Reliability

William Revelle and David M. Condon
Northwestern University

Abstract

Separating the signal in a test from the irrelevant noise is a challenge for all measurement. Low test reliability limits test validity, attenuates important relationships, and can lead to regression artifacts. Multiple approaches to the assessment and improvement of reliability are discussed. The advantages and disadvantages of several different approaches to reliability are considered. Practical advice on how to assess reliability using open source software is provided.

Contents

---

contact: William Revelle revelle@northwestern.edu
To appear in the Wiley-Blackwell Handbook of Psychometric Testing,
Paul Irwing, Tom Booth and David Hughes (Eds.)
Draft version of February 12, 2014

Introduction

All measures reflect an unknown mixture of interesting signal and uninteresting or irrelevant noise. Separating signal from noise is the primary challenge of measurement and is the fundamental goal of all approaches to reliability theory. What makes this challenge particularly difficult is that what is signal to some is noise to others. In climate science, short term variations in weather mask long term trends in climate. In oceanography, variations in waves mask tidal effects, waves and tides in turn mask long term changes in sea level. Within psychology, stable individual differences in affective traits contaminate state measures of momentary affective states; acquiescence and extreme response tendencies contaminate trait measures; moment to moment or day to day fluctuations in alertness or motivation affect measures of ability. All of these examples may be considered as problems of reliability: separating signal from noise. They also demonstrate that the classification of signal depends on what is deemed relevant. For indeed, meteorologists care about the daily weather, climate scientists about long term trends in climate; similarly, emotion researchers care about one's current emotion, personality researchers care about stable consistencies and long term trends.

Whether recording the time spent walking to work or the number of questions answered on an exam, people differ. They differ not only from each other, but from measure to measure. Thus, one of us walks to work in 16 minutes one day, but 15.5 minutes the next and 16.5 on the third day. We can say that his mean time is 16 minutes with a standard deviation of .5 minutes. When asked how long it takes him to get to work, we should say that our best estimate is 16 minutes but we would expect to observe anything between 15 and 17 minutes. We tend to emphasize his central tendency (16 minutes) and consider the variation in his walking rate as irrelevant noise. The expected score across multiple replications that minimizes squared deviations is just the arithmetic average score. In the "classical test theory" (CTT) as originally developed by Spearman (1904), the expected score across an infinite number of replications is known as the *true score*. True score defined this way should not be confused with *Platonic Truth* for if there is any systematic bias in the observations, then the mean score will show this bias (Lord & Novick, 1968).

By defining true score, θ, as the expected observed score, $\theta = \mathcal{E}(x)$, and error as the deviation of an observed score from this expectation, $\varepsilon = x - \mathcal{E}(x)$, error is independent of observed score for it will have an expected value of 0 for all values of true score. That is, for raw scores, X, and deviation scores $x = X - \bar{X}$ the scores for an individual may be expressed as

$$X_i = \Theta_i + E_i \iff x_i = \theta_i + \varepsilon_i \tag{1}$$

and because across individuals the covariance of true and error score, $\sigma_{\theta\varepsilon} = 0$,

$$\sigma_x^2 = \sigma_\theta^2 + \sigma_\varepsilon^2 + \cancel{2\sigma_{\theta\varepsilon}} = \sigma_\theta^2 + \sigma_\varepsilon^2. \tag{2}$$

Just as we can decompose the observed scores into two components, so can we decompose the observed score variance into the variance of the expected (true) scores and the variance of error scores.

Furthermore, because observed scores are the sum of expected scores and error scores, the covariance of observed score with the expected score is just σ_θ^2 and the correlation of true scores with observed scores will be

$$\rho_{\theta x} = \frac{\sigma_{\theta x}}{\sigma_\theta \sigma_x} = \frac{\sigma_\theta^2}{\sigma_\theta \sigma_x} = \frac{\sigma_\theta}{\sigma_x}. \tag{3}$$

This means that the squared correlation of true scores with observed scores (which is the amount of variance shared between observed and true scores) is the ratio of their respective variances:

$$\rho_{\theta x}^2 = \frac{\sigma_\theta^2}{\sigma_x^2}. \tag{4}$$

The *reliability* of a test is defined as this squared correlation of true score with observed score or as the ratio of true to observed variances. Expressing this in engineering terms of the ratio of signal (S) in a test to noise (N) (Brennan & Kane, 1977; Cronbach & Gleser, 1964)

$$\frac{S}{N} = \frac{\rho_{\theta x}^2}{1 - \rho_{\theta x}^2}. \tag{5}$$

Finally, from regression, the true deviation score predicted from the observed score is just

$$\hat{\theta} = \beta_{\theta x} x = \frac{\sigma_\theta^2}{\sigma_x^2} x = \rho_{\theta x}^2 x. \tag{6}$$

That is, the estimated true deviation score is just the observed deviation score times the test reliability. The estimated true score will be the mean of the raw scores + this estimated true score:

$$\hat{\Theta} = \bar{X} + \beta_{\theta x} x = \bar{X} + \frac{\sigma_\theta^2}{\sigma_x^2} x = \bar{X} + \rho_{\theta x}^2 x.$$

Reliability and Validity

Although the validity of a test reflects the content of the items and the particular criterion of interest, it is important to note that a test can not correlate with *any* criterion, y, more than it can correlate with the latent variable that it measures, θ. That is $r_{xy} \leq \rho_{\theta x}$. This logically is the upper bound of the validity of a test and thus validity must be less than or equal to the square root of the reliability $r_{xy} \leq \rho_{\theta x} = \sqrt{\rho_{\theta x}^2}$.

Using reliability

There are three primary reasons to be concerned about a measure's reliability. The first is that the relationship between any two constructs will be *attenuated* by the level of reliability of each measure: two constructs can indeed be highly related at a latent level, but if the measures are not very reliable, the observed correlation will be reduced. The second reason that understanding reliability is so important is the problem of *regression to the mean*. Failing to understand how reliability affects the relationship between observed scores and their expected values plagues economists, sports fanatics, and military training officers. The final reason to examine reliability is to estimate the true score given an observed score and to establish confidence intervals around this estimate based upon the *standard error* of the observed scores.

Correction for attenuation

The original development of reliability theory was to estimate the correlation of latent variables in terms of observed correlations "corrected" for their reliability (Spearman, 1904). Measures of "mental character" showed almost the same correlations (.52) between pairs of brothers as did various physical characteristics (.52), but when the mental characters correlations were corrected for their reliability, they were shown to be much more related (.81) (Spearman, 1904).

The logic of correcting for attenuation is straight forward. For even if observed scores are contaminated by error, the errors are independent of the true scores, and the covariance between the observed scores of two different measures, x and y, just reflects the covariance of their true scores. Consider two observed variables, x and y, which are imperfect measures of two latent traits, θ and ψ:

$$x = \theta + \varepsilon \qquad y = \psi + \zeta$$

with variances

$$\sigma_x^2 = \sigma_{\theta+\varepsilon}^2 \qquad \sigma_y^2 = \sigma_{\psi+\zeta}^2$$

and reliabilities

$$\rho_{\theta x}^2 = \frac{\sigma_\theta^2}{\sigma_x^2} \qquad \rho_{\psi y}^2 = \frac{\sigma_\psi^2}{\sigma_y^2}$$

and covariance

$$\sigma_{xy} = \sigma_{(\theta+\varepsilon)(\psi+\zeta)} = \sigma_{\theta\psi} + \cancel{\sigma_{\theta\zeta} + \sigma_{\psi\varepsilon} + \sigma_{\varepsilon\zeta}} = \sigma_{\theta\psi}.$$

Then the correlation between the two latent variables may be expressed in terms of the observed correlation and the two reliabilities:

$$\rho_{\theta\psi} = \frac{\sigma_{\theta\psi}}{\sigma_\theta \sigma_\psi} = \frac{\sigma_{xy}}{\sqrt{\rho_{\theta x}^2 \sigma_x^2 \rho_{\psi y}^2 \sigma_y^2}} = \frac{r_{xy}}{\sqrt{\rho_{\theta x}^2 \rho_{\psi y}^2}}.$$

That is, the correlation between the true parts of any two tests will the ratio of their observed correlation to the square root of their respective reliabilities. This *correction for attenuation* is perhaps the most important use of reliability theory, for it allows for an estimate of the true correlation between two constructs when the constructs are perfectly measured, without error. It does require, however, that we find the reliability of the separate tests.

The concept that observed covariances reflect true covariances is the basis for structural equation modeling in which relationships between observed scores are expressed in terms of relationships between latent scores and the reliability of the measurement of the latent variables. By correcting for unreliability in this way we are able to determine the underlying latent relationships without the distraction of measurement error.

Regression to the mean

First considered by Galton (1886, 1889) as he was developing the correlation coefficient, reversion to mediocrity was the observation that the offspring of tall parents tended to be shorter, just as those of short parents tended to be taller. Although originally interpreted as of interest only to geneticists, the concept of *regression to the mean* is a classic problem of reliability theory that is unfortunately not as well recognized as it should be (Stigler, 1986, 1997). Whenever groups are selected on the basis of being extreme on an observed variable, the scores on a retest will be closer to the mean than they were originally. Classic examples include the tendency of companies with award winning CEOs to become less successful than comparable companies whose CEOs do not win the award (Malmendier & Tate, 2009), for flight instructors to think that rewarding good pilots is counter productive because they get worse on their next flight (Kahneman & Tversky, 1973), for athletes who make the cover of *Sports Illustrated* to do less well following the publication (Gilovich, 1991), for training programs for disadvantaged children to help their students (Campbell & Kenny, 1999) and for the breeding success of birds to improve following prior failures (Kelly & Price, 2005). Indeed the effect of regression to mean artifacts on the market value of baseball players was the subject of the popular book and movie, *Moneyball* (Lewis, 2004). A critical review of various examples of regression artifacts in chronobiology has the the impressive title of "how to show that unicorn milk is a chronobiotic" and provides thoughtful simulated examples (Atkinson, Waterhouse, Reilly, & Edwards, 2001). Regression effects should be controlled for when trying to separate placebo from treatment effects in behavioral and drug intervention studies (Davis, 2002).

So, if it is so well known, what is it? If observed score is imperfectly correlated with true score (Equation 3) then it is also correlated with error because

$$\sigma_x^2 = \sigma_\theta^2 + \sigma_\varepsilon^2$$

$$\sigma_{\varepsilon x} = \sigma_{\varepsilon(\theta+\varepsilon)} = \cancel{\sigma_{\varepsilon\theta}} + \sigma_\varepsilon^2 = \sigma_\varepsilon^2$$

and thus

$$\rho_{\varepsilon x} = \frac{\sigma_{\varepsilon x}}{\sqrt{\sigma_\varepsilon^2 \sigma_x^2}} = \frac{\sigma_\varepsilon^2}{\sqrt{\sigma_\varepsilon^2 \sigma_x^2}} = \frac{\sigma_\varepsilon}{\sigma_x}. \tag{7}$$

That is, individuals with extreme observed scores might well have extreme true scores, but are most likely to also have extreme error scores. From Equation 6 we see that for any observed score, the expected true score is regressed towards the mean with a slope of $\rho_{x\theta}^2$.

In the case of pilot trainees, if the reliability of flying skill is .5, with a mean score of 50 and a standard deviation of 20, the top 10% of the fliers will have an average score of 75.6 on their first trial but will regress 12.8 points (the reliability value * their deviation score) towards the mean or have a score of 62.8 on the second trial. The bottom 10%, on the other hand, will have scores far below the mean with an average of 24.4 but improve 12.8 points on their second flight to 37.2. Flight instructors seeing this result will falsely believe that punishing those who do badly leads to improvement, perhaps due to heightened effort, while rewarding those who do well leads to a decrease in effort (Kahneman & Tversky, 1973). Similarly, because the mean batting average in baseball is $\approx .260$ with a standard deviation of $\approx .0275$ and has a year to year reliability of $\approx .38$, those who have a batting average of 3σ or .083 above the average in one year (.343 instead of .260) are expected to be just $1.1\sigma \approx .031$) above the average or .291 in the succeeding year (Schall & Smith, 2000). That is, a spectacular year is most likely followed by a return, not to the overall average, but rather to the player's average.

Standard Error of Observed Score

Given an observation for one person of x_i, what is our best estimate of that person's true score and what is the standard deviation of that estimate? The estimate of true score was found before (Equation 6) and is just $\hat{\theta}_i = \rho^2_{\theta x} x_i$. Since the variance of the error scores is the unreliability times the observed score variance, $(1-\rho^2_{\theta x})\sigma^2_x$, the standard deviation of our estimated true score will be

$$\sigma_\varepsilon = \sigma_x \sqrt{1-\rho^2_{\theta x}}. \tag{8}$$

This means that the 95% confidence interval of a true score will be

$$\rho^2_{\theta x} x_i \pm 1.96 \sigma_x \sqrt{1-\rho^2_{\theta x}}. \tag{9}$$

Note that this confidence interval is symmetric around the regressed score. Thus, for our flight instructor example with a reliability of .5 and a standard deviation of 20, a pilot with an observed score of 70 will have an estimated true score of $60 \pm 1.96 * 20 * \sqrt{1-.50} = 32$ to 88 and our baseball player who was batting .343 had a 95% confidence interval of $.291 \pm 1.96 * .0275 * \sqrt{1-.38} = .249$ to .333! Given this amount of expected variation, it is not surprising that so many baseball aficionados develop superstitious explanations for baseball success; stellar performance one year is not very predictive of performance in the subsequent year.

True score theory

Estimating reliability using parallel tests

Unfortunately, all of the analyses discussed so far are of no use unless we have some way of estimating σ^2_θ and σ^2_ε. With one test, it is obviously impossible to find a unique decomposition into true scores and error scores.

Spearman's basic insight was to recognize that if there are two (or more) observed measures (x and x') that have true scores in common but independent error scores with equal variances, then the correlation of these two measures is a direct estimate of σ^2_θ (Spearman, 1904). For if both x and $x'$are measures of θ, and both have equal amounts of independent error,

$$\sigma^2_e = \sigma^2_{e'}$$

then

$$r_{xx'} = \frac{\sigma_{\theta x}\sigma_{\theta x'}}{\sigma_x \sigma_{x'}} = \frac{\sigma^2_{\theta x}}{\sigma_x \sigma_{x'}}. \tag{10}$$

But since both x and x' are thought to be measures of the same latent variable, and their error variances are equal, then $\sigma_x = \sigma_{x'}$ and $\sigma_{\theta x} = \sigma_{\theta x'} = \sigma_\theta$ and thus the correlation of two parallel tests is the squared correlation of the observed score with the true score:

$$r_{xx'} = \frac{\sigma^2_{\theta x}}{\sigma^2_x} = \frac{\sigma^2_\theta}{\sigma^2_x} = \rho^2_{x\theta} \tag{11}$$

which is defined as the reliability of the test. Reliability is the correlation between two *parallel tests* or measures of the same construct and is the ratio of the amount of true variance in a test to the total variance of the test and is the square of the correlation between either measure and the true score.

Estimating reliability using τ *equivalent measures*

The previous derivation requires the assumption that the two measures of the latent (unobserved) trait are exactly equally good and that they have equal error variances. These are very strong assumptions for "unlike the correlation coefficient, which is merely an observed fact, the reliability coefficient has embodied in it a belief or point of view of the investigator" (Kelley, 1942, p 75). Kelley, of course, was commenting upon the assumption of parallelism as well as the assumption that the test means the same thing as when it is given again. With the assumption of parallelism it is possible to solve the 3 equations (two for variances and one for the covariance) shown in the first two rows of Table 1 for the three unknowns ($\sigma^2_\theta, \sigma^2_\varepsilon$, and λ_1). A relaxation of the exact parallelism assumption is to assume that the covariances of observed scores with true scores are equal ($\lambda_1 = \lambda_2 = \lambda_3$), but that the error variances are unequal (Table 1 lines 1-3). With this assumption of equal covariances with true score (known as *tau equivalence*) we have six equations (one for each correlation between the three tests, and one for each variance) and five unknowns (σ^2_θ, $\lambda_1 = \lambda_2 = \lambda_3$, and the three error variances, $\varepsilon^2_1, \varepsilon^2_1, \varepsilon^2_1$) and we can solve using simple algebra.

Table 1: Estimating the parameters of parallel, τ equivalent, and congeneric tests. To solve for two parallel tests (lines 1-2) require the assumption of equal true ($\lambda_1\sigma_\theta = \lambda_2\sigma_\theta$) and error ($\varepsilon^2_1 = \varepsilon^2_2$) variances. To solve the six equations for three τ equivalent tests (lines 1-3) we can relax this assumption, but require the assumption of equal error variances. Congeneric measures (four or more tests) can be solved with no further assumptions.

Observed correlations and modeled parameters

Variable	$Test_1$	$Test_2$	$Test_3$	$Test_4$
$Test_1$	$\sigma^2_{x_1} = \lambda_1\sigma^2_\theta + \varepsilon^2_1$			
$Test_2$	$\sigma_{x_1x_2} = \lambda_1\sigma_\theta\lambda_2\sigma_\theta$	$\sigma^2_{x_2} = \lambda_2\sigma^2_\theta + \varepsilon^2_2$		
$Test_3$	$\sigma_{x_1x_3} = \lambda_1\sigma_\theta\lambda_3\sigma_\theta$	$\sigma_{x_2x_3} = \lambda_2\sigma_\theta\lambda_3\sigma_\theta$	$\sigma^2_{x_3} = \lambda_3\sigma^2_\theta + \varepsilon^2_3$	
$Test_4$	$\sigma_{x_1x_4} = \lambda_1\sigma_\theta\lambda_4\sigma_t$	$\sigma_{x_2x_4} = \lambda_2\sigma_\theta\lambda_4\sigma_\theta$	$\sigma_{x_3x_4} = \lambda_3\sigma_\theta\lambda_4\sigma_\theta$	$\sigma^2_{x_4} = \lambda_4\sigma^2_\theta + \varepsilon^2_4$

Estimating reliability using congeneric measures

If there are at least four tests, it is possible to solve for the unknown parameters (covariances with true score, true score variance, error score variances) without any further assumptions other than that all of the tests are imperfect measures of the same underlying construct (Table 1). In terms of factor analysis, the *congeneric* model merely assumes that all measures load on one common factor. Indeed, with four or more measures of the same construct it is possible to evaluate how well each measure reflects the construct, λ_i and the amount of error variance in each measure, $r^2_{x_i\theta} = \frac{\lambda^2_i}{\sigma^2_{x_i}}$.

Reliability over what?

The previous paragraphs discuss reliability in terms of the correlations between two or more measures. What is unstated is when or where are these measures given, as well as the meaning of alternative measures. Reliability estimates can be found based upon variations in the overall test, variations over time, variation over items in a test, and variability associated with who is giving the test. Each of these alternatives has a different meaning and sometimes a number of different estimates. In the abstract case of parallel tests or congeneric measurement, the domain of generalization (time, form, items) is not specified. It is possible, however, to organize reliability coefficients in terms of a simple taxonomy (Table 2). Each of these alternatives is discussed in more detail in the subsequent sections.

Table 2: Reliability is the ability to generalize about individual differences across alternative sources of variation. Generalizations within a domain of items use internal consistency estimates. If the items are not necessarily internally consistent, reliability can be estimated based upon the worst split half, β, the average split (corrected for test length) or the best split, λ_4. Reliability across forms or across time is just the Pearson correlation. Reliability across raters depends upon the particular rating design and is one of the family of Intraclass correlations. Functions in R may be used to find all of these coefficients. Except for `cor`, all functions are in the *psych* package.

Generalization over	Type of reliability	R function	Name
Unspecified	Parallel tests	`cor(xx')`	r_{xx}
	Tau equivalent tests	`cov(xx')` and `fa`	r_{xx}
	Congeneric tests	`cov(xx')` and `fa`	r_{xx}
Forms	Alternative form	`cor(x,y)`	r_{xx}
Time	Test-retest	`cor(`$time_1 time_2$`)`	r_{xx}
Split halves	random split half	`splitHalf`	r_{xx}
	worst split half	`iclust` or `splitHalf`	β
	best split half	`splitHalf`	λ_4
Items	Internal consistency		
	general factor (g)	`omega` or `omegaSem`	ω_h
	average	`alpha` or `scoreItems`	α
	smc	`alpha` or `scoreItems`	λ_6
	all common (h^2)	`omega` or `omegaSem`	ω_t
Raters	Single rater	`ICC`	ICC_2, ICC_2, ICC_3
	Average rater	`ICC`	$ICC_{1k}, ICC_{2k}, ICC_{3k}$

Reliability over alternate forms

Perhaps the easiest two to understand because they are just raw correlations are the reliability of *alternate forms* and the reliability of tests over time (*test-retest*). *Alternative form reliability* is just the correlation between two tests measuring the same construct, both measures of which are thought to measure the construct equally well. Such tests might be the same items presented in a different order to avoid cheating on an exam, or made up different items with similar but not identical content (e.g., 2 + 4 = ? and 4 + 2 = ?). Ideally, to be practically useful, such equivalent forms should have equal means and equal variances. Although the intuitive expectation might be that two such tests would correlate perfectly, they won't, for all of the reasons previously discussed. Indeed, the correlation between two such alternate forms gives us an estimate of the amount of variance that each test shares with the latent construct being measured. If we have three or more alternate forms, then their correlations may be treated as if they were τ equivalent or congeneric measures and we can use factor analysis to find each test's covariance (λ_i) with the latent factor, the square of which will be the reliability.

The construction of such alternate forms can be done formulaically by randomizing items from one form to prepare a second or third form, or by creating quasi-matching pairs of items across forms ("the capital of Brazil is ?" and "Brazilia is the capital of ?"). To control for subtle differences in difficulty, multiple groups of items can be matched across forms (e.g., 4 * 9 = ? and 3 * 7 = ? might be easier than 9 * 4 = ?, and 7 * 3 = ?, so form A could have 4 * 9 = ? and 7 * 3

= ? while form B could have 9 * 4 = ? and 3 * 7 = ?).

With the ability to computer generate large sets of equivalently difficult ability items (e.g., Condon & Revelle, 2014; Embretson, 1998; Leon & Revelle, 1985), the construction of alternate forms becomes amazingly straight forward. The typical use of such alternate forms of measurement is to enable equivalent tests to be given to different groups over time without worrying about the particular test items being disclosed by earlier test takers to later test takers.

Stability over time

The second type of reliability (*test-retest*) that is a correlation between two forms is the correlation of the same test given at two different occasions. Unlike the correlations between alternate forms, which should be high, the expected test-retest correlation depends upon the construct being measured. A fundamental question in measuring any construct is its stability over time. Some measures should be stable over time; others should not. Traits such as ability, extraversion, or the propensity to experience positive affect are thought to be relatively consistent across time and space. Although there might be changes in mean scores over time, rank orders of people on these tests should be relatively stable.

There is, however, at least one serious difficulty with test retest measures:

> The retest coefficient on the same form gives, in general, 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)

In addition, test-retest measurement of many constructs (e.g., state measures of positive or negative emotion) are not expected to show any consistency over time. Indeed, the very concept of a state is that it is not consistent over time. Perhaps the earliest discussion of dispositions or propensities (traits) and states may be found in Cicero's *Tuscalan Disputations* in 45 BCE (Cicero, 1877; Eysenck, 1983). Among later but still early work distinguishing between states and traits perhaps Allport & Odbert (1936) is the most influential. More recently, Fleeson (2001) conceived of traits as density distributions of states and Revelle (1986) as the rate of change of achieving a state. The state-trait distinction is used in longitudinal studies by the State-Trait-Occasion model (Cole, Martin, & Steiger, 2005) which explicitly decomposes measures over time into their state and trait components.

When examining the correlates of a putative stable trait measured at one time with a subsequent measure of another trait predicted by the first, the natural question is to what extent is the trait measure at time 1 the same as if it were measured at the later time. Consider the case of intelligence at age 11 predicting subsequent risk for mortality (Deary, Whiteman, Starr, Whalley, & Fox, 2004). How stable is the measure taken at age 11? Correlations of .66 and .54 with performance on the identical exam 68 years and 79 years later (Deary, Pattie, & Starr, 2013; Gow et al., 2011) suggest that the test is remarkably stable. These correlations are even higher when corrected for restriction of range of those taking the retest (there was differential attrition associated with IQ).

The stability of intelligence measures across 68-79 years is in marked contrast to the much lower (but non-zero) correlations of affect over a few years. That positive state affect among high school students is related .34 to positive state affect three years later suggests that the measure reflects not just a state component, but rather a reliable trait component as well (Kendall, 2013).

Split half reliability: the reliability of composites

For his dissertation research at the University of London, William Brown (1910) examined the correlations of a number of simple cognitive tasks (e.g., crossing out e's and r's from jumbled French text, adding up single digits in groups of ten) given two weeks apart. For each task, he measured the test-retest reliability by correlating the two time periods and then formed a composite based upon the average of the two scores. He then wanted to know the reliability of these composites so that he could correct the correlations with other composites for their reliability. That is, given a two test composite, $\mathbf{X}$, with a reliability for each test, ρ, what would the composite correlate with a similar (but unmeasured) composite, $\mathbf{X}'$?

Consider $\mathbf{X}$ and $\mathbf{X}'$, both made up of two subtests. The reliability of $\mathbf{X}$ is just its correlation with $\mathbf{X}'$ and can be thought of in terms of the variance-covariance matrix, $\Sigma_{XX'}$:

$$\Sigma_{XX'} = \begin{pmatrix} \mathbf{V_x} & \vdots & \mathbf{C_{xx'}} \\ \cdots & \cdots & \cdots \\ \mathbf{C_{xx'}} & \vdots & \mathbf{V_{x'}} \end{pmatrix} \tag{12}$$

and letting $V_\mathbf{x} = \mathbf{1 V_x 1'}$ and $C_\mathbf{XX'} = \mathbf{1 C_{XX'} 1'}$ where $\mathbf{1}$ is a 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 into two subtests ($\mathbf{X_1}$ and $\mathbf{X_2}$) and their respective variances and covariances. The structure of the two tests seen in Equation 12 becomes

$$\Sigma_{XX'} = \left(\begin{array}{ccc|ccc} \mathbf{V_{x_1}} & \vdots & \mathbf{C_{x_1x_2}} & \mathbf{C_{x_1x'_1}} & \vdots & \mathbf{C_{x_1x'_2}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{C_{x_1x_2}} & \vdots & \mathbf{V_{x_2}} & \mathbf{C_{x_2x'_1}} & \vdots & \mathbf{C_{x_2x'_1}} \\ \hline \mathbf{C_{x_1x'_1}} & \vdots & \mathbf{C_{x_2x'_1}} & \mathbf{V_{x'_1}} & \vdots & \mathbf{C_{x'_1x'_2}} \\ \mathbf{C_{x_1x'_2}} & \vdots & \mathbf{C_{x_2x'_2}} & \mathbf{C_{x'_1x'_2}} & \vdots & \mathbf{V_{x'_2}} \end{array}\right) \tag{13}$$

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

$$r_{xx'} = \frac{2r_1}{1+r_1}. \tag{14}$$

Equation 14 is known as the *split half* estimate of reliability. It is important to note that the split half reliability is not the correlation between the two halves, but rather is adjusted upwards by Equation 14.

In the more general case where the two splits do not have equal variance, ($\mathbf{V_{x_1}} \neq \mathbf{V_{x_2}}$) equation 14 becomes a little more complicated and may be expressed in terms of the total test variance as well as the covariance between the two subtests, or in terms of the subtest variances and correlations (J. Flanagan as cited in Rulon, 1939):

$$r_{xx'} = \frac{4C_{x_1x_2}}{V_x} = \frac{4C_{x_1x_2}}{2C_{x_1x_2}+V_{x_1}+V_{x_2}} = \frac{4r_{x_1x_2}s_{x_1}s_{x_2}}{2C_{x_1x_2}+Vx_1+V_{x_2}} = \frac{4r_{x_1x_2}s_{x_1}s_{x_2}}{2r_{x_1x_2}s_{x_1}\sigma_{x_2}+s^2_{x_1}+s^2_{x_2}}. \tag{15}$$

Because the total variance $V_{x_1+x_2} = V_{x_1}+V_{x_2}+2C_{x_1x_2}$, and the variance of the differences is $V_{x_1-x_2} = V_{x_1}+V_{x_2}-2C_{x_1x_2}$, then $C_{x_1x_2} = \frac{V_{x_1}+V_{x_2}-V_{x_1-x_2}}{2}$, and we can express reliability as a function of the variances of differences scores between the splits and the variances of the two splits

$$r_{xx'} = \frac{4C_{x_1x_2}}{V_x} = \frac{2(V_{x_1}+V_{x_2}-V_{x_1-x_2})}{V_{x_1}+V_{x_2}+2C_{x_1x_2}} = \frac{2(V_{x_1}+V_{x_2}-V_{x_1-x_2})}{V_{x_1}+V_{x_2}+V_{x_1}+V_{x_2}-V_{x_1-x_2}} = \frac{V_{x_1}+V_{x_2}-V_{x_1-x_2}}{V_{x_1}+V_{x_2}-\frac{V_{x_1-x_2}}{2}}. \tag{16}$$

When calculating correlations was tedious compared to finding variances, Equation 16 was a particularly useful formula because it just required finding variances of the two halves as well as the variance of their differences. It is still useful, for it expresses reliability in terms of test variances and recognizes that unreliability is associated with the variances of the difference scores (perfect reliability implies that $V_{x_1-x_2} = 0$).

But how to decide how to split a test? Brown compared the scores at time one with those at time two and then formed a composite of the tests taken at both times. But estimating reliability based upon stability over time implies no change in the underlying construct over time. This is reasonable if measuring speed of processing but is a very problematic assumption if measuring something more complicated:

> ...the reliability coefficient has embodied in it a belief or point of view of the investigator. Consider the score resulting from the item, "Prove the Pythagorean theorem." One teacher asserts that this is a unique demand and that there is no other theorem in geometry that can be paired with it as a similar measure. It cannot be paired with itself if there is any memory, conscious or subconscious, of the first attempt at proof at the time the second attempt is made, for then the mental processes are clearly different in the two cases. The writer suggests that anyone doubting this general principle take, say, a contemporary-affairs test and then retake it a day later. He will undoubtedly note that he works much faster and the depth and breadth of his thinking is much less, – he simply is not doing the same sort of thing as before. (Kelley, 1942, p 75-76)

The alternative to estimating composite reliability by repeating the measure to get two splits is to split the test items from one administration. Thus, it is possible to consider splits such as the odd versus even items of a test. This would reflect differences in speed of taking a test in a different manner than would splitting a test into a first and second part (Brown, 1910). Unfortunately, the number of ways to split a n item test into two is an explosion of possible combinations $\binom{n}{n/2} = \frac{n!}{2(n/2)!^2}$. A 16 item test has 6,435 possible 8 item splits and a 20 item test has 92,378 10 item splits. Most of these possible splits will yield slightly different split half estimates. Consider all possible splits

of the 16 cognitive ability items in the `ability` data set included in the *psych* package (Revelle, 2014) in R (R Core Team, 2014). The split half reliabilities found from equation 15 range from .73 to .87 with an average of .83 (Figure 1).

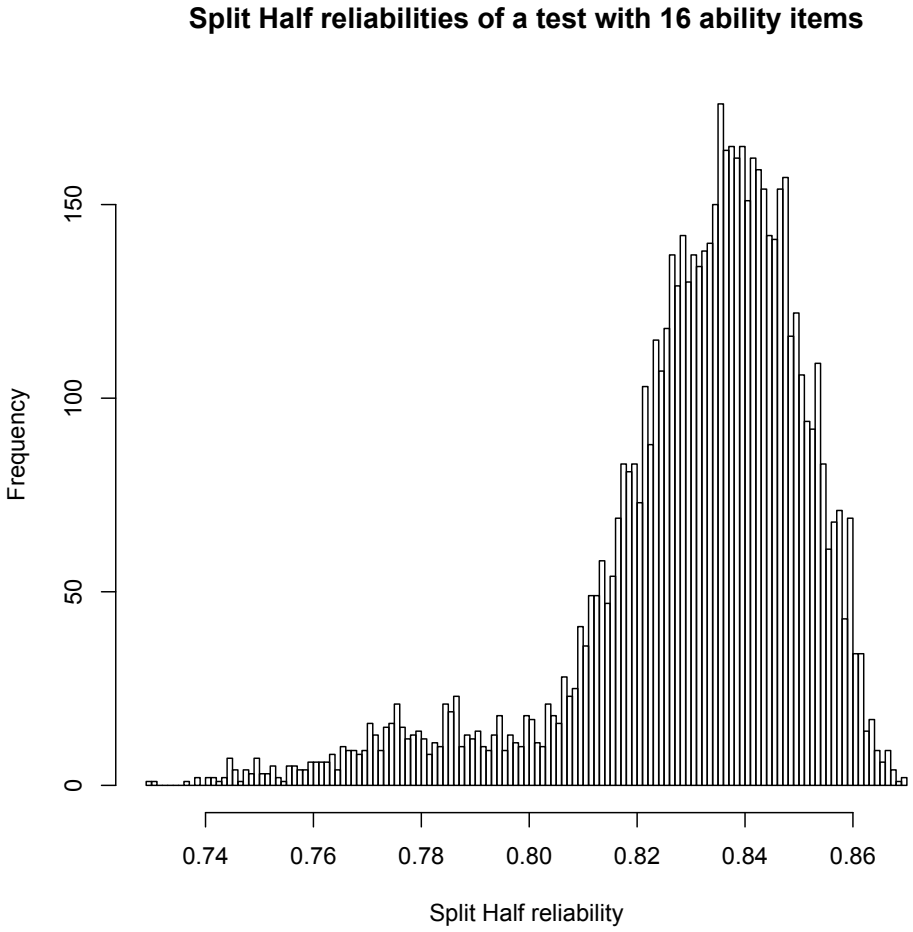

Figure 1. There are 6,435 possible eight item splits of the 16 ability items of the `ability` data set. Of these, the maximum split half reliability is .87, the minimum is .73 and the average is .83. All possible splits were found using the `splitHalf` function.

Internal consistency estimates of reliability

The generalization of Equation 14 to predict the reliability of a composite made up of n tests with average intercorrelation of $\bar{r}_{ij}$ was developed by both (Brown, 1910) and (Spearman, 1910) and has become known as the *Spearman-Brown prophecy formula*.

$$r_{xx} = \frac{n\bar{r}_{ij}}{1+(n-1)\bar{r}_{ij}}. \tag{17}$$

Expressed in terms of the average covariance, $\bar{c}_i$, the unstandardized reliability is

$$r_{xx} = \frac{n\bar{c}_{ij}}{1+(n-1)\bar{c}_{ij}}. \tag{18}$$

That is, the reliability of a composite of n tests (or items) increases as a function of the number of items and the average intercorrelation or covariance of the tests (items). By combining items, each of which is a mixture of signal and noise, the ratio of signal to noise (S/N) increases linearly with the number of items and the resulting composite is a purer measure of signal (Cronbach & Gleser, 1964). If we think of every item as a very weak thread (the amount of signal is small compared to the noise), we can make a very strong rope by binding many threads together (Equation 17).

Considering how people differ from item to item and from trial to trial, Guttman (1945) defined reliability as variation over trials.

> Using this definition, no assumptions of zero means for errors or zero correlations are needed to prove that the total variance of the test is the sum of the error variance and the variance of expected scores; this relationship between variances is an algebraic identity. Therefore, the reliability coefficient is defined without assumptions of independence as the complement of the ratio of error variance to total variance (Guttman, 1945, p 257).

That is,

$$r_{xx} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}. \tag{19}$$

KR-20, λ_3, and α as indicators of internal consistency

Although originally developed to predict the reliability of a composite where the reliability of the subtests is found from their test-retest correlation, the Spearman-Brown methodology was quickly applied to estimating reliability based upon the internal structure of a particular test. Because of the difficulty of finding the average between-item correlation or covariance in equations 17 or 18, reliability was expressed in terms of the total test variance, V_x, and a function of the item variances, Vx_i. For dichotomous items with a probability of being correct, p, or being wrong, q, $V_{x_i} = p_i q_i$ (Kuder & Richardson, 1937). This approach was subsequently generalized to polytomous and continuous items by Guttman (1945) and by Cronbach (1951).

The approach to find σ_e^2 for dichotomous items taken by Kuder & Richardson (1937) was to recognize that for an n-item test, that the average covariance between items estimates the reliable variance of each item, and the error variance for each item will therefore be

$$\sigma_{e_i}^2 = \sigma_{x_i}^2 - \bar{\sigma_{ij}} = \sigma_{x_i}^2 - \frac{\sigma_x^2 - \Sigma\sigma_{x_i}^2}{n(n-1)} = p_i q_i - \frac{\sigma_x^2 - \Sigma p_i q_i}{n(n-1)}$$

and thus

$$r_{xx} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = \frac{\sigma_x^2 - \Sigma(p_i q_i - \frac{\sigma_x^2 - \Sigma p_i q_i}{n(n-1)})}{\sigma_x^2}$$

or

$$r_{xx} = \frac{\sigma_x^2 - \Sigma(p_i q_i)}{\sigma_x^2} \frac{n}{n-1}. \tag{20}$$

The derivation in terms of the total item variance and the sum of the (dichotomous) item variances, Equation 20 was the 20th equation in Kuder & Richardson (1937) and is thus is known as the Kuder-Richardson (20) or KR_{20} formula for reliability. Generalizing this to the polytomous or continous item case, it is known either as α (Cronbach, 1951) or as λ_3 (Guttman, 1945):

$$r_{xx} = \alpha = \lambda_3 = \frac{\sigma_x^2 - \Sigma\sigma_{x_i}^2}{\sigma_x^2} \frac{n}{n-1}. \tag{21}$$

Guttman (1945) considered six different ways to estimate reliability from the pattern of item correlations. His λ_3 coefficient used the average inter-item covariance as an estimate of the reliable variance for each item. He also suggested an alternative, λ_6 which is to use the amount of an item's variance which is predictable by all of the other variables. That is, to find the *squared multiple correlation* or *smc* of the item with all the other items and then find the shared variance as $V_{s_i} = smc_i V_{x_i}$

$$\lambda_6 = \frac{V_x - \Sigma V_{x_i} + \Sigma V_{x_{s_i}}}{V_x}. \tag{22}$$

Guttman (1945) also considered the maximum split half reliability (λ_4). Both λ_4 and λ_6 are obviously more complicated to find than λ_3 or α. To find λ_4 requires finding the maximum among many possible splits and λ_6 requires taking the inverse of the correlation matrix to find the *smc*. But with modern computational power, it is easy to find λ_6 using the `alpha`, `scoreItems` or `splitHalf` functions in the *psych* package. It is a little more tedious to find λ_4 but this can be done by comparing all possible splits for up to 16 items or by sampling thousands of times for larger data sets using the `splitHalf` function.

Consider the 16 ability items with the range of split half correlations as shown in (Figure 1). Using the `splitHalf` function we find that the range of possible splits is from .73 to .87 with an average of .83, $\alpha = .83$, $\lambda_6 = .84$ and a maximum (λ_4) of .87.

Standard error of alpha

There are at least two ways to find the standard error of the estimated α. One is through bootstrapping, the other is through normal theory. Consider the variability in values of α for the 16 ability items for 1525 subjects found in the `ability` data set. Using the `alpha` function to bootstrap by randomly resampling the data (with replacement) 10,000 times yields a distribution of alpha that ranges for .802 to .848 with a mean value of .829 (Figure 2). Compare this to the observed value of .829.

Using the assumption of multivariate normality, Duhachek & Iacobucci (2004) showed that the standard error of α, ASE, is a function of the covariance matrix of the items, $\mathbf{V}$, the number of items, n, and the sample size, N. Defining Q as

$$Q = \frac{2n^2}{(n-1)^2(\mathbf{1'V1})^3}[\mathbf{1'}V\mathbf{1}(tr\mathbf{V}^2 + tr^2\mathbf{V}) - 2tr\mathbf{V}(\mathbf{1'V}^2\mathbf{1})] \tag{23}$$

where *tr* is the trace of a matrix (the sum of the diagonal of a matrix), and $\mathbf{1}$ is a row vector of 1's, then the standard error of α is

$$ASE = \sqrt{\frac{Q}{n}} \tag{24}$$

and the resulting 95% confidence interval is

$$\alpha \pm 1.96\sqrt{\frac{Q}{n}}. \tag{25}$$

These confidence intervals are reported in the `alpha` function. For the 1525 subjects in the 16 `ability` data set, the 95% confidence interval using normal theory is from 0.8123 to 0.8462 which is very similar to the empirical bootstrapped estimates of 0.8166 to 0.8403.

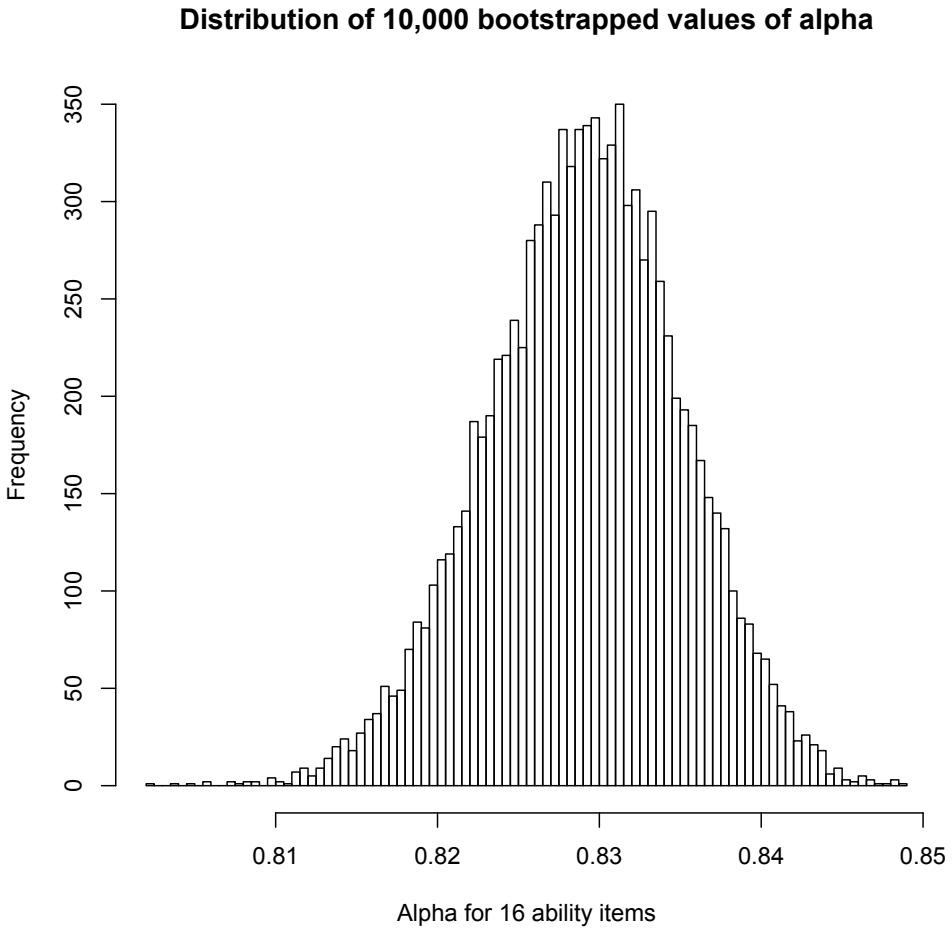

Figure 2. The value of α for the `ability` data set varies across 10,000 bootstrapped resamplings from .80 to .85. For the 1525 subjects in the 16 item `ability` data set, the 95% confidence interval using normal theory is from 0.8123 to 0.8462 which is very similar to the empirical bootstrapped estimates of 0.8168 to 0.8405.

Reliability and item analysis

The α reliability of a scale is a function of the number of items in the scale as well as the average inter-item correlation in the scale (Equations 17-21). Thus, even if the items do not correlate very highly, the reliability of the total scale can be increased by merely adding items. Consider scales ranging in length from 1 to 100 items with average inter-item correlations of .01, .05, .1, .2, .3, and .4 (left hand panel of Figure 3). An α of .9 may be achieved by using 14 highly correlated items ($\bar{r} = .4$), while to achieve this same level of reliability it would take 21 items with a somewhat lower inter-correlation ($\bar{r} = .3$) or 36 with an even lower value ($\bar{r} = .2$). For reference purposes, the average correlation of the 16 ability items in the `ability` data set have an $\bar{r} = .23$ while the five item scales measuring "Big 5 " constructs in the `bfi` data set have average rs ranging from .23 (for openness/intellect) to .46 (for emotional stability).

The ratio of reliable variance to unreliable variance is known as the Signal/Noise ratio and

is just $\frac{S}{N} = \frac{\rho^2}{1-\rho^2}$, which for the same assumptions as for α, will be

$$\frac{S}{N} = \frac{n\bar{r}}{1-\bar{r}}. \tag{26}$$

That is, the S/N ratio increases linearly with the number of items as well as with the average intercorrelation. By thinking in terms of this ratio, the benefits of increased reliability due to increasing the number of items is seen not to be negatively accelerated as it appears when thinking just in reliability units (Equation 21). Indeed, while the S/N ratio is linear with the number of items, it is an accelerating function of the conventional measures of reliability. That is, while the S/N = 1 for a test with a reliability of .5, it is 2 for a test with a reliability of .66, 3 for .75, and 4 for .8, it is 9 for a test with a reliability of .9 and 19 for a reliability of .95 (right hand panel of Figure 3). Depending upon whether the test is norm referenced (comparing two individuals) or domain referenced (comparing an individual to a criterion), there are several different S/N ratios to consider, but all follow this same general form (Brennan & Kane, 1977).

It is not unusual when creating a set of items thought to measure one construct to have some items that do not really belong. This is, of course, an opportunity to use *factor analysis* to explore the structure of the data. If just a few items are suspect, it is possible to find α and λ_6 for all the subsets found by dropping out one item. That is, if an item doesn't really fit, the α and λ_6 values of a scale without that item will actually be higher (Table 3). In the example, five items measuring Agreeableness and one measuring Conscientiousness were scored using the `alpha` function. Although the α and λ_6 values for all six items was .66 and .65 respectively, if item C1 is dropped, the values become .70 and .68. For all other single items, dropping the item leads to a decrease in α and λ_6 either because it reduces the average r, (items A2 - A5) and also because the test length is less (items A1-A5). Note that the `alpha` function recognizes that one item (A1) needs to be reversed scored. If it were not reversed scored, the overall α value would be .44. This reverse scoring is done by finding the sign of the loading of each item on the first principal component of the item set and then reverse scoring those with a negative loading.

If internal consistency were the only goal when creating a test, clearly reproducing the same item many times will lead to an extraordinary reliability. (It will not be one because given the same item repeatedly, some people will in fact change their answers.) But this kind of *tautological consistency* is meaningless and should be avoided. Items should have similar domain content, but not identical content.

Reliability of scales formed from dichotomous or polytomous variables

Whether using using true/false items to assess ability or 4-6 level polytomous (Likert-like) items to assess interests, attitudes, or temperament, the inter-item correlations are reduced from what would be observed with continuous measures. The *tetrachoric* and *polychoric* correlation coefficients are estimates of what the relationship would be between two bivariate, normally distributed items if they had not be dichotomized (tetrachoric) or trichotomized, tetrachotomized, pentachotomized or otherwise broken into discrete but ordered categories (Pearson, 1901). The use of tetrachoric correlations to model what would be the case if the data were in fact bivariate normal had they not be dichotomized is not without critics. The most notable was Yule (1912) who suggested that some phenomena (vaccinated, not vaccinated, alive vs. dead) were truly dichotomous while Pearson & Heron (1913) defended the use of his tetrachoric correlation.

Some have proposed that one should use tetrachoric or polychoric correlations when finding the reliability of categorical scales (Gadermann, Guhn, & Zumbo, 2012; Zumbo, Gadermann, & Zeisser, 2007). We disagree. The Zumbo et al. (2007) procedure estimates the correlation between

Table 3: Item analysis of five Agreeableness items and one Conscientiousness item from the `bfi` data set using the `alpha` function. Note that one item is automatically reversed. Without reverse scoring item A1, $\alpha = .44$ and $\lambda_6 = .52$. The items are: A1: "Am indifferent to the feelings of others.", A2: "Inquire about others' well-being", A3: "Know how to comfort others", A4: "Love children", A5: "Make people feel at ease" and C1: "Am exacting in my work." The item statistics include the number of subjects who answered the item, the raw correlation (inflated by item overlap), a correlation that corrects for scale unreliability and item overlap, the correlation with the scale without that item, the mean and standard deviation for each item. Examining the effect of dropping one item at a time or by looking at the correlations of the item with the scale, item C1 does not belong to this set of items.

```
> alpha(bfi[1:6])

Reliability analysis
Call: alpha(x = bfi[1:6])

  raw_alpha std.alpha G6(smc) average_r   ase mean  sd
      0.66      0.66    0.65      0.25 0.014  4.6 0.8

 lower alpha upper     95% confidence boundaries
0.63 0.66 0.68

 Reliability if an item is dropped:
    raw_alpha std.alpha G6(smc) average_r alpha se
A1-      0.66      0.66    0.64      0.28    0.015
A2       0.56      0.56    0.55      0.20    0.018
A3       0.55      0.55    0.53      0.20    0.018
A4       0.61      0.62    0.61      0.24    0.016
A5       0.58      0.58    0.56      0.22    0.017
C1       0.70      0.71    0.68      0.33    0.014

 Item statistics
       n    r r.cor r.drop mean  sd
A1- 2784 0.52  0.35   0.27  4.6 1.4
A2  2773 0.72  0.67   0.55  4.8 1.2
A3  2774 0.74  0.71   0.57  4.6 1.3
A4  2781 0.61  0.48   0.39  4.7 1.5
A5  2784 0.69  0.61   0.49  4.6 1.3
C1  2779 0.37  0.13   0.11  4.5 1.2

Non missing response frequency for each item
      1    2    3    4    5    6 miss
A1 0.33 0.29 0.14 0.12 0.08 0.03 0.01
A2 0.02 0.05 0.05 0.20 0.37 0.31 0.01
A3 0.03 0.06 0.07 0.20 0.36 0.27 0.01
A4 0.05 0.08 0.07 0.16 0.24 0.41 0.01
A5 0.02 0.07 0.09 0.22 0.35 0.25 0.01
C1 0.03 0.06 0.10 0.24 0.37 0.21 0.01
```

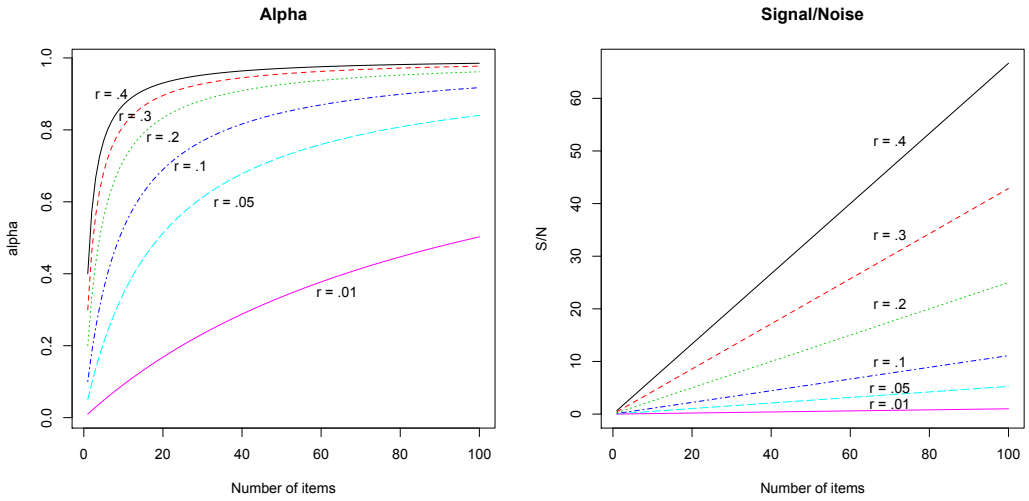

Figure 3. α or λ_3 reliability is an increasing function of the number of items and the inter-item correlation. Just 6 highly correlated items (r=.4) are needed to achieve an $\alpha = .8$ which requires 16 items with more typical correlations (r=.2). Even with barely related items (e.g., r = .05), an α of .8 may be achieved with 76 items. α values of .90 require 14, 21, 36 and 81 for intercorrelations of .4, .3, .2, and .1 respectively. Although α is a decelerating function of the number of items, the relationship between signal/noise ratio and both the inter-item correlations and the number of items is linear.

unobserved continuous scores and true scores rather than the correlation of the observed scores (formed by dichotomizing the unobserved continuous scores) with the latent true score. Reliability is the squared correlation between observed score and true score, not an unobserved score with true score. With a simple simulation it is easy to see that the use of the ϕ or Pearson r provides reliability estimates that closely match the squared correlation of observed and latent but that using the tetachoric or polychoric correlation inflates the reliability estimate.

Partially following the simulation of Zumbo et al. (2007), we simulated 14 items for 10,000 participants using the `sim.congeneric` function. For each participant, a normally distributed latent score was used to generate a probability of response. This latent score was then used to generate 14 different scores broken into 1 of n categories where n ranged from 2 to 7. All items were set to have the same difficulty. The item factor loadings were varied to produce three different sets of data with different levels of reliability (Table 4). Several conclusions can be drawn from this simulation: 1) With the same underlying distribution, inter-item r and thus α or λ_3 increase as the number of response categories increases. 2) The correlation of observed scores with the latent scores also increases as more categories are used. 3) If we are concerned with how well our test scores correlate with the latent scores from which they were generated, the squared correlation of observed scores based upon either simply summing the items or by doing an Item Response Theory based scoring (not shown) is almost exactly the same as the α found using the raw correlations. This is indeed what we would expect given Equations 17-21. This is not the case when we use the tetrachoric or polychoric correlations. The suggestion that we should use "ordinal α" seems incorrect.

Table 4: The average inter-item correlation, and thus α varies as a function of the number of categories in a scale as well as the discrimination parameter (factor loadings) of the items. α based upon the raw correlations more closely approximates the squared correlation of the observed scores with the latent score, $\rho^2_{o\theta}$ than does the α based upon the polychoric correlations. The ratio of alpha to the squared correlation is shown for both the raw, α/ρ^2, and the polychoric based α, α_{poly}/ρ^2. Simulated data using the `sim.congeneric` function.

Simulated results for 10,000 cases.

Factor loading	Number of categories	$\bar{r}$	α	$\rho_{o\theta}$	$\rho^2_{o\theta}$	$r_{\bar{poly}}$	α_{poly}	$\rho_{p\theta}$	$\rho^2_{p\theta}$	α/ρ^2	α_{poly}/ρ^2
	2	0.06	0.48	0.69	0.47	0.10	0.60	0.65	0.43	1.02	1.41
	3	0.07	0.53	0.73	0.53	0.10	0.61	0.71	0.51	1.00	1.20
	4	0.08	0.56	0.75	0.56	0.10	0.60	0.74	0.54	0.99	1.12
.33	5	0.09	0.58	0.76	0.59	0.10	0.61	0.75	0.57	0.99	1.08
	6	0.09	0.58	0.76	0.58	0.10	0.60	0.76	0.58	1.00	1.03
	7	0.09	0.59	0.77	0.59	0.10	0.61	0.76	0.57	1.00	1.06
	2	0.14	0.69	0.83	0.69	0.22	0.80	0.83	0.69	1.00	1.16
	3	0.16	0.73	0.85	0.73	0.22	0.80	0.85	0.73	1.00	1.10
	4	0.18	0.76	0.87	0.77	0.22	0.80	0.87	0.76	0.99	1.04
.47	5	0.19	0.77	0.88	0.77	0.22	0.80	0.88	0.77	0.99	1.03
	6	0.20	0.78	0.88	0.78	0.22	0.80	0.88	0.78	1.00	1.02
	7	0.21	0.79	0.89	0.79	0.22	0.80	0.89	0.79	1.00	1.01
	2	0.26	0.83	0.90	0.80	0.39	0.90	0.90	0.81	1.03	1.12
	3	0.29	0.85	0.92	0.85	0.39	0.90	0.91	0.84	1.00	1.08
	4	0.33	0.87	0.93	0.87	0.39	0.90	0.93	0.87	1.00	1.04
.63	5	0.35	0.88	0.94	0.88	0.39	0.90	0.94	0.88	1.00	1.02
	6	0.36	0.89	0.94	0.89	0.39	0.90	0.94	0.89	1.00	1.01
	7	0.37	0.89	0.94	0.89	0.39	0.90	0.94	0.89	1.00	1.01

Domain sampling theory and structural measures of reliability

A great deal of space has been devoted to finding λ_3 or α. This is not because we recommend the routine use of either, for we don't. They are important to discuss both for historical reasons and because so many applied researchers use them. It would seem that one can not publish a paper without reporting "Cronbach's α". This is unfortunate, for as we (Revelle, 1979; Revelle & Zinbarg, 2009) and many others (e.g., Bentler, 2009; Green & Yang, 2009; Lucke, 2005; Schmitt, 1996; Sijtsma, 2009), including Cronbach & Shavelson (2004), have discussed, α is neither a measure of how well a test measures one thing (Revelle, 1979; Revelle & Zinbarg, 2009; Zinbarg, Revelle, Yovel, & Li, 2005), nor the greatest lower bound for reliability (Bentler, 2009).

The basic problem is that α assesses neither the reliability of a test, nor the internal consistency of a test *unless* the test items all represent just one factor. This is generally not the case. When we think about a test made up of specific items thought to measure a construct, we are concerned not so much with those particular items as we are with how those items represent the larger (perhaps infinite) set of possible items that reflect that construct. Thus, extraversion is not just responding with strong agreement to an item asking about enjoying lively parties, but it also reflects a preference for talking to people rather than reading books, to seeking out exciting situations, to taking charge, and many, many more affective, behavioral, cognitive and goal directed items (Wilt

& Revelle, 2009). Nor is general intelligence just the ability to do spatial rotation problems, to do number or word series, or the ability to do a matrix reasoning task (Gottfredson, 1997). Items will correlate with each other not just because they share a common core or *general factor*, but also because they represent some subgroups of items which share some common affective, behavioral or cognitive content, i.e., *group factors*. Tests made up of such items will correlate with other tests to the extent they both represent the general core that all items share, but also to the extent that specific group factors match across tests.

By a general factor, we mean a factor on which most if not all of the items have a substantial loading. It is analogous to the general 3° background radiation in radio astronomy used as evidence for the "Big Bang". That is, it pervades all items (Revelle & Wilt, 2013). Group factors, on the other hand, represent item clusters where only some or a few items share some common variance in addition to that shared with the general factor. These group factors represent systematic content (e.g., party going behavior vs. talkativeness in measures of extraversion, spatial and verbal content in measures of ability) over and above what is represented by the general factor. Typically, when we assign a name to a scale we are implicitly assuming that a substantial portion of that scale does in fact reflect one thing: the general factor.

Reliability is both the ratio of true score variance to observed variance as well as the correlation of a test with a test *just like it* (Equation 11). But what does it mean to be a test just like another test? If we are concerned with a test made up of a set of items sampled from a domain, then the other test should also represent samples from that same domain. If we are interested in what is common to all the items in the domain, we are interested in the general factor saturation of the test. If we are interested in a test that shares general as well as group factors with another test, then we are concerned with the total reliability of the test.

Seven measures of internal consistency: $\alpha, \lambda_3, \lambda_6, \beta, \omega_g, \omega_t$, *and* λ_4

This distinction between general, group, and total variance in a test, and the resulting correlations with similar tests has led to at least seven different coefficients of internal consistency. These are: α (Cronbach, 1951) (Equation 21) and its equivalent, λ_3 (Guttman, 1945) (Equation 21, which are estimates based upon the average inter item covariance, λ_6, an estimate based upon the squared multiple correlations of the items (Equation 22); β, defined as the worst split half reliability Revelle (1979); ω_g (McDonald, 1999; Revelle & Zinbarg, 2009; Zinbarg et al., 2005), the amount of general factor saturation; ω_t, the total reliable variance estimated by a factor model; and λ_4, the greatest split half reliability.

As an example of the use of these coefficients, consider the 16 ability items discussed earlier (Figure 1). We have already shown that this set has $\alpha = \lambda_3 = .83$ with a $\lambda_6 = .84$ and a $\lambda_4 = .87$. To find the other coefficients requires either cluster analysis for β or factor analysis for the two ω coefficients. A parallel analysis of random data (Horn, 1965) suggests that two principal components or four factors should be extracted. When four factors are found, the resulting structure may be seen in the left hand panel of Figure 4. But these factors are moderately correlated, and when the matrix of factor correlations is in turn factored, the hierarchical structure may be seen in the right hand panel of Figure 4.

Although hierarchical, higher level, or bifactor models of ability have been known for years (Holzinger & Swineford, 1937, 1939; Schmid & Leiman, 1957), it is only relatively recently that these models have been considered when addressing the reliability of a test (McDonald, 1999; Revelle & Zinbarg, 2009; Zinbarg et al., 2005). Rather than consider the reliable variance of a test as reflecting

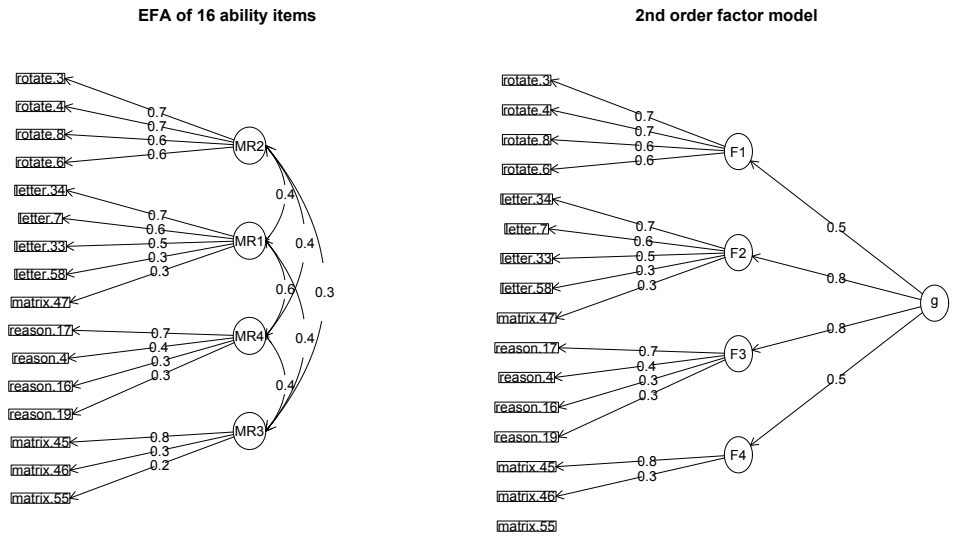

Figure 4. An exploratory factor analysis of 16 ability items shows four moderately correlated factors (left panel). When these in turn are factored, a second order general factor is shown to account for much of the variance of the items (right panel and Table 5).

just one factor, $\mathbf{F}$ with correlations modeled as $\mathbf{R} = \mathbf{FF'} + \mathbf{U^2}$ and reliability as

$$\rho_{xx} = \frac{\mathbf{1FF'1'}}{\mathbf{1FF'1'} + tr(\mathbf{U^2})} \tag{27}$$

the hierarchical approach decomposes test variance in that due to a *general factor*, $\mathbf{g}$, and a number of independent *group factors*, $\mathbf{G_i}$ with a correlation matrix of $\mathbf{R} = (\mathbf{g} + \mathbf{G})(\mathbf{g} + \mathbf{G})' + \mathbf{U^2}$. This representation leads to two different measures of reliability ω_g and ω_t where

$$\omega_g = \frac{\mathbf{1gg'1'}}{\mathbf{1gg'1'} + \mathbf{1GG'1'} + tr(\mathbf{U^2})} = \frac{(\mathbf{1g1'})^2}{\mathbf{1R1'}} \tag{28}$$

and

$$\omega_t = \frac{\mathbf{1gg'1'} + \mathbf{1GG'1'}}{\mathbf{1gg'1'} + \mathbf{1GG'1'} + tr(\mathbf{U^2})} = \frac{\mathbf{1gg'1'} + \mathbf{1GG'1'}}{\mathbf{1R1'}}. \tag{29}$$

ω_g represents that percentage of the variance of a test which is due to the general factor that is common to all of the items in the test, while ω_t is the total amount of reliable variance in the test. When ω_g is estimated using a Schmid & Leiman (1957) transformation or by using a higher order model, it is also known (Revelle & Zinbarg, 2009; Zinbarg et al., 2005) as ω_h for $\omega_{hierarchical}$ to reflect that it represents a hierarchical model. To make the terminology even more confusing, (McDonald, 1999, equations 6.20a and 6.20b) who introduced ω used equations 28 and 29 to define ω without distinguishing between these as two very different models.

The approach of Schmid & Leiman (1957) is to extract a number of factors, $\mathbf{F}$, rotate them obliquely, and then extract one, general, factor, g_h, from the resulting factor intercorrelation matrix. The loadings of the original variables on this higher order factor are found by the product $\mathbf{g} = \mathbf{g_h'F}$. That is, the g loadings are fully mediated by the lower order factors (Gignac, 2007). The original loadings in $\mathbf{F}$ are then residualized by $\mathbf{F^*} = \mathbf{F} - \mathbf{g}_h F$. This results in the model

$$\mathbf{R} = (\mathbf{g} + \mathbf{F^*})(\mathbf{g'} + \mathbf{F^{*'}}) + \mathbf{U^2}$$

where $\mathbf{F}^*$ represents the residualized group factors and $\mathbf{g}$ the matrix of factor coefficients of the original variables. Then ω_g and ω_t are found by equations 28 and 29. This approach is implemented in the `omega` function in the *psych* package. The solution for the 16 ability items is shown in Table 5 where $\omega_h = 0.65$ and $\omega_t = 0.86$.

Table 5: An analysis of the hierarchical structure of 16 ability items shows a general factor and four lower level factors. $\omega_h = 0.65, \alpha(\lambda 3) = 0.83, \lambda_{6^*} = 0.84, \omega_t = 0.86$

An omega analysis table from the psych package in R

Variable	g	F1*	F2*	F3*	F4*	h2	u2	p2
reason.4	0.50			0.27		0.34	0.66	0.73
reason.16	0.42			0.21		0.23	0.77	0.76
reason.17	0.55			0.47		0.52	0.48	0.57
reason.19	0.44			0.21		0.25	0.75	0.77
letter.7	0.52		0.35			0.39	0.61	0.69
letter.33	0.46		0.30			0.31	0.69	0.70
letter.34	0.54		0.38			0.43	0.57	0.67
letter.58	0.47		0.20			0.28	0.72	0.78
matrix.45	0.40				0.66	0.59	0.41	0.27
matrix.46	0.40				0.26	0.24	0.76	0.65
matrix.47	0.42		0.15			0.23	0.77	0.79
matrix.55	0.28				0.14		0.88	0.65
rotate.3	0.36	0.61				0.50	0.50	0.26
rotate.4	0.41	0.61				0.54	0.46	0.31
rotate.6	0.40	0.49				0.41	0.59	0.39
rotate.8	0.32	0.53				0.40	0.60	0.26
SS loadings	3.04	1.32	0.46	0.42	0.55			

Alternatively, a $\omega_{g_{bi}}$ from a *bifactor* solution (Holzinger & Swineford, 1937, 1939) may be found directly by using a confirmatory factor model where $\mathbf{g}$ loads on all variables and the $\mathbf{G}$ matrix has a cluster structure such that items load on one and only one of multiple groups. This approach is implemented in the `omegaSem` function in the *psych* package and makes use of the *sem* package (Fox, Nie, & Byrnes, 2013) to do the confirmatory fit.

Unfortunately, these two approaches do not always agree. ω_{g_h} as found with the Schmid & Leiman (1957) transformation using `omega` is .65 while $\omega_{g_{bi}}$ found with the `omegaSem` function is .75. The reason for the difference is that that the bifactor `sem` solution tends to find the general factor as almost equivalent to the first group factor found with the hierarchical solution. The two approaches differ most obviously in the case of a very small to no general factor (see "where α goes wrong").

Another approach to finding the general factor reliability is through the use of hierarchical cluster analysis with the `ICLUST` algorithm (Revelle, 1979) (implemented in the *psych* package as `iclust`). This approach is very simple: 1) Find the overall correlation matrix; 2) combine the two most similar items into a new (composite) item; 3) find the correlation of this new item with the remaining items; 4) repeat steps 2 and 3 until the worst split half correlation, β, fails to increase. β for a cluster is found by the correlation between the two lower level parts of the cluster (corrected by Equation 15). Because the only variance that the two worst splits share will be general variance,

β is an estimate of the general factor saturation. β found by `iclust` will usually, but not always agree with the estimated ω_h found by the `omega` function.

When α goes wrong: the misuse of α

α is frequently reported without any evidence for scale homogeneity. The assumption is made that if a test has a medium to high value of α it must automatically measure one thing. This is, unfortunately, not correct. For α is just a measure of the average inter item correlation and the number of items. It does not measure homogeneity. We have shown how a 12 item scale with average correlations of .3 (and thus $\alpha = .84$) can represent one general factor in which all the items correlate .3, two correlated but distinct groups with within-group correlations .42 or .54 but between-group correlations of .2 or .1 respectively, or even two unrelated sets with within-group correlations of .66 and between-group correlations of 0 (Revelle & Wilt, 2013).

Consider 10 items from the Big Five Inventory, five of which measure *emotional stability* and five of which measures *intellect* or *openness*. These 10 items are included as part of the `bfi` data set in the *psych* package. Their correlations are shown in Table 6. Using the `alpha` function on these items yields $\alpha = .70$ with an average intercorrelation of .18 (Table 7). This is not particularly impressive, but is not atypical of personality items and meets an arbitrary standard of an "adequate" α. When we examine this result more closely, however, we see that α is not very informative. For ease of demonstration, we reverse code the three items (O1, O3, and O4) that are flagged by the `alpha` function as needing to be reversed keyed. Then plot the resulting correlation matrix using a "heat map" plot from the `cor.plot` function (Figure 5). When this done we see that we have two sub scales (as expected), one measuring (Lack of) Emotional Stability or Neuroticism, the other measuring openness-intellect. The two sub scales have α reliabilities separately of .81 and .61, but only correlate .07 for a split half reliability of the entire 10 items of .14. $(2 * .07/(1 + .07))$. Indeed, although the average correlation within the N scale is .47 and within the O scale is .24, the average inter-item correlation between the two parts is .035. Expressed in terms of *factor analysis* this is an example of where a test has two large *group factors* but not a very large *general factor*. Indeed, $\omega_h = .17$ and $\omega_t = .76$. The value of $\beta = .14$ found by `iclust` agrees exactly with the worst split found by `splitHalf`.

This was obviously a case with two factors that should not be combined into one construct even though the α reliability was adequate. How often this happens in published studies is hard to know, but unless evidence is provided that the test is indeed homogenous, one should treat studies that just report α with great skepticism. If the scale is indeed unifactorial, then α is quite adequate but this needs to be shown rather than assumed.

Other approaches

Reliability is typically considered to be the correlation between two equivalent tests sampled from a domain of items. It is also a variance decomposition: how much of test variance is due to signal, how much to noise. Indeed, the ratio of signal to noise is a simple transformation of the conventional measures of reliability (Equation 5). Recognizing that there are other sources of variation that are systematic but not associated with the signal that concerns us leads to the concept of *generalizability theory* (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach, & Rajaratnam, 1965; Rajaratnam, Cronbach, & Gleser, 1965) which essentially takes a variance decomposition approach to the problem of reliability (Brennan, 1997; Shavelson, Webb, & Rowley, 1989).

Table 6: The correlation matrix of the 10 BFI items suggests two different clusters of content. Note that three items (O1, O3, O4) have been reversed keyed. The items have been rearranged to show the structure more clearly. See also the heat map (Figure 5). The items are N1: Get angry easily. N2: Get irritated easily N3:Have frequent mood swings N4: Often feel blue. N5: Panic easily. O1:Am full of ideas O2: Avoid difficult reading material O3: Carry the conversation to a higher level. O4: Spend time reflecting on things. and O5: Will not probe deeply into a subject. The average correlation within the N set of items is .47, and is .24 within the O set. However, the average inter-item correlation between the two sets is just .035.

`lowerCor(reverse.code(keys = c("O1","O3","O4"),bfi[16:25]))`

Variable	N1	N2	N3	N4	N5	O1-	O3-	O5	O2	O4-
N1	1.00									
N2	0.71	1.00								
N3	0.56	0.55	1.00							
N4	0.40	0.39	0.52	1.00						
N5	0.38	0.35	0.43	0.40	1.00					
O1-	0.05	0.05	0.03	0.05	0.12	1.00				
O3-	0.05	0.03	0.03	0.06	0.08	0.40	1.00			
O5	0.11	0.04	0.06	0.04	0.14	0.24	0.31	1.00		
O2	0.13	0.13	0.11	0.08	0.20	0.21	0.26	0.32	1.00	
O4-	-0.08	-0.13	-0.18	-0.21	-0.11	0.18	0.19	0.18	0.07	1.00

Generalizability theory: reliability over facets

When doing any study in which there are multiple sources of variance, it is important to know their relative contributions in order to improve the quality of the measurement. For example, if student performance is nested within teachers whom are nested within schools, and the tests are given at different times, then all of these terms and their interactions are potential sources of variance in academic performance. If we want to track changes due to an intervention and correct for errors in reliability, we need to know what are the relevant sources of variance in performance. Should we increase the number of students per class room, the number of classrooms, or the number of schools? Similarily, if clinicians rate patients on various symptoms, then we want to know the variance associated with patients, that with symptoms, that with clinicians, as well as the interactions of each. Is it better to use more clinicians or better to have them each rate more symptoms? The procedure as discussed by Cronbach et al. (1972) is first to do an analysis of variance in the *generalizability (G) study* to estimate all of the variance components. Then determine which variance components are relevant for the application in the *decision (D) study* in which one is trying to use the measure (Cronbach et al., 1972). Similarly, the components of variance associated with parts of a test can be analyzed in terms of the generalizability of the entire test.

Consider the data shown in the top of Table 8 which has been adapted from Gleser et al. (1965). 12 patients were rated on six symptoms by two clinicians in a G study. Clearly the patients differ in their total scores (ranging from 13 to 44), and the symptoms differ in their severity (ratings ranging from 9 to 35). The two clinicians seem to agree fairly highly with each other. The ANOVA table (bottom section of Table 8) suggests that there are meaningful interactions of people by items and judges by items. The analysis of variance approach to the measurement of reliability focuses on the relevant facets in an experimental design and decomposes these facets in terms of their contribution to the total variance. The application to the D study uses knowledge gained in the

Table 7: Using `alpha` and `splitHalf` functions to examine the structure of 10 items from the Big 5 Inventory. Although the $\alpha = .7$ might be thought of as satisfactory, the worst split half reliability of .14 suggests that making one scale out of these 10 items is probably a mistake. In fact, the items were chose to represent two relatively independent scales of five items each.

```
This input
> alpha(bfi[16:25],keys = c("O1","O3","O4"))
> splitHalf(bfi[16:25],keys = c("O1","O3","O4"))
produces this output

Reliability analysis
Call: alpha(x = bfi[16:25], keys = c("O1", "O3", "O4"))
 raw_alpha std.alpha G6(smc) average_r   ase mean   sd
       0.7      0.68    0.73      0.18 0.011  2.8 0.75
 lower alpha upper     95% confidence boundaries
0.68 0.7 0.72

 Reliability if an item is dropped:
    raw_alpha std.alpha G6(smc) average_r alpha se
N1       0.64      0.62    0.67      0.16    0.013
N2       0.64      0.63    0.67      0.16    0.013
N3       0.64      0.63    0.68      0.16    0.013
N4       0.66      0.65    0.70      0.17    0.012
N5       0.65      0.64    0.70      0.17    0.012
O1-      0.69      0.67    0.72      0.18    0.011
O2       0.68      0.66    0.72      0.18    0.012
O3-      0.69      0.66    0.71      0.18    0.011
O4-      0.73      0.72    0.76      0.22    0.010
O5       0.69      0.66    0.72      0.18    0.011

 Item statistics
       n    r  r.cor r.drop mean  sd
N1  2778 0.65 0.6574  0.550  2.9 1.6
N2  2779 0.61 0.6121  0.510  3.5 1.5
N3  2789 0.61 0.5916  0.508  3.2 1.6
N4  2764 0.54 0.4766  0.412  3.2 1.6
N5  2771 0.58 0.5148  0.456  3.0 1.6
O1- 2778 0.46 0.3486  0.256  2.2 1.1
O2  2800 0.49 0.3862  0.305  2.7 1.6
O3- 2772 0.48 0.3789  0.270  2.6 1.2
O4- 2786 0.18 0.0086 -0.045  2.1 1.2
O5  2780 0.48 0.3751  0.284  2.5 1.3

Split half reliabilities
Call: splitHalf(r = bfi[16:25], keys = c("O1", "O3", "O4"))
Maximum split half reliability (lambda 4) =  0.78
Guttman lambda 6                          =  0.73
Average split half reliability            =  0.68
Guttman lambda 3 (alpha)                  =  0.68
Minimum split half reliability  (beta)    =  0.14
```

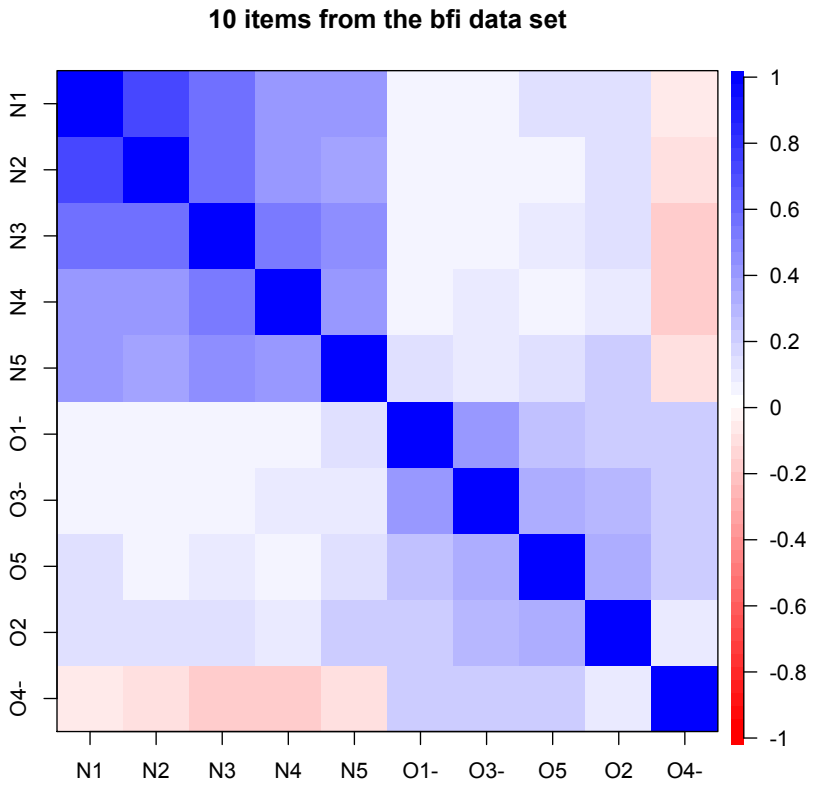

Figure 5. The ten items from the `bfi` data set represent five from the Neuroticism scale and five from the Openness/Intellect scale. Although the overall $\alpha = .70$ is marginally acceptable for a ten item inventory, in fact two subsets correlate .07 and α values of .81 and .61 respectively. The two factor structure is easily identifiable by showing the correlations in a "heat map" where the darker the color, the higher the correlation. Plot done with the `cor.plot` function.

original G study to consider the sources of variance relevant to the particular inference. Examining the components of variance, we can see that people differ a great deal ($\hat{V}_p = .42$), but that there is also a great deal of variance associated with the various symptoms being examined ($\hat{V}_i = .47$). There is negligible variance associated with the mean level of the raters (judges), although there is some degree of interaction between the raters and the patients. There is substantial variation left over in the residual ($\hat{V}_{e_{pij}} = .62$). If the Decision study is concerned with generalizing to the universe of judges but for the same six symptoms, then the ratio of the expected universe score variance (that due to individuals and that due to the interaction of individuals with items) to the expected observed score variance (which includes all terms involving individuals) is $\frac{.419+.427/6}{.419+.427/6+.192/2+.621/12} = .768$. On the other hand, if the generalization is to any pair of judges and any set of six items, the ratio will be $\frac{.419}{.419+.427/6+.192/2+.621/12} = .657$.

Table 8: An example of a Generalization study, adapted from Gleser et al. (1965). 12 patients are rated by two clinicians on six symptoms with a severity ranging from 0 to 6. A simple ANOVA provides the Sums of Squares (not shown) and the Mean Squares. From these, it is possible to estimate the respective variance components to be used in the Decision study.

The raw data

	Item 1		Item 2		Item 3		Item 4		Item 5		Item 6		
Patient	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	total
1	0	0	2	1	2	0	2	1	1	1	1	2	13
2	0	0	2	1	2	0	1	2	2	1	2	1	14
3	0	0	1	1	3	3	2	1	2	1	1	2	17
4	2	0	2	1	2	2	2	1	2	1	4	1	20
5	0	0	1	2	2	0	2	3	3	3	3	3	22
6	2	0	2	1	2	0	4	1	3	3	3	1	22
7	0	1	3	1	3	1	3	4	2	2	2	3	25
8	0	0	0	1	4	3	3	4	2	3	3	3	26
9	1	2	2	1	3	6	1	3	2	3	2	1	27
10	0	1	2	4	3	3	2	2	3	5	3	1	29
11	3	4	2	2	3	2	4	5	3	3	5	5	41
12	1	1	2	4	4	4	3	3	4	6	6	6	44
Total	9	9	21	20	33	24	29	30	29	32	35	29	300

With associated estimated components of variance

Source	df		MS		Estimated Variance Components	
Persons	n-1	11	MS_p	7.65	$\hat{V}_p = (MS_p - MS_{pi} - MS_{pj} + MS_r)/km$	.419
Items	k-1	5	MS_i	12.93	$\hat{V}_i = (MS_i - MS_{ij} - MS_{pi} + MS_r)/nm$	.471
Judges	m-1	1	MS_j	1.00	$\hat{V}_j = (MS_j - MS_{ij} - MS_{pj} + MS_r)/nk$	-0.14*
Persons:Items	(n-1)(m-1)	55	MS_{pi}	1.48	$\hat{V}_{pi} = (MS_{pi} - MS_r)/k$	.427
Persons:Judges	(n-1)(k-1)	11	MS_{pj}	1.77	$\hat{V}_{pj} = (MS_{pi} - MS_r)/m$	.192
Items:Judges	(k-1)(m-1)	5	MS_{ij}	0.87	$\hat{V}_{ij} = (MS_{pi} - MS_r)/n$	.021
Persons:Items:Judges	(n-1)(k-1)(m-1)	55	MS_r	0.62	$\hat{V}_{e_{pij}} = MS_r$	.621

*Negative variance estimates are typically replaced with 0.

A special case of generalizability theory: the Intraclass correlations and the reliability of ratings

The components of variance approach associated with *generalizability theory* is particularly appropriate when considering the reliability of multiple raters or judges. By forming appropriate ratios of variances, various *intraclass correlation coefficients* may be found (Shrout & Fleiss, 1979). The term *intraclass* is used because judges are seen as indistinguishable members of a "class". That is, there is no logical way of distinguishing them.

For example, six subjects are given some score by four different judges (Table 9). The judges differ in their mean leniency and in their range. Values of six different ICC coefficients, their probability of occurring, and confidence intervals for the estimates are reported by the `ICC` function in the *psych* package. `ICC` reports the variance between subjects (MS_b), the variance within subjects (MS_w), the variances due to the judges (MS_j), and the variance due to the interaction of judge by subject (MS_e). The variance within subjects is based upon the pooled SS_j and the SS_e. The reliability estimates from this *generalizability analysis* will depend upon how the scores from the judges are to be used in the *decision analysis*.

If one wants to know how well the scores of a particular rater will match those of another particular rater, then the appropriate ICC is that of a single rater (ICC_{11}). If however, raters are selected at random from a population of raters, the measure of similarity of scores will be ICC_{21}. Both of these measures reflect the fact that raters can differ in their means. If these effects are removed by considering deviations from each judge's average rating, then the agreement between two fixed raters will be the ICC_{31}. The effect of pooling raters is seen in the ICC_{1k}, ICC_{2k} and ICC_{3k} coefficients which benefit in the same way as the Spearman-Brown formula predicts an increase in reliability by pooling items.

Reliability of composites

A common problem is to assess the reliability of a set of tests that are thought to measure one construct. In this case, it is possible to assess the reliability of each test, to examine their intercorrelations, and to estimate the reliability of the overall composite score. This problem is conceptually identical to the estimation of a general factor and group factor contributions to overall reliability (see Equations 28 and 29). Consider two tests X_1 and X_2 with reliabilities $r_{x_1x_1}$ and $r_{x_2x_2}$ and correlation $c_{x_1x_2}$. We want to find the correlation of this composite with another composite of similar subtests and similar covariances. Unlike the assumption we made of parallel tests (Equation 13), here we assume that the covariances between the subtests is not the same as the true variance within each subtest, but we do assume that the variances and covariances for the alternate form will match those of the original two subtests.

$$\left(\begin{array}{cc|cc} \mathbf{V_{x_1}} & \mathbf{C_{x_1x_2}} & \rho_{x_1x'_1}\mathbf{V_{x_1}} & \mathbf{C_{x_1x'_2}} \\ \mathbf{C_{x_1x_2}} & \mathbf{V_{x_2}} & \mathbf{C_{x_2x'_1}} & \rho_{x_2x'_2}\mathbf{V}_{x_2} \\ \hline \rho_{x_1x'_1}\mathbf{V_{x_1}} & \mathbf{C_{x_2x'_1}} & \mathbf{V_{x'_1}} & \mathbf{C_{x'_1x'_2}} \\ \mathbf{C_{x_1x'_2}} & \rho_{x_2x'_2}\mathbf{V}_{x_2} & \mathbf{C_{x'_1x'_2}} & \mathbf{V_{x'_2}} \end{array}\right) \tag{30}$$

For simplicity, we consider the standardized solution expressed in correlations rather than in variances and covariances. Then the correlation between two such tests and thus the reliability of the composite test will be

$$r_{(x_1+x_2)(x_1+x_2)} = \frac{r_{x_1x_1} + r_{x_2x_2} + 2r_{x_1x_2}}{2(1+r_{x_1x_2})}. \tag{31}$$

Table 9: The Intraclass Correlation Coefficient (ICC) measures the correlation between multiple observers when the observations are all of the same class. It is a special case of generalizability theory. The ICC is found by doing an analysis of variance to identify the effects due to subjects, judges, and their interaction. These are combined to form the appropriate ICC. There are at least six different ICCs, depending upon the type of generalization that is to be made. The data and formulae are adapted from Shrout & Fleiss (1979). The analysis was done with the `ICC` function.

Six subjects and 4 raters

Subject	J1	J2	J3	J4	Total
S1	1	3	2	6	12
S2	1	2	6	7	16
S3	2	4	7	6	19
S4	2	5	8	9	24
S5	4	6	8	8	26
S6	5	6	9	10	30
Total	15	26	40	46	127

Produces the following Analysis of Variance Table

Source	Df	SS	MS	Label
Subjects	5	56.21	11.24	MS_b
Within Subjects	18	112.75	6.26	MS_w
Judges	3	97.46	32.49	MS_j
Residuals	15	15.29	1.02	MS_e

Number of subjects (n) = 6 Number of raters (k) = 4

The ANOVA can then be used to find 6 different ICCs.

Variable	type	Formula	ICC	F	df1	df2	p
Single raters absolute	ICC_{11}	$\frac{MS_b - MS_w}{MS_b + (k-1)MS_w}$	0.17	1.79	5	18	0.16
Single random raters	ICC_{21}	$\frac{MS_b - MS_e}{MS_b + (k-1)MS_e + k(MS_j - MS_e)/n}$	0.29	11.03	5	15	0.00
Single fixed raters	ICC_{31}	$\frac{MS_b - MS_e}{MS_b + (k-1)MS_e}$	0.71	11.03	5	15	0.00
Average raters absolute	ICC_{1k}	$\frac{MS_b - MS_w}{MS_b}$	0.44	1.79	5	18	0.16
Average random raters	ICC_{2k}	$\frac{MS_b - MS_e}{MS_b + (MS_j - MS_e)/n}$	0.62	11.03	5	15	0.00
Average fixed raters	ICC_{3k}	$\frac{MS_b - MS_e}{MS_b}$	0.91	11.03	5	15	0.00

In the case that the reliabilities of the two subtests match their intercorrelation, this is identical to Equation 14. It is perhaps useful to note that a composite made up of two reliable but unrelated subtests will have a reliability of the average of the two subtests, even though there is no common factor to the two subtests! For example, a composite of six items of German speaking ability with six items measuring knowledge of sailboat racing, with α reliabilities for the two subtests of .8, and intercorrelation of 0 will still be expected to correlate .8 (have a reliability of .8) with another 12 item composite of six parallel German and six parallel sailing items. The pooled α reliability of such a test will be $\alpha = .68$, even though the $\omega_g = 0$.

Reliability of difference scores

A related problem is the reliability of a difference score. Replacing the $\mathbf{C}_{x_1x_2}$ in Equation 30 with $-\mathbf{C}_{x_1x_2}$ leads to a change in sign of the corrections in Equation 31 and we find that the reliability of a difference score is an inverse function of the correlation between the two tests:

$$r_{(x_1-x_2)(x_1-x_2)} = \frac{r_{x_1x_1} + r_{x_2x_2} - 2r_{x_1x_2}}{2(1 - r_{x_1x_2})}. \tag{32}$$

That is, as the correlation between the two tests tends towards their reliabilities, the reliability of the difference tends toward 0. This is particularly a problem when one wants to interpret differential deficits in cognitive processing by finding the difference, for example, between verbal and spatial abilities, each of which is reliably measured, but which are also highly correlated. Consider the case where $r_{vv} = .9, r_{ss} = .9$ and $r_{vs} = .6$. Then while the composite V+S measure has a reliability of $r_{(v+s)(v+s)} = \frac{.9+.9+2*.6}{2(1+.6)} = \frac{3.0}{3.2} = .9375$, the reliability of the difference V-S is $r_{(v-s)(v-s)} = \frac{.9+.9-2*.6}{2(1-.6)} = \frac{.6}{.8} = .75$. But if the $r_{vs} = .8$ the reliability of the composite will increase only slightly ($r_{(v+s)(v+s)} = .94$ but the reliability of the difference scores decreases considerably ($r_{(v-s)(v-s)} = .5$).

Conclusion

All signals are contaminated by noise. The effect of such contamination is to attenuate latent relationships and to raise the threat of regression artifacts. Perhaps because psychological measures are so threatened with lack of reliability, psychologists have spent more than a century trying to understand the challenges of reliability theory (Traub, 1997). Even as IRT approaches become more prevalent (Bock, 1997; Embretson & Hershberger, 1999; Wright, 1997) the study of reliability is a worthwhile enterprise, for even today, there remains confusion about the ways of estimating and correcting for reliability.

References

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(211).

Atkinson, G., Waterhouse, J., Reilly, T., & Edwards, B. (2001). How to show that unicorn milk is a chronobiotic: The regression-to-the mean statistical artifact. *Chronobiology International: The Journal of Biological & Medical Rhythm Research*, *18*(6), 1041-1053.

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

Bock, R. D. (1997). A brief history of item theory response. *Educational Measurement: Issues and Practice*, *16*(4), 21–33. Retrieved from http://dx.doi.org/10.1111/j.1745-3992.1997.tb00605.x doi: 10.1111/j.1745-3992.1997.tb00605.x

Brennan, R. L. (1997). A perspective on the history of generability theory. *Educational Measurement: Issues and Practice*, *16*(4), 14–20. Retrieved from http://dx.doi.org/10.1111/j.1745-3992.1997.tb00604.x doi: 10.1111/j.1745-3992.1997.tb00604.x

Brennan, R. L., & Kane, M. T. (1977). Signal/noise ratios for domain referenced tests. *Psychometrika*, *42*(4), 609-625.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296-322.

Campbell, D. T., & Kenny, D. (1999). *A primer on regression artifacts*. Guilford Press.

Cicero, M. T. (1877). *Tusculan disputations* (translated by C.D. Yonge, Ed.). New York: Harper & Brothers (Originally published c 45 BCE).

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.

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, (in press).

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

Cronbach, L. J., & Gleser, G. C. (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, *24*(3), 467-480.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

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.

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

Davis, C. E. (2002). Regression to the mean or placebo effect? In H. A. Guess (Ed.), *The science of the placebo: Toward an interdisciplinary research agenda* (p. 158-166). London: BMJ Books.

Deary, I. J., Pattie, A., & Starr, J. M. (2013). The stability of intelligence from age 11 to age 90 years: The Lothian Birth Cohort of 1921. *Psychological Science*, *24*(12), 2361-2368. doi: 10.1177/0956797613486487

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.

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. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380 - 396.

Embretson, S. E., & Hershberger, S. L. (1999). *The new rules of measurement : what every psychologist and educator should know.* Mahwah, N.J.: L. Erlbaum Associates.

Eysenck, H. J. (1983). Cicero and the state-trait theory of anxiety: Another case of delayed recognition. *American Psychologist*, *38*(1), 114-115. doi: 10.1037/0003-066X.38.1.114

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011-1027.

Fox, J., Nie, Z., & Byrnes, J. (2013). sem: Structural equation models [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=sem (R package version 3.1-3)

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1-13.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246-263.

Galton, F. (1889). *Natural inheritance.* London: Macmillan.

Gignac, G. E. (2007). Multi-factor modeling in individual differences research: Some recommendations and suggestions. *Personality and Individual Differences*, *42*(1), 37-48.

Gilovich, T. (1991). *How we know what isn't so.* Free Press.

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

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, p. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79 - 132. doi: DOI:10.1016/S0160-2896(97)90014-3

Gow, A. J., Johnson, W., Pattie, A., Brett, C. E., Roberts, B., Starr, J. M., & Deary, I. J. (2011). Stability and change in intelligence from age 11 to ages 70, 79, and 87: The Lothian Birth Cohorts of 1921 and 1936. *Psychology and Aging*, *26*(1), 232 - 240.

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

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

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

Holzinger, K., & Swineford, F. (1939). *A study in factor analysis: the stability of a bi-factor solution.* (No. 48). Department of Education, University of Chicago. Chicago, Il.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi: 10.1007/BF02289447

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

Kelley, T. L. (1942). The reliability coefficient. *Psychometrika*, *7*(2), 75-83.

Kelly, C., & Price, T. D. (2005). Correcting for regression to the mean in behavior and ecology. *The American Naturalist*, *166*(6), pp. 700-707. doi: 10.1086/497402

Kendall, A. D. (2013). *Depressive and anxiety disorders: Results from a 10-wave latent trait-state modeling study.* Unpublished master's thesis, Northwestern University.

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

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

Lewis, M. (2004). *Moneyball: The art of winning an unfair game.* WW Norton & Company.

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

Lucke, J. F. (2005). The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*, *29*(1), 65–81.

Malmendier, U., & Tate, G. (2009). Superstar CEOs. *The Quarterly Journal of Economics*, *124*(4), 1593-1638.

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

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal*, *6*(2), 559-572.

Pearson, K., & Heron, D. (1913). On theories of association. *Biometrika*, *9*(1/2), 159-315.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0)

Rajaratnam, N., Cronbach, L., & Gleser, G. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*(1), 39-56. doi: 10.1007/BF02289746

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

Revelle, W. (1986). Motivation and efficiency of cognitive performance. In D. R. Brown & J. Veroff (Eds.), *Frontiers of motivational psychology: Essays in honor of J. W. Atkinson* (p. 105-131). New York: Springer.

Revelle, W. (2014). psych: Procedures for personality and psychological research [Computer software manual]. http://cran.r-project.org/web/packages/psych/. Retrieved from http://cran.r-project.org/web/packages/psych/ (R package version 1.4.1)

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., & 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

Rulon, P. J. (1939). A simplified procedure for determing the reliability of a test by split-halves. *Harvard Educational Review*, *9*, 99-103.

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.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350-353.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922 - 932.

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

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

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

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271-295.

Stigler, S. M. (1986). *The history of statistics : The measurement of uncertainty before 1900*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research*, *6*(2), 103-114.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8–14. Retrieved from http://dx.doi.org/10.1111/j.1745-3992.1997.tb00603.x doi: 10.1111/j.1745-3992.1997.tb00603.x

Wilt, J., & Revelle, W. (2009). Extraversion. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (p. 27-45). New York, N.Y.: Guilford Press.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45. Retrieved from http://dx.doi.org/10.1111/j.1745-3992.1997.tb00606.x doi: 10.1111/j.1745-3992.1997.tb00606.x

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *LXXV*, 579-652.

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

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*, 21-29.

Appendix

R functions called

The examples shown made use of various functions and data sets in the *psych* package (Revelle, 2014) in the R statistical system (R Core Team, 2014). The particular functions used were:

`alpha` A function to find α and λ_6 as well as total scores for a set of items.

`scoreItems` A function to find α and λ_6 as well as total scores for multiple scales.

`splitHalf` A function to find all possible split half reliabilities for 16 or fewer items or to sample $> 10,000$ possible splits for more than 16 items. This includes the lowest β and highest λ_4 splits.

`fa` A function for exploratory factor analysis using maximum likelihood, minimal residual, or principal axis factor extraction and a large number of orthogonal and oblique rotations.

`omega` A function to find ω_{g_h} and ω_t for an item set using exploratory factor analysis.

`omegaSem` A function to find $\omega_{g_{bi}}$ and ω_t for an item set using confirmatory factor analysis.

`ICC` A function to find Intraclass Correlation Coefficients.

A number of functions that are convenient for analysis include:

`fa.diagram` Graphically show a factor analysis structure

`cor.plot` Show a heat map of correlations.

`lowerCor` Find and display the lower off diagonal correlation matrix.

`reverse.code` Reverse code specified items

Data sets used for demonstrations:

`bfi` 25 items measuring Extraversions, Agreeableness, Conscientiousness, Emotional Stability, and Openness/Intellect. Adapted from the International Personality Item Pool (Goldberg, 1999) and administered as part of the Synthetic Aperture Personality Assessment (SAPA) project. Number of observations = 2,800.

`ability` 16 ability items given as part of the SAPA project. See Condon & Revelle (2014) for details on this and other open source measures of ability. Number of observations = 1,525.

Sample R code for basic reliability calculations

In order to use the *psych* package functions, it is necessary to install the package. This needs to be done only once, but it is recommended to get the latest version from CRAN at least every six months as R and the *psych* package get updated. Then, for each new session of R, it is necessary to make the *psych* package active by issuing the `library` command. The following examples are done with two built in data sets: `bfi` which includes 25 items taken from the International Personality Item Pool and used as part of the `sapa-project.org` online personality assessment Condon & Revelle (2014). The other example is a set of 16 ability items also collected as part of the `sapa-project.org`. More detail may be found in the package vignette "An Overview of the

psych package" which is included when downloading the package. For all functions, if more help is needed, consult the help menu for that function by ?function (e.g., ? alpha).

For more extensive examples are found in the *psych* package vignette as well as various tutorials at the Personality-Project : http://personality-project.org/r.

```
install.packages(list(c("GPArotation","mvtnorm","MASS")  #do this once
library(psych) #do this every time R is started
```

Once the package is installed, a data set to be analyzed may be read into R using the `read.file` command, or just read in a text editor/spreadsheet and copied into the clipboard. The data may be then be pasted into R using the `read.clipboard` command. In the listing below, the # symbol denotes a comment and the > symbol an R command.

```
#first copy the data to the clipboard then
> my.data <- read.clipboard()
#or, if copying from a spreadsheet
> my.data <- read.clipboard.tab()
#or, use a built in data set such as ability
> my.data <- ability
#then, to see if the data have been entered correctly,
#find out the dimensions of the data set and some descriptive statistics.
> dim(my.data)
> describe(my.data)
```

```
> dim(my.data)
[1] 1525   16
> describe(my.data)
          var    n mean   sd median trimmed mad min max range  skew kurtosis   se
reason.4    1 1442 0.68 0.47      1    0.72   0   0   1     1 -0.75    -1.44 0.01
reason.16   2 1463 0.73 0.45      1    0.78   0   0   1     1 -1.02    -0.96 0.01
reason.17   3 1440 0.74 0.44      1    0.80   0   0   1     1 -1.08    -0.84 0.01
reason.19   4 1456 0.64 0.48      1    0.68   0   0   1     1 -0.60    -1.64 0.01
letter.7    5 1441 0.63 0.48      1    0.67   0   0   1     1 -0.56    -1.69 0.01
letter.33   6 1438 0.61 0.49      1    0.63   0   0   1     1 -0.43    -1.82 0.01
letter.34   7 1455 0.64 0.48      1    0.68   0   0   1     1 -0.59    -1.65 0.01
letter.58   8 1438 0.47 0.50      0    0.46   0   0   1     1  0.12    -1.99 0.01
matrix.45   9 1458 0.55 0.50      1    0.56   0   0   1     1 -0.20    -1.96 0.01
matrix.46  10 1470 0.57 0.50      1    0.59   0   0   1     1 -0.28    -1.92 0.01
matrix.47  11 1465 0.64 0.48      1    0.67   0   0   1     1 -0.57    -1.67 0.01
matrix.55  12 1459 0.39 0.49      0    0.36   0   0   1     1  0.45    -1.80 0.01
rotate.3   13 1456 0.20 0.40      0    0.13   0   0   1     1  1.48     0.19 0.01
rotate.4   14 1460 0.22 0.42      0    0.15   0   0   1     1  1.34    -0.21 0.01
rotate.6   15 1456 0.31 0.46      0    0.27   0   0   1     1  0.80    -1.35 0.01
rotate.8   16 1460 0.19 0.39      0    0.12   0   0   1     1  1.55     0.41 0.01
>
```

Find the α and λ_6 estimates of reliability for this data set.

```
> alpha(ability)
```

```
Reliability analysis
Call: alpha(x = ability)

  raw_alpha std.alpha G6(smc) average_r    ase mean   sd
      0.83      0.83    0.84      0.23 0.0086 0.51 0.25

 lower alpha upper     95% confidence boundaries
0.81 0.83 0.85

 Reliability if an item is dropped:
          raw_alpha std.alpha G6(smc) average_r alpha se
reason.4       0.82      0.82    0.82      0.23   0.0093
reason.16      0.82      0.82    0.83      0.24   0.0091
```

```
reason.17      0.82      0.82    0.82      0.23   0.0093
reason.19      0.82      0.82    0.83      0.24   0.0091
letter.7       0.82      0.82    0.82      0.23   0.0092
letter.33      0.82      0.82    0.83      0.24   0.0092
letter.34      0.82      0.82    0.82      0.23   0.0093
letter.58      0.82      0.82    0.82      0.23   0.0092
matrix.45      0.82      0.83    0.83      0.24   0.0090
matrix.46      0.82      0.82    0.83      0.24   0.0091
matrix.47      0.82      0.82    0.83      0.24   0.0091
matrix.55      0.83      0.83    0.83      0.24   0.0089
rotate.3       0.82      0.82    0.82      0.23   0.0092
rotate.4       0.82      0.82    0.82      0.23   0.0092
rotate.6       0.82      0.82    0.82      0.23   0.0092
rotate.8       0.82      0.82    0.83      0.24   0.0091

 Item statistics
             n    r r.cor r.drop mean   sd
reason.4  1442 0.58  0.54   0.50 0.68 0.47
reason.16 1463 0.50  0.44   0.41 0.73 0.45
reason.17 1440 0.57  0.54   0.49 0.74 0.44
reason.19 1456 0.52  0.47   0.43 0.64 0.48
letter.7  1441 0.56  0.52   0.48 0.63 0.48
letter.33 1438 0.53  0.48   0.44 0.61 0.49
letter.34 1455 0.57  0.53   0.49 0.64 0.48
letter.58 1438 0.57  0.52   0.48 0.47 0.50
matrix.45 1458 0.48  0.42   0.38 0.55 0.50
matrix.46 1470 0.49  0.43   0.40 0.57 0.50
matrix.47 1465 0.52  0.47   0.43 0.64 0.48
matrix.55 1459 0.42  0.35   0.32 0.39 0.49
rotate.3  1456 0.54  0.51   0.44 0.20 0.40
rotate.4  1460 0.58  0.56   0.48 0.22 0.42
rotate.6  1456 0.56  0.53   0.46 0.31 0.46
rotate.8  1460 0.51  0.47   0.41 0.19 0.39

Non missing response frequency for each item
             0    1 miss
reason.4  0.32 0.68 0.05
reason.16 0.27 0.73 0.04
reason.17 0.26 0.74 0.06
reason.19 0.36 0.64 0.05
letter.7  0.37 0.63 0.06
letter.33 0.39 0.61 0.06
letter.34 0.36 0.64 0.05
letter.58 0.53 0.47 0.06
matrix.45 0.45 0.55 0.04
matrix.46 0.43 0.57 0.04
matrix.47 0.36 0.64 0.04
matrix.55 0.61 0.39 0.04
rotate.3  0.80 0.20 0.05
rotate.4  0.78 0.22 0.04
rotate.6  0.69 0.31 0.05
rotate.8  0.81 0.19 0.04
```

Find the ω estimates of reliability. Specify that you want to try a four factor solution.

```
> omega(ability,4)

Omega
Call: omega(m = ability, nfactors = 4)
Alpha:                 0.83
G.6:                   0.84
Omega Hierarchical:    0.65
Omega H asymptotic:    0.76
Omega Total            0.86
```

```
Schmid Leiman Factor loadings greater than  0.2
             g   F1*   F2*   F3*   F4*   h2   u2   p2
reason.4  0.50              0.27       0.34 0.66 0.73
reason.16 0.42              0.21       0.23 0.77 0.76
reason.17 0.55              0.47       0.52 0.48 0.57
reason.19 0.44              0.21       0.25 0.75 0.77
letter.7  0.52        0.35             0.39 0.61 0.69
letter.33 0.46        0.30             0.31 0.69 0.70
letter.34 0.54        0.38             0.43 0.57 0.67
letter.58 0.47        0.20             0.28 0.72 0.78
matrix.45 0.40                   0.66 0.59 0.41 0.27
matrix.46 0.40                   0.26 0.24 0.76 0.65
matrix.47 0.42                        0.23 0.77 0.79
matrix.55 0.28                        0.12 0.88 0.65
rotate.3  0.36  0.61                  0.50 0.50 0.26
rotate.4  0.41  0.61                  0.54 0.46 0.31
rotate.6  0.40  0.49                  0.41 0.59 0.39
rotate.8  0.32  0.53                  0.40 0.60 0.26

With eigenvalues of:
   g  F1*  F2*  F3*  F4*
3.04 1.32 0.46 0.42 0.55

general/max  2.3   max/min =   3.17
mean percent general =  0.58    with sd =  0.2 and cv of  0.35
Explained Common Variance of the general factor =  0.53

The degrees of freedom are 62  and the fit is  0.05
The number of observations was  1525  with Chi Square =  70.19  with prob <  0.22
The root mean square of the residuals is  0.01
The df corrected root mean square of the residuals is  0.03
RMSEA index =  0.01  and the 90 % confidence intervals are  NA 0.019
BIC =  -384.25

Compare this with the adequacy of just a general factor and no group factors
The degrees of freedom for just the general factor are 104  and the fit is  0.78
The number of observations was  1525  with Chi Square =  1186.18  with prob <  5e-183
The root mean square of the residuals is  0.09
The df corrected root mean square of the residuals is  0.13

RMSEA index =  0.083  and the 90 % confidence intervals are  0.078 0.087
BIC =  423.88

Measures of factor score adequacy
                                                 g  F1*   F2*   F3*  F4*
Correlation of scores with factors            0.83 0.80  0.53  0.56 0.71
Multiple R square of scores with factors      0.69 0.64  0.28  0.32 0.50
Minimum correlation of factor score estimates 0.37 0.28 -0.45 -0.36 0.00

 Total, General and Subset omega for each subset
                                                  g  F1*  F2*  F3*  F4*
Omega total for total scores and subscales    0.86 0.77 0.69 0.64 0.53
Omega general for total scores and subscales  0.65 0.23 0.52 0.47 0.27
Omega group for total scores and subscales    0.13 0.53 0.17 0.17 0.26
```