	.57		.50	.05	.27	`iclust`
average r	.39	.33		.32	.11	.19	`alpha`
Other forms of reliability							
ICC	.87	.83		.90	.51	.76	`ICC`
kappa							

*Trait Anxiety x State Anxiety

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[5] 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, 2017; 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 is

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

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 et al., 2015) for R. As an example of such a variance decomposition consider the 10 overlapping items in the STAI and MSQ discussed earlier (Table 2). 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. From these components of variance, we can find several different reliability estimates (Cranford et al., 2006; Shrout & Lane, 2012). The first is the reliability of the total score for each individual across the 10 overlapping items if the test is thought of as composed of those (fixed) 10 items.

$$R_{1F} = \frac{\sigma^2_{Person} + \frac{\sigma^2_{Person\ x\ Item}}{k}}{\sigma^2_{Person} + \frac{\sigma^2_{Person\ x\ Item}}{k} + \frac{\sigma^2_{Residual}}{k}} \tag{12}$$

The second is the reliability of the average of the two measurement occasions[6] this is

$$R_{kF} = \frac{\sigma^2_{Person} + \frac{\sigma^2_{Person\ x\ Item}}{k}}{\sigma^2_{Person} + \frac{\sigma^2_{Person\ x\ Item}}{k} + \frac{\sigma^2_{Residual}}{2k}}. \tag{13}$$

[5] 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).

[6] Functionally, just the Spearman-Brown prophecy formula applied to R_{1F}. This will be discussed later in terms of split half reliability.

Additional estimates can be found for the reliability of a single item (R_{1R}) or the average of an item across both time points (R_{2R}) (Shrout & Lane, 2012).

Table 2: A variance decomposition of the 10 overlapping items from the STAI and MSQ measures (N= 200). Data taken from the example for the `testRetest` function in the *psych* package. Four different variance ratios and reliabilities may be found from these (Cranford et al., 2006; Shrout & Lane, 2012).

Multilevel components of variance due to		
	Variance	Percent
Subjects	0.34	0.19
Time	0.44	0.25
Items	0.17	0.10
Subjects x time	0.24	0.13
Subjects x items	0.01	0.01
time x items	0.34	0.19
Residual	0.23	0.13
Total	1.78	1.00

This leads to four reliability estimates:

R_{1F}	= 0.94	Reliability of average of all items for one time (Random time effects)
R_{kF}	= 0.97	Reliability of average of all ratings across all items and times (Fixed time effects)
R_{1R}	= 0.33	Generalizability of a single time point across all items (Random time effects)
R_{2R}	= 0.49	Generalizability of average time points across all items (Fixed time effects)

Multi-level modeling approaches are particularly appropriate if repeating the same measure multiple times in for instance an experience sampling study (e.g., Bolger & Laurenceau, 2013; Fisher, 2015; Mehl & Conner, 2012; Mehl & Robbins, 2012; Wilt et al., 2011, 2016a,b). 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., 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, 2017). 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 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 et al., 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 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:

$$r_{x_1x_2} = \frac{\sigma_T^2 + \sigma_S^2}{V_X} = \frac{\sigma_T^2 + \sigma_S^2}{\sigma_T^2 + \sigma_S^2 + \sigma_s^2 + \sigma_e^2}$$

If given with a long enough delay, then the state effect will tend towards zero and the alternate form correlation will be an estimate of the percentage of trait variance in each of the two test forms.

$$r_{x_1x_2} = \frac{\sigma_T^2}{V_X} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_S^2 + \sigma_s^2 + \sigma_e^2}. \tag{14}$$

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 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`. Note how these correlation are less than when we actually repeat the same items by correlating the overlapping items of the `sai` and `msqR` (.85).

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):

$$r_{xx} = \frac{2 * r_{x_1x_2}}{1 + r_{x_1x_2}}. \tag{15}$$

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 15)

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 4 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) where the greatest split half reliability was .87, the average is .83, and the lowest is .73 (Figure 4 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 4 part C). This supports the criticism that the EPI E scale tends to measure 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 .84, a mean of .8, and a minimum of .68, providing strong evidence for a relatively homogeneous scale (Figure 4 part D)

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 et al., 2018) and allow for the many reliability estimates discussed before.

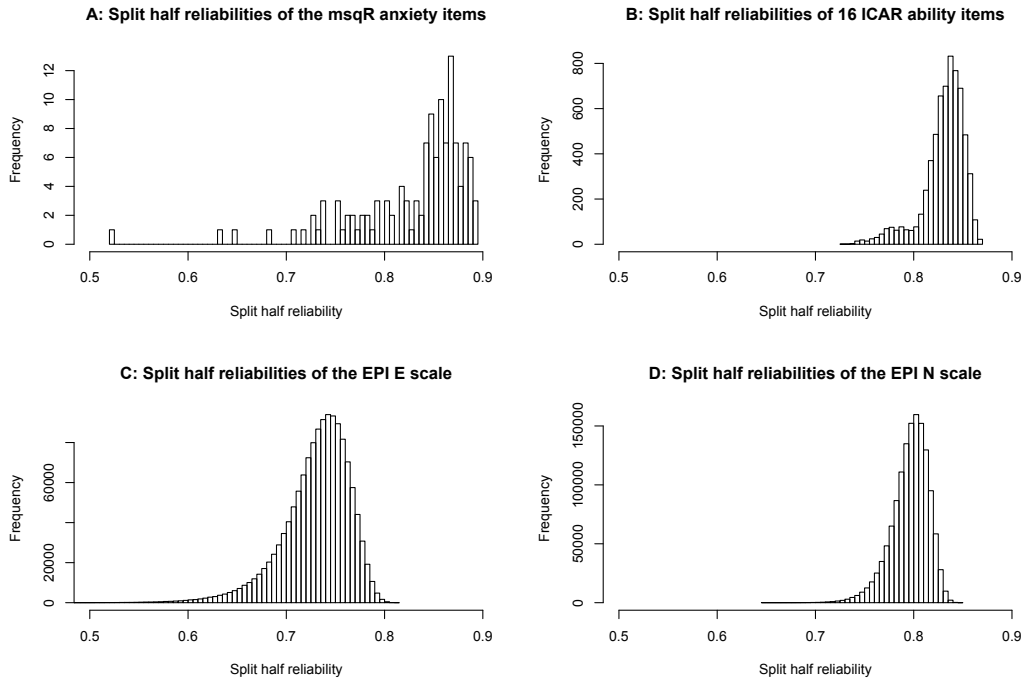

Figure 4. The distribution of 126 split half reliabilities for the 10 state anxiety items (panel A) and the 1,352,078 splits of the 24 EPI Extraversion items (panel C) suggests that the tests are not univocal while that of the 6,435 splits of the ICAR ability items (panel B) and the 1,352,078 splits of the EPI N scale (panel D) suggests greater homogeneity .

In order to understand these procedures, it is useful to think about what goes into the correlation between two tests or two times. Consider two tests, $\mathbf{X}$ and $\mathbf{X'}$, both made up of two subtests. The simple correlation $r_{xx'} = \frac{C_{xx'}}{\sqrt{V_x V_{x'}}}$ may be expressed in terms of elements (items) in X and X'. The reliability of $\mathbf{X}$ is just its correlation with $\mathbf{X'}$ and can be thought of in terms of the elements of the variance-covariance matrix, $\Sigma_{XX'}$:

$$\Sigma_{XX'} = \begin{pmatrix} \mathbf{V_x} & \vdots & \mathbf{C_{xx'}} \\ \cdots & \cdots & \cdots \\ \mathbf{C_{xx'}} & \vdots & \mathbf{V_{x'}} \end{pmatrix} \tag{16}$$

and letting $V_{\mathbf{x}} = \mathbf{1'V_x1}$ and $C_{\mathbf{XX'}} = \mathbf{1'C}_{XX'}\mathbf{1}$ where $\mathbf{1}$ is a column vector of 1s and $\mathbf{1'}$ is its transpose, the correlation between the two tests will be

$$\rho_{xx'} = \frac{C_{xx'}}{\sqrt{V_x V_{x'}}}.$$

But the variance of a test is simply the sum of the true covariances and the error variances and we can break up each test ($\mathbf{X}$ and $\mathbf{X}'$) into their individual items ($\mathbf{x_1}$ $\mathbf{x_2}$ ... $\mathbf{x_i}$) and their respective variances and covariances. We can split each test into two parts and then the structure of the two tests seen in Equation 16 becomes

$$\Sigma_{XX'} = \left(\begin{array}{ccc|ccc} \mathbf{V_{x_1}} & \vdots & \mathbf{C_{x_1x_2}} & \mathbf{C_{x_1x'_1}} & \vdots & \mathbf{C_{x_1x'_2}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{C_{x_1x_2}} & \vdots & \mathbf{V_{x_2}} & \mathbf{C_{x_2x'_1}} & \vdots & \mathbf{C_{x_2x'_2}} \\ \hline \mathbf{C_{x_1x'_1}} & \vdots & \mathbf{C_{x_2x'_1}} & \mathbf{V_{x'_1}} & \vdots & \mathbf{C_{x'_1x'_2}} \\ \mathbf{C_{x_1x'_2}} & \vdots & \mathbf{C_{x_2x'_2}} & \mathbf{C_{x'_1x'_2}} & \vdots & \mathbf{V_{x'_2}} \end{array}\right) \tag{17}$$

But what if we don't have two tests? We need to make assumptions about the structure of what the covariance between a test ($\mathbf{X}$) and a test just like it ($\mathbf{X}'$) would be based upon what we know about test $\mathbf{X}$.

Because the splits are done at random and the second test is parallel with the first test, the expected covariances between splits are all equal to the true score variance of one split ($\mathbf{V_{t_1}}$), and the variance of a split is the sum of true score and error variances:

$$\Sigma_{XX'} = \left(\begin{array}{ccc|ccc} \mathbf{V_{t_1}} + \mathbf{V_{e_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} + \mathbf{V_{e_1}} & \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} \\ \hline \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t'_1}} + \mathbf{V_{e'_1}} & \vdots & \mathbf{V_{t'_1}} \\ \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t'_1}} & \vdots & \mathbf{V_{t'_1}} + \mathbf{V_{e'_1}} \end{array}\right)$$

The correlation between a test made up of two halves with intercorrelation ($r_1 = V_{t_1}/V_{x_1}$) with another such test is

$$r_{xx'} = \frac{4V_{t_1}}{\sqrt{(4V_{t_1} + 2V_{e_1})(4V_{t_1} + 2V_{e_1})}} = \frac{4V_{t_1}}{2V_{t_1} + 2V_{x_1}} = \frac{4r_1}{2r_1 + 2}$$

and thus

$$r_{xx'} = \frac{2r_1}{1 + r_1}. \tag{18}$$

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):

$$\frac{2 * r}{1 + r}. \tag{19}$$

Unfortunately, as we showed in Figure 4 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 test times the average covariance of the items in the test. Considering items as measuring ability (e.g., right with probability p and wrong with probability $q = 1 - p$), 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):

$$r_{tt} = \frac{n}{n-1}\frac{\sigma_t^2 - n\overline{pq}}{\sigma_t^2}. \tag{20}$$

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)

$$\alpha = \lambda_3 = \frac{n}{n-1}\frac{V_t - \sum v_i}{V_t} \tag{21}$$

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 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, 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. 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 et al., 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, 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 et al., 2005, 2006). As we saw in Table 1, 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, 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 and loadings on these factors:

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

But the covariances of items reflect the factors common to all or some of the items (**g** and **f**) and the variance/covariance matrix (**C**) of a test made up n items is just the sum of the product of the **c** vector and its transpose and the Group vectors (**A**) and their transpose. The total test variance (V_x) is just the sum of the elements of this matrix

$$V_X = \mathbf{1'C1} = \mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1} + \mathbf{1'ee'1}$$

where **1** is a vector of 1s and with matrix multiplication allows us to find the sum of the elements. The communality of an item is the amount of its variance modeled by the common factors and is

$$h^2 = c_i^2 + \Sigma A_{ij}^2.$$

The unique variance for each item is its total variance less the common variance:

$$u_i^2 = \sigma_i^2(1 - h_i^2).$$

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

$$\omega_t = \frac{\mathbf{1'cc'1} + \mathbf{1'AA'1}}{V_x} = 1 - \frac{\Sigma(1 - h_i^2)}{V_x} = 1 - \frac{\Sigma u_i^2}{V_x}. \tag{22}$$

The percentage of the total variance that is due to the general factor (ω_g, McDonald, 1999) is

$$\omega_g = \frac{\mathbf{1'cc'1}}{V_X} = \frac{\mathbf{1'cc'1}}{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1} + \mathbf{1'ee'1}}. \tag{23}$$

That is, the sum of the squared loadings on the **g** factor divided by the sum of the correlations or covariances of all of the items. In summation notation this is just

$$\omega_g = \frac{\Sigma_{i=1}^{n} g_i^2}{\Sigma_{i=1}^{n}\Sigma_{j=1}^{n} R_{ij}}.$$

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):

$$\omega_t = \frac{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1}}{V_X} = \frac{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1}}{\mathbf{1'cc'1} + \mathbf{1'AA'1} + \mathbf{1'DD'1} + \mathbf{1'ee'1}}. \tag{24}$$

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 et al., 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 bi-factor 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 et al., 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 because 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.

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 earlier. We show the correlation matrix as well as different reliability estimates in Table 3. 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 (Table 1).

The ω_t based value of .88 agrees closely with the greatest split half of .91 or the duplicate item estimate of .92. These are all estimates of the total reliable variance. The worst split half .57 and ω_g values of .59 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 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 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. The temptation to use tetrachoric or polychoric correlations when finding the reliability of a test using any of the formulas in Table 3 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 3 are functions of test length and will tend asymptotically towards 1 as the number of items increases. Examining the equations in Table 3 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, 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 becomes less as a fraction of the total test variance. Thus, the limit of the glb, $\lambda_4, \omega_t, \lambda_6, \alpha$ 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

Table 3: 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* fuctions (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	0.46	0.49	0.52	0.27	0.22	0.24	-0.01	0.08	0.24	0.35		0.59	0.47
jittery	0.46	1.00	0.46	0.47	0.16	0.25	0.31	-0.01	0.04	0.33	0.43		0.43	0.37
nervous	0.49	0.46	1.00	0.56	0.37	0.28	0.32	0.11	0.13	0.32	0.44		0.58	0.53
tense	0.52	0.47	0.56	1.00	0.48	0.36	0.38	0.11	0.19	0.42	0.55		0.57	0.63
upset	0.27	0.16	0.37	0.48	1.00	0.32	0.26	0.19	0.27	0.32	0.39		0.31	0.25
at.ease-	0.22	0.25	0.28	0.36	0.32	1.00	0.63	0.45	0.55	0.60	0.68	0.43		0.65
calm-	0.24	0.31	0.32	0.38	0.26	0.63	1.00	0.35	0.44	0.58	0.72	0.27		0.59
confident-	-0.01	-0.01	0.11	0.11	0.19	0.45	0.35	1.00	0.60	0.38	0.22	0.71		0.55
content-	0.08	0.04	0.13	0.19	0.27	0.55	0.44	0.60	1.00	0.45	0.34	0.73		0.65
relaxed-	0.24	0.33	0.32	0.42	0.32	0.60	0.58	0.38	0.45	1.00	0.71	0.29		0.59
SMC	0.37	0.36	0.42	0.52	0.29	0.55	0.48	0.40	0.48	0.49				
r_{ii}	0.73	0.67	0.66	0.74	0.73	0.61	0.62	0.67	0.72	0.57				

	Formula	Calculation	Reliability measure
Total variance = $V_X = \Sigma(R_{ij}) = 39.80$			
Total reliable item variance = $\Sigma r_{ii} = 6.71$	$glb = \frac{V_x - tr(R) + \Sigma(r_{ii})}{V_x}$	$\frac{39.80-10+6.71}{38.90}$	= .917
r best split (A= 1, 4, 6, 7, 10 vs B = 2,3,5, 8, 9) = .834	λ_4 = best split half = $\frac{2r_{ab}}{1+r_{ab}}$	$\frac{2*834}{1+.834}$	= .909
Total common variance = $\Sigma h_i^2 = 5.27$	$\omega_t = \frac{V_x - tr(R) + \Sigma h_i^2}{V_x}$	$\frac{39.80-10+5.27}{39.80}$	= .881
Total squared multiple correlations $\Sigma(SMC) = 4.36$	$\lambda_6 = \frac{V_x - tr(R) + \Sigma(SMC)}{V_x}$	$\frac{39.80-10+4.36}{39.80}$	= .858
	$\alpha = \frac{n}{n-1}\frac{V_x - tr(R)}{V_x}$	$\frac{10}{9}\frac{39.80-10}{39.80}$	= .832
Average correlation = $\frac{V_X - tr(V_X)}{n*(n-1)} = 0.331$	$\alpha = \frac{n\bar{r}}{1+(n-1)\bar{r}}$	$\frac{10*.331}{1+9*.331}$	= .832
r worst split (A = 1-5 vs. B= 6-10) = .397	β = worst split half = $\frac{2r_{ab}}{1+r_{ab}}$	$\frac{2*.397}{1+.397}$	= .569
Sum of g loadings = 4.84 (bi-factor)	$\omega_g = \frac{(\Sigma g_i)^2}{V_X}$	$\frac{4.84^2}{39.80}$	= .589
Sum of g loadings = 4.21 (Schmid-Leiman)	$\omega_h = \frac{(\Sigma g_i)^2}{V_X}$	$\frac{4.21^2}{39.80}$	= .446

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 ($\omega_{t_1} = \bar{h}_i^2$), or the average SMC ($\lambda_{6_1} = \overline{SMC_i}$).

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 et al., 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 4). 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. 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.

In Case 1 each subject is rated by k randomly chosen judges. The variance of the ratings is thus a mixture of between person and between judges. We can estimate these variance components from a one way analysis of variance treating subjects as random effects. Within person variance is an unknown mixture of rater and and residual (which includes error and the interaction) effects. Reliability for a single judge (ICC_1) is the ratio

of person variance to total variance, while reliability for multiple judges (ICC_{1k}) adjusts the residual variance (σ_w^2) by the number of judges

$$ICC_1 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_w^2} \qquad ICC_{1k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_w^2}{k}}. \tag{25}$$

Case 2 is more typical, in that we are still using the ratings of k randomly chosen judges, but each judge rates all subjects. We are trying to generalize to another set of randomly chosen judges. This is a two way random effects model where both subjects and raters are chosen at random. By partitioning out the raters effects (σ_r^2) from the residual, we improve our estimate for the person variance (σ_p^2). Once again, by having multiple raters, the residual term ($\sigma_r^2 + \sigma_e^2$) is reduced by the number of raters ($\frac{\sigma_r^2+\sigma_e^2}{k}$):

$$ICC_2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_r^2 + \sigma_e^2} \qquad ICC_{2k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_r^2+\sigma_e^2}{k}}. \tag{26}$$

Case 3 is unusual when considering judges, but is typical when considering items. It assumes judges are fixed rather than random effects. Thus, this is a two-way mixed model (subjects are random, judges are fixed). The estimate of the person variance is the same as in Case 2, but by assuming judges are fixed, the variance associated with judges is removed from the divisor of our reliability coefficient:

$$ICC_3 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2} \qquad ICC_{3k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{k}}. \tag{27}$$

Each of these cases can be examined for the reliability of a single judge, or for k judges. The effect of pooling judges is identical to the effect of pooling items and is just the Spearman-Brown correction. When applied to items, the ICC_{3k} is the same as α because we typically associate item differences as fixed effects. The `ICC` function reports 6 different relliability estimates: three for the case of single judges, three for the case of multiple judges. It also reports the results in terms of a traditional analysis of variance aa well as a mixed effects linear model as well as confidence intervals for each coefficient (Table 5).

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 et al., 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:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{f_o - f_e}{N - f_e} \tag{28}$$

Table 4: An example of rater reliability. Five judges (raters) rate 10 subjects on a trait. The subjects differ in their overall mean score across judges, the judges differ in their mean ratings. These data produce six different estimates of reliability (Table 5).

	Judge or Rater					
Subject	J1	J2	J3	J4	J5	Mean
S1	1	1	3	2	3	2.0
S2	1	1	3	2	5	2.4
S3	1	2	3	2	4	2.4
S4	4	1	2	6	2	3.0
S5	2	4	3	5	3	3.4
S6	3	4	3	4	6	4.0
S7	1	4	6	4	6	4.2
S8	3	4	4	4	6	4.2
S9	3	5	4	6	5	4.6
S10	5	5	5	6	5	5.2
Mean	2	2.4	2.4	3	3.4	4.0

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). 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).

Consider a study where four coders are asked to rate 10 narratives for three mutually exclusive categories: Achievement, Intimacy, and Power. A hypothetical example of such data is shown in Table 6. Raters 1 and 2 and 3 and 4 show high agreement, but there is no agreement between raters 2 and 4.

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"

Table 5: Intra class correlations summarize the amount of variance due to subjects, raters, and their interactions. Depending upon the type of generalization to be made, one of six different reliability coefficients is most appropriate. Scores may be analyzed as a one way (Case 1) or two way (Cases 2 and 3) ANOVAs with random (Cases 1 and 2) or mixed effects (Case 3). Variance components may be derived from the MS from ANOVA or directly from the `ICC` output: For Case 1, $\sigma_p^2 = \frac{MS_p - MS_w}{k}$. Similarly, for Cases 2 and 3 , $\sigma_p^2 = \frac{MS_p - MS_{residual}}{k}$. ICCs may be based upon 1 rater or k raters. Data from Table 4 are analyzed using the `ICC` function from the *psych* package.

Analysis of Variance and the resulting decomposition into variance components

	df	SS	MS	Variance	Value by case			% of total		
					C1	C2	C3	C1	C2	C3
Person	9	51.22	5.69	σ_p^2	0.76	0.87	0.87	29	32	40
Within	40	75.20	1.88	σ_w^2	1.88			71		
Rater	4	27.32	6.83	σ_r^2		0.55			20	
Residual (P x R)	36	47.88	1.33	σ_e^2		1.33	1.33		48	60
Total				σ_t^2	2.64	2.75	2.20	100	100	100

Intraclass correlations and their confidence intervals (From the `ICC` function).

Variable	type	ICC	F	df1	df2	p	lower bound	upper bound
Single_raters_absolute	ICC1	0.29	3.03	9	40	0.01	0.04	0.66
Single_random_raters	ICC2	0.32	4.28	9	36	0.00	0.09	0.67
Single_fixed_raters	ICC3	0.40	4.28	9	36	0.00	0.13	0.74
Average_raters_absolute	ICC1k	0.67	3.03	9	40	0.01	0.19	0.91
Average_random_raters	ICC2k	0.70	4.28	9	36	0.00	0.32	0.91
Average_fixed_raters	ICC3k	0.77	4.28	9	36	0.00	0.42	0.93

Number of subjects = 10 Number of raters = 5

Table 6: Cohen's κ can be used to assess the change corrected agreement between raters for categorical data. κ adjusts observed agreement by expected agreement. It is found using the `cohen.kappa` function.

Hypothethetical ratings from four raters for 10 subjects on three strivings.

Subject	R1	R2	R3	R4
1	Achieve	Achieve	Achieve	Power
2	Achieve	Achieve	Intimacy	Power
3	Achieve	Achieve	Intimacy	Power
4	Achieve	Achieve	Power	Power
5	Achieve	Intimacy	Achieve	Achieve
6	Intimacy	Achieve	Achieve	Achieve
7	Intimacy	Intimacy	Intimacy	Intimacy
8	Intimacy	Power	Intimacy	Intimacy
9	Power	Power	Intimacy	Intimacy
10	Power	Power	Power	Power

Produces this measure of % agreement

Rater	R1	R2	R3	R4
R1	100	70	50	40
R2	70	100	40	30
R3	50	40	100	70
R4	40	30	70	100

Which, when adjusted for chance becomes
Kappa by each pair of raters
(Unweighted below the diagonal, weighted above)

Rater	R1	R2	R3	R4
R1	1.00	0.78	0.30	-0.14
R2	0.52	1.00	0.29	-0.17
R3	0.24	0.13	1.00	0.52
R4	0.15	-0.01	0.57	1.00

Average Cohen kappa for all raters	0.27
Average weighted kappa for all raters	0.26

(.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 et al. (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 et al., 2006; Mehl & Conner, 2012; Wilt et al., 2011, 2016b). 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 et al., 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, 2017). 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*.

Table 7: An abbreviated data set (adapted from Shrout & Lane (2012). Four subjects give responses to three items over four time points. See also Figure 5.

Variable	Person	Time	Item1	Item2	Item3
1	1	1	3	4	4
2	2	1	6	6	5
3	3	1	3	4	3
4	4	1	7	8	7
5	1	2	5	7	7
6	2	2	6	7	8
7	3	2	3	5	9
8	4	2	8	8	9
9	1	3	4	6	7
10	2	3	7	8	9
11	3	3	5	6	7
12	4	3	6	7	8
13	1	4	5	9	7
14	2	4	8	9	9
15	3	4	8	7	9
16	4	4	6	8	6

Shrout & Lane (2012) discuss six reliability (or generalizability) coefficients that may be found from these variance components. Although the next six equations are probably challenging, it might be a relief to realize that the resulting estimates are all reported in the `multilevel.reliablity` function.

If we are interested in how stable the between person differences are when averaged over all the (m) items and all (k) occasions, then we need to compare the variance due to people and to people by items to those variances plus error variance:

$$R_{kF} = \frac{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m}}{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m} + \frac{\sigma_e^2}{km}}. \tag{29}$$

But, if the interest is individual differences from one randomly chosen time point (R), then

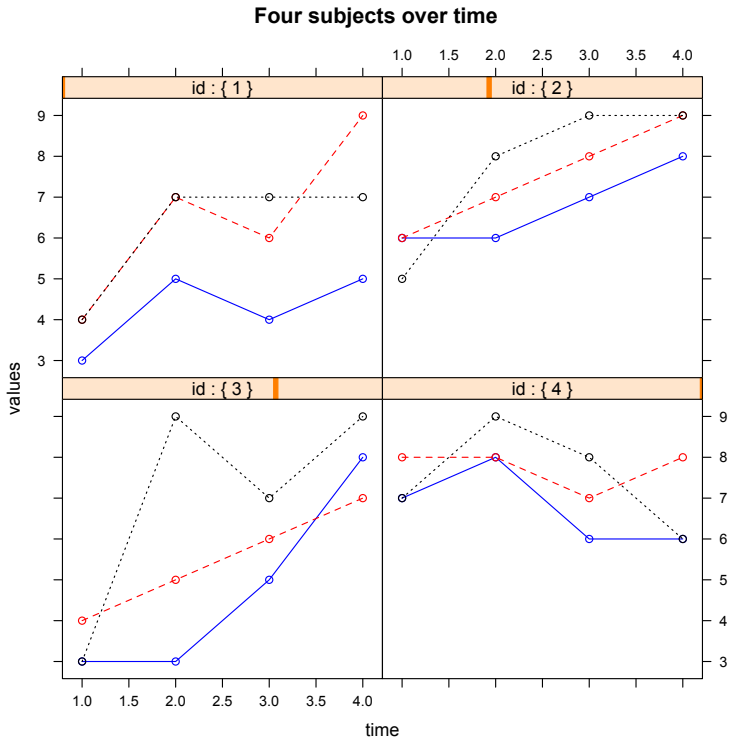

Figure 5. Four subjects are measured over four time points on three variables. The data are shown in Table 7 and adapted from Shrout & Lane (2012). Figure drawn using the `mlPlot` function in *psych*.

we need to add time and its interaction with person in the denominator and we find

$$R_{1R} = \frac{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m}}{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m} + \sigma_t^2 + \sigma_{p*t}^2 + \frac{\sigma_e^2}{m}}. \tag{30}$$

An extension of the reliability coefficient from one randomly chosen time point (R_{1R}) to the average of k times (R_{kR}) is analogous to the benefit of Spearman-Brown formula and is

$$R_{kR} = \frac{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m}}{\sigma_p^2 + \frac{\sigma_{p*i}^2}{m} + \frac{\sigma_t^2 + \sigma_{p*t}^2}{k} + \frac{\sigma_e^2}{km}}. \tag{31}$$

To measure the reliability of within individual change, we do not need to consider between

person variability, just variability within people over time:

$$R_C = \frac{\sigma^2_{p*t}}{\sigma^2_{p*t} + \frac{\sigma^2_e}{m}}. \tag{32}$$

The four equations above assume that time points are fixed and that all subjects are measured at the same time points (e.g., perhaps every evening, or at fixed times through out the day). But if the timing differs across subjects we need to think of time as nested within subjects and we derive two more reliabilities, that between subjects and that within subjects:

$$R_{kRN} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{t(p)}}{k} + \frac{\sigma^2_e}{km}}. \tag{33}$$

$$R_{CN} = \frac{\sigma^2_{t(p)}}{\sigma^2_{t(p)} + \frac{\sigma^2_e}{m}}. \tag{34}$$

The previous six equations might seem daunting, but are included to show the logic of generalizability theory as applied to the problems associated with measuring individual differences in mood over time. All six are included as the output for the `multilevel.reliabiity` function, see Table 8.

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 Table 8. In the Appendix, we show multi-level reliabilities 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 such a composite score. Such a composite is sometimes referred to as a stratified test, where the strata may be difficulty or content based (Cronbach et al., 1965). The stratified reliability (ρ_{xx_s}) of such a 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:

$$\rho_{xx_s} = \frac{V_t - \Sigma v_i + \Sigma p_{xx_i} v_i}{V_t} \tag{35}$$

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).

Table 8: Alternative estimates of reliability based upon Generalizability theory for the example data set. Analysis done by the `multilevel.reliability` function.

RkF	= 0.92	Reliability of average of all ratings across all items and times (Fixed time effects)
R1R	= 0.25	Generalizability of a single time point across all items (Random time effects)
RkR	= 0.57	Generalizability of average time points across all items (Random time effects)
Rc	= 0.79	Generalizability of change (fixed time points, fixed items)
RkRn	= 0.44	Generalizability of between person differences averaged over time (time nested within people)
Rcn	= 0.73	Generalizability of within person variations averaged over items (time nested within people)

The data had 4 observations taken over 4 time intervals for 3 items.

Source		Variance	Percentage
Person	σ_p^2	0.57	0.15
Time	σ_t^2	0.82	0.21
Items	σ_i^2	0.48	0.12
Person x Time	σ_{pt}^2	0.84	0.22
Person x Items	σ_{pi}^2	0.12	0.03
Time x items	σ_{ti}^2	0.31	0.08
Residual	σ_e^2	0.68	0.18
Total	σ_T^2	3.82	1.00
	Nested model		
Person	σ_p^2	0.38	0.11
Time(person)	$\sigma_{p(t)}^2$	1.46	0.43
Residual	σ_e^2	1.58	0.46
Total	σ_T^2	3.43	1.00

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.

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. 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 a test 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 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. Classical test theory has been likened to a "flogging wall" where we count the number of whips hitting subjects as

they move down a conveyer belt as a measure of height rather than calibrating items to the targets (Lumsden, 1976).

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 tests are maximally reliable if all of the items are equally difficult, such a test 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 aproaches, 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; 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.

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 much smaller 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 observed scores are symmetric around the expected score (Equation 5), but therefore are not symmetric around the observed score. Combining these two equations we see that the confidence interval for an observed score, X, with a sample variance of Vx, mean of $\bar{X}$ and estimated reliability of ρ_{xx} is

$$\rho_{xx}(X-\bar{X}) - \sqrt{V_x(1-\rho_{xx})} + \bar{X} < \rho_{xx}(X-\bar{X}) < \rho_{xx}(X-\bar{X}) + \sqrt{V_x(1-\rho_{xx})} + \bar{X}$$

which is probably easier to understand in terms of deviation scores ($x = X - \bar{X}$):

$$\rho_{xx}(x) - \sqrt{V_x(1-\rho_{xx})} < \rho_{xx}(x) < \rho_{xx}(x) + \sqrt{V_x(1-\rho_{xx})}. \tag{36}$$

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 (Table 9).

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? Reliability is not a function of a 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 responses that 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.

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 a test. 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. Some subjects may be much more reliable than others.

If the measures are nor 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 9) may all be found in the *psych* package (Revelle, 2018) for the open source statistics environment, R (R Core Team, 2018). *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. We show the specific commands to use to find all of these coefficients in the appendix 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 the correlation of a test with a test just like it, or alternatively, the fraction of a test 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 9). 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 and maximum split halfs, six ICCs, and six repeated measure reliabilities are all available with one or two simple commands. (See the appendix 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 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.

Table 9: Steps toward reliability analysis: choosing the appropriate function to find reliability. All functions except for the `cfa` 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		`omega`
	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.reliabiity`
	variance components	$\sigma_p^2, \sigma_i^2, \sigma_t^2$	`multilevel.reliabiity`
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`

References

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* (pp. 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

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).

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

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.

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., 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.

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. (PMID: 23288382) doi: 10.1176/appi.ajp.2012.12091189

Gleser, G., Cronbach, L., & 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. (PMID: 27880048) 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 01). 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. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/10705510701758406 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.

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), 1073191117746501. (PMID: 29239208) 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.

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.

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. (2018). 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. (PMID: 24049214) 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. (2018, April). psych: Procedures for personality and psychological research [Computer software manual]. https://cran.r-project.org/web/packages=psych. Retrieved from https://CRAN.R-project.org/package=psych (R package version 1.8.4)

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. (2017). Analyzing dynamic data: a tutorial. *Personality and Individual Differences*. 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.

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. 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. Retrieved from http://www.jstor.org/stable/2746450

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, May 01). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher*, *22*(2), 209–213. Retrieved from https://doi.org/10.1007/s40299-013-0075-z 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. (2016a). 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. (2016b). Velocity explains the links between personality states and affect. *Journal of Research in Personality*, *xx*, xx-xx doi: 10.1016/j.jrp.2016.06.008. doi: http://dx.doi.org/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

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement*, *30*(2), 121-144. doi: 10.1177/0146621605278814

Appendix

Here we include R code (R Core Team, 2018) to find the various reliability estimates discussed above. All of these examples require installing the *psych* (Revelle, 2018) package and then making it active. To install R on your computer, go to https://cran.r-project.org and install the most recent version that is appropriate for your computer (PC, MacOS, Linux). For details on installing and using R, go to the http://personality-project.org/r. Many people find that RStudio (RStudio Team, 2016) is a very convenient interface to R. It may be downloaded from https://www.rstudio.com.

First steps: installing psych and making it active

R code

```
install.packages("psych",dependencies = TRUE)  #just need to do this once
library(psych) #make the psych package active-- need to do this everytime you start R
```

Detailed instructions for *psych* may be found by reading the accompanying *vignettes* or reading the series of "HowTo"s from the personality-project.org. As is true of all R packages, help for individual functions may be obtained by entering ? followed by the command you do not understand. All functions in R operate upon objects and then return the result as another object. This is the real power of R for it allows us to do a particular analysis and then do a subsequent analyiss on the results. Most functions in R have default values for certain options. These can be changed by specifying the option by name and giving the desired value. To find the complete list of options for any functions, you can ask for help for that function. RStudio will prompt with the available options when you type the name of the function.

Entering your data

For the examples below, we will use datasets already available in R. However, it is important to know how to enter your own data into R as well. The easiest way of doing this is to read from an external file where the first row of the file gives the names of the

variables and the subsequent rows are one row for each subject. If you have a data file that is a text file with the suffix .text, or .txt, or .csv, or that has been saved from e.g., SPSS as a .sav file, then you can read the data using the `read.file` command. This will open a search window on your computer and you can locate the file. Alternatively, you can copy the data to your clipboard and use the `read.clipboard` command.

By default, `read.file` and `read.clipboard` assume that the first line of the file includes "header" information. That is, the names of the variables. If this is not true, then specify that header=FALSE. Examples of the code one might use to enter your data are given below. All of the 'R code' chunks in this file can be copied and pasted directly into the R console.

R code

```
my.data <- read.file() #opens a search window and reads the file
#or first copy your data to the clipboard and then
my.data <- read.clipboard()
```

Specifying the items we want

For these examples we use smaller subsets of the larger `msqR` data set and then specify which items to score for which analysis.

R code

```
?msqR   #ask for information about the sai data set
table(msqR$study,msqR$time)   #show the study names and  sample sizes
#Now, select some subsets for analysis using the subset function.
msq1 <- subset(msqR,msqR$time == 1)  #just the first day measures
sai1 <- subset(sai,sai$time==1)  #just the first set of observations
rim <- subset(sai,sai$study=="RIM")  #choose the RIM study for test retest over days
vale <- subset(sai,sai$study=="VALE") #choose the VALE study for multilevel analysis

#create keying information for several analyses
sai.alternate.forms <- list( pos1 =c( "at.ease","calm","confident","content","relaxed"),
  neg1 = c("anxious", "jittery", "nervous" ,"tense"  ,   "upset"),
  anx1 = c("anxious", "jittery", "nervous" ,"tense", "upset","-at.ease" ,  "-calm" ,
       "-confident", "-content","-relaxed"),
  pos2=c(  "secure","rested","comfortable" ,"joyful" , "pleasant" ),
  neg2=c("regretful","worrying", "high.strung","worried", "rattled" ),
  anx2 = c("regretful","worrying", "high.strung","worried", "rattled", "-secure",
    "-rested", "-comfortable", "-joyful",  "-pleasant" ))
 anx.keys <- sai.alternate.forms$anx1   #the keys to use for scoring

 select <- selectFromKeys (anx.keys)  #to be used later in alpha
```

Consistency using the `testRetest` function

To run the `testRetest` function, the data need to be in one of two forms: two data objects with an equal number of rows or one object where the subjects are identified with

an identification number and the time (1 or 2) of testing is specified. Here we show the second way of doing this. We use the `sai` example data file included in *psych* package and then extract just a small subset (the RIM data set measured State Anxiety on two different days).

R code

```
rim.test.retest <-  testRetest(rim,keys=anx.keys)  #do the analysis
rim.test.retest    #show the results
```

This results in the following output

```
Test Retest reliability
Call: testRetest(t1 = rim, keys = anx.keys)

Number of subjects =  342  Number of items =  10
 Correlation of scale scores over time 0.33      <---  this is the test-retest correlation
 Alpha reliability statistics for time 1 and time 2
       raw G3 std G3   G6 av.r  S/N   se lower upper var.r
Time 1   0.86   0.86 0.89 0.39 6.35 0.04  0.75  0.92  0.03
Time 2   0.88   0.88 0.91 0.43 7.62 0.03  0.80  0.92  0.03

 Mean between person, across item reliability =  0.26
 Mean within person, across item reliability =  0.45
with standard deviation of  0.39

 Mean within person, across item d2 =  1.09
R1F  =  0.79 Reliability of average of all items for one  time (Random time effects)
RkF  =  0.88 Reliability of average of all items and both times (Fixed time effects)
R1R  =  0.51 Generalizability of a single time point across all items (Random time effects)
Rc   =  0.72 Generalizability of change (fixed time points, fixed items)
Multilevel components of variance
             variance Percent
ID               0.11    0.10
Time             0.00    0.00
Items            0.20    0.19
ID x time        0.09    0.09
ID x items       0.20    0.19
time x items     0.11    0.10
Residual         0.35    0.33
Total            1.05    1.00

 To see the item.stats, print with short=FALSE.
To see the subject reliabilities and differences, examine the 'scores' object.
```

Split reliability using the `splitHalf` function

To find split half reliabilities and to graph the distributions of split halves (e.g., Figure 4) requires three lines. Here we use the built in `ability` data set of 16 items for 1,525 participants taken from the Synthetic Aperture Personality Assessment (SAPA) project ((http://sapa-project.org) (Revelle et al., 2010, 2016) and reported in (Condon & Revelle, 2014).

R code

```
sp <- splitHalf(ability,raw=TRUE, brute=TRUE)
sp #show the results
hist(sp$raw,breaks=101, xlab="Split half reliability",
      main="Split half reliabilities of 16 ICAR ability items")
```

```
Split half reliabilities
Call: splitHalf(r = ability, raw = TRUE, brute = TRUE)

Maximum split half reliability (lambda 4) =  0.87
Guttman lambda 6                          =  0.84
Average split half reliability            =  0.83
Guttman lambda 3 (alpha)                  =  0.83
Minimum split half reliability  (beta)    =  0.73
                                             2.5% 50% 97.5%
 Quantiles of split half reliability      =  0.77 0.83 0.86
```

Internal consistency using the `alpha` *and* `omega` *functions*

Although we do not recommend α as a measure of consistency, many researchers want to report it. The `alpha` function will do that. Confidence intervals from normal theory (Duhachek & Iacobucci, 2004) as well as from the bootstrap are reported. We use 10 items from the anxiety inventory as an example. We use all the cases from the `msqR` data set. By default, items that are negatively correlated with the total score are *not* reversed. However, if we specify that `check.keys=TRUE`, then items with negative correlations with the total score are automatically reversed keys. A warning is produced.

R code

```
alpha(msq1[select],check.keys=TRUE)
```

```
Reliability analysis
Call: alpha(x = msq1[select], check.keys = TRUE)

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean   sd
      0.83      0.83    0.86      0.33   5 0.0046    2 0.54

 lower alpha upper     95\% confidence boundaries
0.82 0.83 0.84

 Reliability if an item is dropped:
          raw_alpha std.alpha G6(smc) average_r S/N alpha se    NA
anxious-       0.83      0.83    0.85      0.34 4.7   0.0047 0.026
jittery-       0.83      0.83    0.85      0.35 4.8   0.0047 0.027
nervous-       0.82      0.82    0.84      0.33 4.4   0.0049 0.029
tense-         0.81      0.81    0.83      0.32 4.2   0.0051 0.029
upset-         0.82      0.82    0.85      0.34 4.7   0.0049 0.033
at.ease        0.80      0.80    0.83      0.31 4.1   0.0055 0.028
```

```
calm           0.80      0.81    0.84      0.32 4.2   0.0054 0.030
confident      0.83      0.83    0.85      0.36 5.0   0.0046 0.022
content        0.82      0.82    0.84      0.34 4.6   0.0049 0.025
relaxed        0.80      0.81    0.84      0.31 4.1   0.0055 0.030

 Item statistics
             n raw.r std.r r.cor r.drop mean   sd
anxious-  1871  0.54  0.56  0.51   0.42  2.3 0.86
jittery-  3026  0.52  0.55  0.48   0.41  2.3 0.83
nervous-  3017  0.59  0.64  0.60   0.52  2.6 0.68
tense-    3017  0.67  0.71  0.69   0.60  2.4 0.78
upset-    3019  0.54  0.58  0.50   0.45  2.6 0.68
at.ease   3018  0.77  0.74  0.72   0.67  1.6 0.94
calm      3020  0.74  0.71  0.68   0.63  1.6 0.92
confident 3021  0.54  0.50  0.43   0.38  1.5 0.93
content   3010  0.64  0.59  0.55   0.50  1.4 0.92
relaxed   3023  0.76  0.73  0.70   0.66  1.6 0.91

Non missing response frequency for each item
              0    1    2    3 miss
anxious    0.53 0.29 0.13 0.04 0.38
jittery    0.54 0.31 0.12 0.04 0.00
nervous    0.70 0.22 0.06 0.02 0.00
tense      0.59 0.28 0.10 0.03 0.00
upset      0.74 0.18 0.05 0.02 0.00
at.ease    0.14 0.33 0.35 0.18 0.00
calm       0.14 0.34 0.36 0.17 0.00
confident  0.16 0.33 0.37 0.14 0.00
content    0.17 0.35 0.35 0.13 0.01
relaxed    0.12 0.30 0.40 0.18 0.00
```

Now do it again, using the `omegaSem` function which calls the *lavaan* package to do a SEM analysis and report both the EFA and CFA solutions. `omega` just reports the EFA solution.

R code

```
omegaSem(msq1[select],nfactors = 2)   #specify a two factor solution
```

```
Call: omegaSem(m = msq1[select], nfactors = 2)
Omega
Call: omega(m = m, nfactors = nfactors, fm = fm, key = key, flip = flip,
    digits = digits, title = title, sl = sl, labels = labels,
    plot = plot, n.obs = n.obs, rotate = rotate, Phi = Phi, option = option)
Alpha:                 0.83
G.6:                   0.86
Omega Hierarchical:    0.45
Omega H asymptotic:    0.51
Omega Total            0.87

Schmid Leiman Factor loadings greater than  0.2
             g   F1*   F2*   h2   u2   p2
```

```
anxious-  0.36       -0.57 0.46 0.54 0.28
jittery-  0.35       -0.52 0.40 0.60 0.31
nervous-  0.43       -0.57 0.51 0.49 0.36
tense-    0.50       -0.62 0.63 0.37 0.39
upset-    0.35       -0.29 0.25 0.75 0.50
at.ease   0.52  0.59       0.64 0.36 0.43
calm      0.49  0.47 -0.21 0.51 0.49 0.48
confident 0.31  0.58       0.46 0.54 0.21
content   0.40  0.65       0.59 0.41 0.26
relaxed   0.51  0.48 -0.22 0.53 0.47 0.48

With eigenvalues of:
  g F1* F2*
1.8 1.6 1.5

general/max  1.13   max/min =   1.05
mean percent general =  0.37    with sd =  0.1 and cv of  0.28
Explained Common Variance of the general factor =  0.37

The degrees of freedom are 26  and the fit is  0.24
The number of observations was  3032  with Chi Square =  721.36  with prob <  2.4e-135
The root mean square of the residuals is  0.04
The df corrected root mean square of the residuals is  0.05
RMSEA index =  0.094  and the 10 \% confidence intervals are  0.088 0.1
BIC =  512.92

Compare this with the adequacy of just a general factor and no group factors
The degrees of freedom for just the general factor are 35  and the fit is  1.67
The number of observations was  3032  with Chi Square =  5055.64  with prob <  0
The root mean square of the residuals is  0.21
The df corrected root mean square of the residuals is  0.24

RMSEA index =  0.218  and the 10 \% confidence intervals are  0.213 0.223
BIC =  4775.04

Measures of factor score adequacy
                                                 g  F1*  F2*
Correlation of scores with factors             0.67 0.77 0.76
Multiple R square of scores with factors       0.45 0.60 0.59
Minimum correlation of factor score estimates -0.09 0.19 0.17

 Total, General and Subset omega for each subset
                                                   g  F1*  F2*
Omega total for total scores and subscales    0.87 0.84 0.79
Omega general for total scores and subscales  0.45 0.33 0.30
Omega group for total scores and subscales    0.36 0.51 0.49

 The following analyses were done using the  lavaan  package

 Omega Hierarchical from a confirmatory model using sem =  0.59
 Omega Total  from a confirmatory model using sem =  0.88
With loadings of
             g  F1*  F2*   h2   u2   p2
anxious    0.35      0.59 0.47 0.53 0.26
jittery    0.43      0.43 0.37 0.63 0.50
nervous    0.44      0.58 0.53 0.47 0.37
tense      0.55      0.57 0.63 0.37 0.48
upset      0.39      0.31 0.25 0.75 0.61
```

```
at.ease-   0.69 0.43      0.65 0.35 0.73
calm-      0.72 0.27      0.59 0.41 0.88
confident- 0.22 0.71      0.55 0.45 0.09
content-   0.34 0.73      0.65 0.35 0.18
relaxed-   0.71 0.29      0.59 0.41 0.85

With eigenvalues of:
  g F1* F2*
2.6 1.4 1.3

The degrees of freedom of the confimatory model are  25  and the fit is  529.8061  with p =  0
general/max  1.9   max/min =   1.07
mean percent general =  0.49    with sd =  0.28 and cv of  0.56
Explained Common Variance of the general factor =  0.5

Measures of factor score adequacy
                                                 g  F1*  F2*
Correlation of scores with factors             0.87 0.85 0.80
Multiple R square of scores with factors       0.75 0.72 0.64
Minimum correlation of factor score estimates 0.51 0.43 0.28

 Total, General and Subset omega for each subset
                                                 g  F1*  F2*
Omega total for total scores and subscales    0.88 0.87 0.80
Omega general for total scores and subscales  0.59 0.48 0.35
Omega group for total scores and subscales    0.30 0.39 0.45

To get the standard sem fit statistics, ask for summary on the fitted object
```

Parallel Forms

The `sai` data set includes 20 items. 10 overlap with the `msqR` data set and are used for most examples. But we may also score anxiety from the second set of items. We can use either the `scoreItems` or the `scoreOverlap` functions. The latter function corrects for the fact that the positive and negative subsets of the anxiety scales overlap with the total scale.

R code

```
sai.parallel <- scoreOverlap(sai.alternate.forms,sai1)
sai.parallel
```

```
Call: scoreOverlap(keys = sai.alternate.forms, r = sai1)

(Standardized) Alpha:
pos1 neg1 anx1 pos2 neg2 anx2
0.86 0.82 0.87 0.83 0.73 0.80

(Standardized) G6*:
pos1 neg1 anx1 pos2 neg2 anx2
0.87 0.84 0.78 0.84 0.79 0.72

Average item correlation:
```

```
pos1 neg1 anx1 pos2 neg2 anx2
0.54 0.48 0.40 0.50 0.36 0.28

Number of items:
pos1 neg1 anx1 pos2 neg2 anx2
   5    5   10    5    5   10

Signal to Noise ratio based upon average r and n
pos1 neg1 anx1 pos2 neg2 anx2
 5.9  4.7  6.6  4.9  2.8  4.0

Scale intercorrelations corrected for item overlap and attenuation
 adjusted for overlap correlations below the diagonal, alpha on the diagonal
 corrected correlations above the diagonal:
      pos1  neg1  anx1  pos2  neg2  anx2
pos1  0.86 -0.60 -0.90  0.95 -0.57 -0.95
neg1 -0.50  0.82  0.89 -0.40  0.98  0.81
anx1 -0.77  0.75  0.87 -0.77  0.87  1.00
pos2  0.80 -0.33 -0.66  0.83 -0.41 -0.86
neg2 -0.45  0.76  0.70 -0.32  0.73  0.82
anx2 -0.78  0.66  0.83 -0.70  0.63  0.80

 In order to see the item by scale loadings and frequency counts of the data
 print with the short option = FALSE
```

Inter rater reliability using the `ICC` function

We use the same data as in Table 5. This saved here in a compact form and then analyzed.

R code

```
example <- structure(list(c(1, 1, 1, 4, 2, 3, 1, 3, 3, 5, 2), c(1, 1, 2,
1, 4, 4, 4, 4, 5, 5, 2.4), c(3, 3, 3, 2, 3, 3, 6, 4, 4, 5, 2.4
), c(2, 2, 2, 6, 5, 4, 4, 4, 6, 6, 3), c(3, 5, 4, 2, 3, 6, 6,
6, 5, 5, 3.4), c(2, 2.4, 2.4, 3, 3.4, 4, 4.2, 4.2, 4.6, 5.2,
4)), .Names = c("V1", "V2", "V3", "V4", "V5", "Mean"), row.names = c("S1",
"S2", "S3", "S4", "S5", "S6", "S7", "S8", "S9", "S10", "Mean"), class = "data.frame")

example #show it

 ICC(example[1:10,1:5])  #find the ICCs for the 10 subjects and 5 judges:
```

```
Call: ICC(x = example[1:10, 1:5])

Intraclass correlation coefficients
                         type  ICC   F df1 df2       p lower bound upper bound
Single_raters_absolute   ICC1 0.29 3.0   9  40 0.00748       0.045        0.66
Single_random_raters     ICC2 0.32 4.3   9  36 0.00078       0.085        0.67
Single_fixed_raters      ICC3 0.40 4.3   9  36 0.00078       0.125        0.74
Average_raters_absolute ICC1k 0.67 3.0   9  40 0.00748       0.190        0.91
Average_random_raters   ICC2k 0.70 4.3   9  36 0.00078       0.317        0.91
Average_fixed_raters    ICC3k 0.77 4.3   9  36 0.00078       0.418        0.93

 Number of subjects = 10     Number of Judges =  5
```

Reliability over time: the `multilevelReliability` *function*

The VALE data set has four replications of the anxiety items. We use the `multilevel.reliability` function. The first and third were separated by several days, the second and fourth were 30 minutes following the first and third obserations.

R code

```
vale.mlr <- multilevel.reliability(vale,grp="id",Time="time",items=select)
vale.mlr    #show the output
```

```
Multilevel Generalizability analysis
Call: multilevel.reliability(x = vale, grp = "id", Time = "time", items = select)

The data had  77  observations taken over  4  time intervals for  10 items.

 Alternative estimates of reliabilty based upon Generalizability theory

RkF  =  0.9 Reliability of average of all ratings across all items and  times (Fixed time effects)
R1R  =  0.58 Generalizability of a single time point across all items (Random time effects)
RkR  =  0.84 Generalizability of average time points across all items (Random time effects)
Rc   =  0.37 Generalizability of change (fixed time points, fixed items)
RkRn =  0.7 Generalizability of between person differences averaged over time (time nested within people)
Rcn  =  0 Generalizability of within person variations averaged over items  (time nested within people)

 These reliabilities are derived from the components of variance estimated by ANOVA
             variance Percent
ID               0.04    0.04
Time             0.00    0.00
Items            0.35    0.36
ID x time        0.02    0.02
ID x items       0.28    0.29
time x items     0.01    0.01
Residual         0.28    0.29
Total            0.98    1.00

 The nested components of variance estimated from lme are:
         variance Percent
id        5.5e-02 5.6e-02
id(time)  1.6e-09 1.6e-09
residual  9.2e-01 9.4e-01
total     9.8e-01 1.0e+00

To see the ANOVA and alpha by subject, use the short = FALSE option.
 To see the summaries of the ICCs by subject and time, use all=TRUE
 To see specific objects select from the following list:
 ANOVA s.lmer s.lme alpha summary.by.person summary.by.time ICC.by.person ICC.by.time lmer long Call
```