Psychological Methods
2018, Vol. 23, No. 3, 412–433

1082-989X/18/$12.00 http://dx.doi.org/10.1037/met0000144

Thanks Coefficient Alpha, We'll Take It From Here

Daniel McNeish
Utrecht University and University of North Carolina, Chapel Hill

Abstract

Empirical studies in psychology commonly report Cronbach's alpha as a measure of internal consistency reliability despite the fact that many methodological studies have shown that Cronbach's alpha is riddled with problems stemming from unrealistic assumptions. In many circumstances, violating these assumptions yields estimates of reliability that are too small, making measures look less reliable than they actually are. Although methodological critiques of Cronbach's alpha are being cited with increasing frequency in empirical studies, in this tutorial we discuss how the trend is not necessarily improving methodology used in the literature. That is, many studies continue to use Cronbach's alpha without regard for its assumptions or merely cite methodological articles advising against its use to rationalize unfavorable Cronbach's alpha estimates. This tutorial first provides evidence that recommendations against Cronbach's alpha have not appreciably changed how empirical studies report reliability. Then, we summarize the drawbacks of Cronbach's alpha conceptually without relying on mathematical or simulation-based arguments so that these arguments are accessible to a broad audience. We continue by discussing several alternative measures that make less rigid assumptions which provide justifiably higher estimates of reliability compared to Cronbach's alpha. We conclude with empirical examples to illustrate advantages of alternative measures of reliability including omega total, Revelle's omega total, the greatest lower bound, and Coefficient *H*. A detailed software appendix is also provided to help researchers implement alternative methods.

Translational Abstract

Scales are commonly used in psychological research to measure directly unobservable constructs like motivation or depression. These scales are comprised of multiple items, each aiming to provide information about various aspects of the construct of interest. Whenever a scale is used in a psychological study, it is important to report on its reliability. Since the 1950s, the primary method for capturing reliability has been Cronbach's alpha, a method whose status is perhaps best exemplified by its place as one of the most cited scientific articles of all-time, in any field. Despite its overwhelming popularity, the underlying assumptions of Cronbach's alpha have been questioned recently in the statistical literature because these assumptions were commonplace 65 years ago but have largely disappeared from more modern statistical methods for constructing scales. Though the ideas in these statistical articles have the potential to significantly alter how psychological research is conducted and reported, recommendations from the statistical literature have yet to permeate the psychological literature. In this article, the goal is to demonstrate why Cronbach's alpha is no longer the optimal method for reporting on reliability. To differentiate this article from articles appearing in the statistical literature, we approach issues with Cronbach's alpha with very little focus on mathematical or computational detail so that the deficiencies of Cronbach's alpha are illustrated in words and examples rather than proofs and simulations so that these ideas can impact a larger group of researchers—namely, the researchers who most often report Cronbach's alpha.

Keywords: Cronbach's alpha, internal consistency, reliability, psychometrics

In many areas of psychology and in the behavioral sciences more broadly, variables that are of interest (e.g., motivation, depression, cognitive abilities) are not directly observable and are therefore measured with scales or instruments comprised of a set of items. These items indirectly measure the variable of interest by inferring that some underlying construct manifests itself through these items. For example, an MRI study cannot directly measure the amount of extraversion present in a person's brain. Rather, items are created and administered to an individual. If the individual has high extraversion, this trait manifests itself through certain responses to the items.

Because most measurement in psychology is done through the use of indirect measurement tools, researchers often report a mea-

This article was published Online First May 29, 2017.

Daniel McNeish, Department of Methodology and Statistics, Utrecht University, and Center for Developmental Science, University of North Carolina, Chapel Hill.

I wrote the first version of this article while I was an assistant professor in the Department of Methodology and Statistics, Utrecht University. All subsequent revisions were completed at University of North Carolina, Chapel Hill. The information in this article has not been previously disseminated at a conference or electronically prior to acceptance for publication. I am indebted to Denis Dumas, Greg Hancock, Katherine Muenks, and Kathryn Wentzel for conversations that inspired the motivation for this article. I especially would like to thank Gjalt-Jorn Peters for his expertise and assistance with the R code included in this article.

Correspondence concerning this article should be addressed to Daniel McNeish, Center for Developmental Science, University of North Carolina, Chapel Hill, 100 East Franklin Street Suite 200, Chapel Hill, NC 27599. E-mail: dmcneish@email.unc.edu

sure of reliability to demonstrate that the items composing the measure are *reliable*, meaning that the scores based on the items are reasonably consistent, the responses to the scale are reproducible, and that responses are not simply comprised of random noise. Put another way, a reliability analysis provides evidence that the scale is consistently measuring the same thing (although, this is distinct from concluding that the scale is measuring the intended construct—that is a question of scale validity).

In psychology studies, the most commonly used reliability index, by a wide margin, is Cronbach's alpha. In a review of reliability reporting practices conducted by Hogan, Benjamin, and Brezinski (2000), about two thirds (66%) of studies reporting a reliability measure selected Cronbach's alpha. Of those reporting a type of reliability that requires only a single administration (e.g., not test–retest or interrater reliability), 87% (548 out 633) reported Cronbach's alpha (or the KR-20, which is a special case of alpha where all items are binary; Crocker & Algina, 2008). Indeed, Cronbach's alpha can be universally found in the pages of psychology journals in any subfield. As of October 2014, the seminal Cronbach (1951) article that first introduced Cronbach's alpha was the 64th most cited English language research article on Google Scholar in any field and, within psychology, is only surpassed by the article of Baron and Kenny (1986) on mediation and moderation and the seminal article of Bandura (1977) on self-efficacy (van Noorden, Maher, & Nuzzo, 2014). In the last 20 years, however, many methodological articles have appeared which question how Cronbach's alpha is applied (Bentler, 2007; Cortina, 1993; Crutzen, 2007; Crutzen & Peters, 2015; Dunn, Baguley, & Brunsden, 2014; Geldhof, Preacher, & Zyphur, 2014; Graham, 2006; Green & Hershberger, 2000; Green & Yang, 2009a, 2009b; Peters, 2014; Raykov, 1997a, 1997b, 1998, 2004; Raykov & Shrout, 2002; Revelle & Zinbarg, 2009; Schmitt, 1996; Sijtsma, 2009; Teo & Fan, 2013; Yang & Green, 2011; Zinbarg, Revelle, Yovel, & Li, 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006). These articles argue that the assumptions made by Cronbach's alpha are commonly violated in types of data and models with which psychological researchers work. These arguments have led to the development of alternative reliability measures whose assumptions are more in-line with psychological data (Hancock & Mueller, 2001; Jackson & Agunwamba, 1977; McDonald, 1970, 1999; Revelle, 1979). Software routines for calculating these measures are also available in R packages such as MBESS (Kelley, 2007), psych (Revelle, 2008), or the scaleStructure function in the userfriendlyscience package (Peters, 2014).

The articles to which we referred in the previous paragraphs are actually fairly well-known, even among nonmethodological researchers. For instance, based on Google Scholar citation counts, Sijtsma (2009) has over 800 citations, Zinbarg et al. (2005) over 450, Hancock and Mueller (2001) almost 400, Yang and Green (2011) over 125, and Dunn et al. (2014) over 100 as of October 2016. Although such seemingly high awareness of issues with Cronbach's alpha appears reassuring, it does not appear that there have been substantial changes in the use of Cronbach's alpha.

To provide evidence for this claim and to show the enduring status of Cronbach's alpha, we reviewed articles in three flagship APA journals from educational psychology (*Journal of Educational Psychology, JEP*), social psychology (*Journal of Personality and Social Psychology, JPSP*), and clinical psychology (*Journal of Abnormal Psychology, JAP*) from January 2014 until October 2016. We located studies through Google Scholar by searching for the string "reliability" within these journals. This resulted in 369 total studies (131 from *JEP*, 118 from *JPSP*, and 120 from *JAP*). We filtered out studies that reported types of reliability that are not of interest to this article (e.g., interrater reliability), studies where "reliability" only appeared in the references, or where reliability was not used in in a psychometric sense. This netted 118 total studies (52 from *JEP*, 31 from *JPSP*, and 35 from *JAP*). Of these 118 studies, 109 (92%) solely used Cronbach's alpha to assess reliability of the scales used in their study while nine (8%) reported an alternative reliability measure either by itself or in addition to Cronbach's alpha. Despite the large number of citations of articles calling for alternative reliability measures, reliability reporting in these flagship APA journals (which have stringent methodological requirements) appears unchanged from the results reported in the Hogan et al. (2000) review. In fact, the aforementioned studies advising against Cronbach's alpha were nearly invisible in these APA journals. For example, none of the five aforementioned, highly cited articles which advocate for alternative measures were cited more than once each in the 118 reviewed articles.

This evidence suggests that researchers continue to almost exclusively rely on Cronbach's alpha as a measure of scale reliability. The pattern that methodological studies are well-cited but do not appear in flagship journals may suggest that researchers are aware of the issues with Cronbach's alpha but are reluctant to adopt new methods because these methods are not as widely known or accepted, that reviewers may not be familiar with the alternative methods, that the editorial process does not require more rigorous methods so researchers do not invest time to learn them, or that researchers are unsure how to obtain estimates of alternative measures for their data because many are not offered as popular general software packages like SPSS, SAS, or Stata. This also suggests that the more rigorous methodological work advising against Cronbach's alpha has not impacted psychologists as much as it has psychometricians or statisticians working in psychological domains. Sijtsma (2009) aptly summarizes this by stating

> while much of Cronbach's article was and still is accessible to many psychologists, the work by Lord, Novick, and Lewis and many others since may have gone unnoticed by most psychologists. This is truly an example of the gap that has grown between psychometrics and psychology and that prevents new and interesting psychometric results. (p. 115)

Though it appears promising that methodological articles are highly cited, there is limited evidence that the findings, conclusions, and recommendations are being incorporated in empirical studies. This may be taken to suggest that these studies are either being misinterpreted or not being read in their entirety, possibly because many appear in journals that are aimed at methodologists and statisticians and therefore may be written at too technical a level for empirical researchers with less quantitative training to fully benefit from the arguments being presented. Consistent with recent recommendations from Sharpe (2013) concerning bridging innovations in the use of statistical methods in psychology to empirical researchers, the aim of this tutorial article is to state as plainly and succinctly as possible why Cronbach's alpha is often inappropriate in empirical contexts and why researchers would benefit from abandoning Cronbach's alpha in favor of alternative

measures. Though there are many resources for readers capable of following mathematically based arguments, far fewer resources exist for the large number of psychological researchers operating below such a level of mathematical sophistication. As such, the scope of this article is intended to be very broad to elucidate the general idea that widespread adoption and continued use of Cronbach's alpha is detrimental. We heavily cite previous work in this area that can provide additional technical or nuanced detail on the issues discussed herein.

To outline this article, we first discuss the basics behind Cronbach's alpha including the restrictive assumptions that often obviate its use. We then overview some of the more conceptually clear, leading alternatives that can be employed to yield better estimates of reliability than Cronbach's alpha. This is followed by a brief comparison of scenarios in which these alternatives have specific advantages and disadvantages. Rather than lay out mathematical or logical arguments for why Cronbach's alpha should not be used as has been the primary method of previous articles on the topic, we demonstrate some of the issues with Cronbach's alpha using example analyses from publicly available data sets. We end with a discussion of why prolonged use of Cronbach's alpha is detrimental and how alternative measures are better suited to accomplish the same goal, often to researchers' benefit. We provide a heavily annotated software appendix to help readers employ these methods in their own research so that they can abandon Cronbach's alpha in favor of better alternatives.

Basics of Reliability

From a theoretical standpoint, some observed score X for a trait or construct is considered to have two latent components: the true component T and an error component E such that $X = T + E$. From a classical test theory perspective (Novick & Lewis, 1967), reliability is considered to be greater when the variance of the true score component accounts for a higher proportion of variance in the observed scores relative to the variance attributable to the error component. More formally, reliability is defined by the ratio of the true score variance to the observed score variance, $\rho_{XX'} = [Var(T)/Var(X)]$. Under this more formal definition, reliability can also be interpreted as the correlation between scores on two consecutive administrations, assuming the respondent does not recall their answers from the first administration (hence, the choice of $\rho_{XX'}$ as the symbol for reliability).

Although the definition of reliability is relatively straightforward, obtaining an estimate of reliability is not always so easy. Historically, many methods for assessing reliability (parallel forms, test–retest, test–retest with parallel forms; Crocker & Algina, 2008) required multiple test administrations which were then correlated to form an estimate of reliability. Due to logistical issues of multiple administrations, the ability to calculate reliability from a single test administration was highly desirable. Cronbach (1951) addressed this in his seminal article on *internal consistency reliability*, the type of reliability on which this article focuses. Rather than inspecting the correlation between separate administrations, internal consistency reliability inspects the relation of each item to all other items from a single administration. If respondents provide similar answers to a set of items, then their responses would reasonably generalize to other items from a similar domain, and the set of items would be considered to have high internal consistency reliability. (Crocker & Algina, 2008).

Cronbach's Alpha

Cronbach's alpha (Cronbach, 1951) is by far the most common measure of internal consistency reliability.[1] Cronbach's alpha is calculated by

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_X^2}\right) \quad (1)$$

where k is the number of items, s_i^2 is the variance of individual item i where $i = 1, \ldots, k$, and s_X^2 is the variance for all items on the scale. This formula is often reported in reduced form as $\alpha = (k^2\bar{s}_{ij})/s_X^2$ where $\bar{s}_{ij}$ is the mean covariance between all pairs of items on the scale (Geldhof et al., 2014). One can interpret the value of Cronbach's alpha in one of many different ways:

1. Cronbach's alpha is the correlation of the scale of interest with another scale of the same length that intends to measure the same construct, with different items, taken from the same hypothetical pool of items (Kline, 1986).

2. The square root of Cronbach's alpha is an estimate of the correlation between observed scores and true scores (Nunnally & Bernstein, 1994).

3. Cronbach's alpha is the proportion of the variance of the scale that can be attributed to a common source (DeVellis, 1991).

4. Cronbach's alpha is the average of all possible split-half reliabilities from the set of items (Pedhazur & Schmelkin, 1991).

Under certain assumptions, Cronbach's alpha is a consistent estimate of the population internal consistency; however, these assumptions are quite rigid and are precisely why methodologists have argued against the use of Cronbach's alpha (Gignac, Bates, & Jang, 2007; Graham, 2006; Novick & Lewis, 1967; Revelle & Zinbarg, 2009; Yang & Green, 2011). The assumptions of Cronbach's alpha are:

1. The scale adheres to tau equivalence.

2. Scale items are on a continuous scale and normally distributed.

3. The errors of the items do not covary.

[1] Readers should note that there are several criticisms of Cronbach's alpha about the degree to which it truly measures internal consistency (e.g., Revelle & Zinbarg, 2009; Sijtsma, 2009). These arguments can become rather abstract and theoretical so, given the intent of this article, we will not delve into the specifics and we will use "internal consistency" as a simplification of what Cronbach's alpha intends to measure. Do note, however, that Cronbach's alpha being a true measure of internal consistency has been called into question on multiple occasions.

4. The scale is unidimensional.

These assumptions have been stated in other locations (e.g., Green & Yang, 2009a; Yang & Green, 2011) and demonstrated mathematically (e.g., Bentler, 2009; Sijtsma, 2009) but their importance (and rigidity) may not necessarily be understood or appreciated in empirical work. The following subsections will expound these assumptions.

Assumption 1: Tau equivalence. *Tau equivalence* is the statistically precise way to state that that each item on a scale contributes *equally* to the total scale score. To put this assumption into perspective, imagine that an exploratory factor analysis is run on the scale and a single factor is extracted (as a researcher would desire). For the tau equivalence assumption to be upheld, the standardized factor loadings for each item would need to be nearly *identical* to all other items on the scale. Figure 1 below shows what hypothetical SPSS output would look like for a five-item scale that does meet tau equivalence (left panel) and a scale that does not meet tau equivalence (right panel).

Tau equivalence tends to be unlikely for most scales that are used in empirical research—some items strongly relate to the construct while some are more weakly related. Furthermore, if a scale captures only a single construct, it is unlikely that all the items devised by researchers capture the construct to an equal degree (Cortina, 1993; Yang & Green, 2011). Put more technically, most scales are *congeneric* (Geldhof et al., 2014; Graham, 2006; Peterson & Kim, 2013) which means that the items measure the same construct, but they do so with different degrees of precision (Raykov, 1997a). Such disparities between the quality of the individual items does not mean that the weaker items necessarily need to be removed, but it does violate the assumptions made by Cronbach's alpha with the result being that Cronbach's alpha will be too low (Miller, 1995).

In the likely event that the assumption of tau equivalence is violated, Cronbach's alpha becomes a lower-bound estimate of internal consistency rather than a true estimate, provided that errors are reasonably uncorrelated (Graham, 2006; Sijtsma, 2009; Yang & Green, 2011). This results in Cronbach's alpha estimates that can vastly underestimate the actual value of reliability—even if just a single item on the scale is responsible for the violation of tau equivalence (Raykov, 1997b). A simulation by Green and Yang (2009a) found that Cronbach's alpha may underestimate the true reliability by as much as 20% when tau equivalence is violated (e.g., if the true reliability is 0.70, Cronbach's alpha would estimate reliability in the mid 0.50s). Furthermore, the degree of underestimation is greatest when scales have a fairly small number of items (e.g., less than 10), which is often the case in empirical psychological research (Graham, 2006).

Item	Std. Loading
Q1	0.711
Q2	0.714
Q3	0.716
Q4	0.709
Q5	0.721

Item	Std. Loading
Q1	0.806
Q2	0.790
Q3	0.725
Q4	0.578
Q5	0.523

Figure 1. Hypothetical SPSS exploratory factor analysis output for standardized factor loadings of a five-item scale that meets tau equivalence (left) and that does not meet tau equivalence (right).

Assumption 2: Continuous items with normal distributions. As noted in discussions of Equation 1, Cronbach's alpha is largely based on the observed covariances (or correlations) between items. In most software implementations of Cronbach's alpha (such as in SAS and SPSS), these item covariances are calculated using a Pearson covariance matrix (Gadermann, Guhn, & Zumbo, 2012). A well-known assumption of Pearson covariance matrices is that all variables are continuous in nature. Otherwise, the elements of the matrix can be substantially biased downward (i.e., the magnitudes will be closer to 0 than they should be; Flora & Curran, 2004). However, it is particularly common for psychological scales to contain items that are discrete (e.g., Likert or binary response scales), which violates this assumption. If discrete items are treated as continuous, the covariance estimates will be attenuated, which ultimately results in underestimation of Cronbach's alpha because the relations between items will appear smaller than they actually are.[2]

To accommodate items that are not on a continuous scale, the covariances between items can instead be estimated with a *polychoric covariance* (*or correlation*) *matrix* rather than with a Pearson covariance matrix. Polychoric covariance matrices assume that there is an underlying normal distribution to discrete responses. For instance, imagine a three-category Likert item whose response choices consist of agree, neutral, and disagree. A polychoric covariance matrix first assumes that these response choices map onto a normal distribution whereby there is no longer three distinct categories but a continuous range of "agreement." Then *thresholds* are estimated which can conceptually be thought of as cut-points on the continuous agreement scale that separate the response categories. So, respondents at the 40th percentile or below on the hypothetical agreement continuum may be considered in the "disagree" category, respondents between the 40th and 80th percentile on the hypothetical agreement continuum would correspond to the "neutral" category, and respondents above the 80th percentile would correspond to the "agree" category (the percentile cut-points are estimated and would change for each item). Provided that it is reasonable to assume that a normal distribution underlies the discrete options, the polychoric covariance estimates correct the attenuation that occurs when discrete items are treated as continuous (Carroll, 1961). Gadermann, Guhn, and Zumbo (2012) demonstrate how using a polychoric covariance matrix with Cronbach's alpha can addresses underestimation of reliability attributable to discrete items.

Another related and less commonly considered assumption is that both the true scores and the errors are normally distributed (e.g., van Zyl, Neudecker, & Nel, 2000; Zimmerman, Zumbo, & LaLonde, 1993). Studies investigating the effect of non-normal distributions on Cronbach's alpha have been mixed. Zimmerman et al. (1993) generally conclude that Cronbach's alpha is fairly

[2] Likert scales with many response options can often be treated as continuous without any adverse effects. The definition of how many response options constitutes "many" has been debated in the methodological literature. In latent variable models broadly, Rhemtulla, Broussard-Liard, and Savalei (2012) recommend five. In the specific context of Cronbach's alpha, Gadermann et al. (2012) recommended seven response options.

robust to deviation from normality. On the other hand, Sheng and Sheng (2012) reported that leptokurtic distributions lead to negative bias (i.e., reliability estimates are too low) while platykurtic distributions lead to positive bias (i.e., reliability estimates are too high). In the simulation in Sheng and Sheng (2012), these biases dissipated as sample size and the magnitude of the true reliability increased.

Assumption 3: Uncorrelated errors. Although frequently overlooked (Zumbo & Rupp, 2004), the assumption that errors are uncorrelated is also required when utilizing Cronbach's alpha. Correlated errors occur when sources other than the construct being measured cause item responses to be related to one another. Correlated errors between items may arise for a variety of reasons including the order of the items on the scale (Cronbach & Shavelson, 2004; Green & Hershberger, 2000), speeded tests (Rozeboom, 1966), transient responses where feelings or opinions may change over the course of the scale (Becker, 2000; Green, 2003), or unmodeled multidimensionality of a scale (Steinberg & Thissen, 1996). Unlike the tau equivalence assumption, the impact of correlated errors does not necessarily bias Cronbach's alpha estimates in a predictable direction, meaning that violations can lead to either overestimates or underestimates of reliability. When errors are correlated, the correlations are often positive which will result in Cronbach's alpha overestimating the reliability (Bentler, 2009; Green & Hershberger, 2000; Green & Yang, 2009b). When correlated errors are not accounted for in the calculation of reliability, Cronbach's alpha can be overestimated by as much as 20% (Gessaroli & Folske, 2002).

Some reasons for error covariances are innocuous while others are much more problematic. For instance, if error covariances are necessary because of item order effects, error covariances can be incorporated to yield appropriate estimates. On the other hand, if the error covariances are needed due to unmodeled dimensions in the scale, this eliminates nearly all support for using the scale (i.e., the assumption of unidimensionality is violated—this assumption is discussed next). Unfortunately, considerations for which of these mechanisms is responsible for the covariances is difficult to determine empirically. It is difficult to test whether error covariances are non-null because there are often not sufficient degrees of freedom to include many error covariances into the model. Possible solutions to such a violation are discussed in subsequent sections.

Assumption 4: Unidimensionality. Though Cronbach's alpha is sometimes thought to be a measure of unidimensionality because its colloquial definition is that it measures "how well items stick together," unidimensionality is an assumption that needs to be verified prior to calculating Cronbach's alpha rather than being the focus of what Cronbach's alpha measures (Cortina, 1993; Crutzen & Peters, 2015; Green, Lissitz, & Mulaik, 1977; Schmitt, 1996). Although the terminology is not universally accepted (cf., Sijtsma, 2009), Schmitt (1996) makes the distinction between unidimensionality and internal consistency. He defines internal consistency as the interrelatedness of a set of items while unidimensionality is the degree to which the items all measure the same underlying construct.

Green et al. (1977) note that internal consistency is necessary for unidimensionality but that internal consistency is not sufficient for demonstrating unidimensionality. That is, items that measure different things can still have a high degree of interrelatedness, so a large Cronbach's alpha value does not necessarily guarantee that the scale measures a single construct. As a result, violations of unidimensionality do not necessarily bias estimates of Cronbach's alpha. In the presence of a multidimensional scale, Cronbach's alpha may still estimate the interrelatedness of the items accurately and the interrelatedness of multidimensional items can in fact be quite high (Cortina, 1993; Schmitt, 1996; Sijtsma, 2009).

Many articles (e.g., Crutzen & Peters, 2015; Schmitt, 1996; Green & Yang, 2009a) recommend beginning any reliability analysis with an inspection of the factor structure of the scale, specifically examining whether a one-factor model fits well via inferential tests like the minimum fit function chi square statistic or via fit index values. Though vitally important to the interpretation of scales, a review by Crutzen and Peters (2015) found that only 2.4% of health psychology studies reported any information about the dimensionality of the scale beyond assessments of reliability. Many leading alternatives to Cronbach's alpha (discussed in detail in the next section), make explicit use of the factor analytic approach to reliability, facilitating the presentation of dimensionality and reliability side-by-side.

Alternatives to Cronbach's Alpha

There are many methods available to assess the reliability of scales. Hattie (1985) reviews about 30 such methods and there are undoubtedly many additional methods that have been developed in the 30+ years since this review was published. Our intention is not to update Hattie (1985) by providing a broad overview of all the possible alternatives to Cronbach's alpha that are available. Instead, we focus on three particular methods: omega coefficients, Coefficient *H*, and the greatest lower bound. These three alternatives are selected because (a) they have been shown to perform well in previous studies; (b) they do not make as strict assumptions as Cronbach's alpha; and (c) they are conceptually similar to Cronbach's alpha, so the idea of each should be relatively familiar if one understands Cronbach's alpha.

Omega and Composite Reliability

Composite reliability is conceptually related to Cronbach's alpha in that it assesses reliability via a ratio of the variability explained by items compared with the total variance of the entire scale (Bentler, 2007; Geldhof et al., 2014; Raykov, 1997a, 1997b, 1998). Omega (McDonald, 1970, 1999) is a commonly recommended measure of composite reliability that is available in multiple software programs. Omega is designed for congeneric scales, where the items vary in how strongly they are related to the construct being measured (i.e., in a factor analysis setting, the loadings would not be assumed to be equal). In other words, where tau equivalence is not assumed. Composite reliability is appropriate when the items from a scale are *unit-weighted* to form the total scale score but the scale itself in congeneric (Bentler, 2007; Geldhof et al., 2014). A unit-weighted scale means that the total score of the scale is calculated by adding up the raw scores (or reverse coded raw scores, if appropriate) of the individual items: Each item is weighted equally.

There are multiple variations of omega including omega hierarchical, omega total, and what we will refer to as "Revelle's omega total." Omega hierarchical is useful for scales that may not

be truly unidimensional and may contain additional minor dimensions (Zinbarg et al., 2006). Omega hierarchical attempts to parse out the variability attributable to subfactors and calculates reliability for a general factor that applies to all items. Although highly advantageous, omega hierarchical differs from Cronbach's alpha conceptually, so we will only provide a broad overview here (although we do recommend its use if researchers believe that the items in the scale are organized in hierarchical factors).

Omega total, on the other hand, assumes that the scale is unidimensional and estimates the reliability for the composite of items on the scale (which is conceptually similar to Cronbach's alpha). In the R software environment, two packages (MBESS and psych) calculate versions of omega total. However, they yield different results because MBESS uses a different specification which generally tends to be more conservative and yields estimates closer to Cronbach's alpha (Peters, 2014; Revelle, 2016; Revelle & Zinbarg, 2009). We overview the properties and formulas for each version of omega total in the next subsections. Though both versions are typically referred to as "omega total," we assign different names to each version help keep them distinct. We refer to the omega total value based on the psych R package specification as "Revelle's omega total." We use "omega total" to refer to the version calculated by the MBESS R package (and as presented in many other sources).

Omega total. Under the assumption that the construct variance is constrained to 1 and that there are no error covariances, omega total is calculated from factor analysis estimates such that

$$\omega_{Total} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \theta_{ii}} \tag{2}$$

where λ_i is the factor loading (not necessarily standardized) for the ith item on the scale, θ_{ii} is the error variance for the ith item, and k is the number of items on the scale. Omega total can only be calculated if the scale is first factor analyzed to obtain the factor loadings and error variances. This is necessary because tau equivalence is no longer assumed and the potentially differential contribution of each item to the scale must be assessed.

Although perhaps not immediately intuitive, Equation 2 is identical to the Cronbach's alpha formula in Equation 1 under the condition of tau equivalence (Geldhof et al., 2014). The condensed equation for Cronbach's alpha that appears under Equation 1 can alternatively be written as $\alpha = (k\sum_i \sum_j \sigma_{ij})/\sigma_X^2$ because $\bar{\theta}_{ij} = \frac{\sum\sum \sigma_{ij}}{k}$. From factor analysis path tracing rules, the model-implied covariance for a pair of items (with no error covariances) that load on the same factor is equal to the square of the loadings (times the factor variance which is assumed to be equal to 1). Under tau equivalence, all the loadings are equal, so the total true score variance is equal to the item covariance for a single pair of items, repeated k times. In both Equation 1 and Equation 2, this variance is divided by the total variance of the scale. The denominator in Equation 2 is the factor analysis representation of s_X^2 from Equation 1. As such, omega total is a more general version of Cronbach's alpha and actually subsumes Cronbach's alpha as a special case. More simply, if tau equivalence is met, omega total will yield the same result as Cronbach's alpha but omega total has the flexibility to accommodate congeneric scales, unlike Cronbach's alpha.

Similar to Cronbach's alpha, omega total overestimates reliability if errors have a positive covariance. The omega total formula in Equation 2 assumes that errors are uncorrelated, though it can be generalized to cases where this assumption is violated by altering the denominator term to account for error covariance such that,[3]

$$\omega_{TCov} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \theta_{ii} + 2\sum_{i=2}^{k}\sum_{j=1}^{i} \theta_{ij}} \tag{3}$$

If the residual covariances may be attributable to additional minor dimensions, then omega hierarchical will yield a more accurate estimate of the reliability of the scale (Zinbarg et al., 2006). Extensions of omega total are also available for cases where the factor variance is not assumed to be 1 (Raykov, 2004) and when the data contain multiple groups (Zinbarg, Revelle, & Yovel, 2007). These extensions, however, are outside the scope of this introduction and will not be discussed further.

Revelle's omega total. Though similar in name and idea, Revelle's omega total can yield quite different (and typically larger) estimates of reliability than omega total. This is due to a different, more sophisticated variance decomposition that is used. In Revelle's omega total, a factor model is estimated as with omega total. However, the solution is then transformed with a *Schimd-Leiman rotation* (Schmid & Leiman, 1957). Though we will not go into full detail regarding this rotation because it is rather technical and full detail is outside the scope of this article (for full details, see Mansolf & Reise, 2016 or Wolff & Preising, 2005), the general idea is to rotate the factor solution to a bifactor model where there is one general factor and several minor factors. More specifically, each item will load on the single general factor (g), one or more group factors (f), and an item-specific factor (s). The communality is then calculated by squaring the loadings of the general factor and the group factor(s) but not the item-specific factors (Revelle, 2016).

The formula for Revelle's omega total is essentially the same as Equation 2; however, it is more complex to account for the differential variance decomposition and additional minor factors. Namely, Revelle's omega is equal to

$$\omega_{RT} = \frac{\left(\sum_{i=1}^{k} \lambda_{gi}\right)^2 + \left(\sum_{f=1}^{F}\sum_{i=1}^{k_f} \lambda_{fi}\right)^2}{V_X} \tag{4}$$

where λ_{gi} is the loading of the ith item on the general factor, λ_{fi} is the standardized loading of the ith item on the fth group factor, k is the total number of items, F is the total number of group factors,

[3] Note that, although the inclusion of the error covariances in the denominator appropriately takes the extra source of variation into account, it does not solve the broader issue of *why* there is error covariance. That is, whether the error covariance is attributable to a model misspecification where an important factor has been omitted from the model (Green & Hershberger, 2000) or whether design-driven aspects of the scale led to the correlated errors (e.g., speeded tests; Cole, Ciesla, & Steiger, 2007). Bentler (2009) nicely summarizes this issue by stating "It would seem that the question of whether to consider correlated errors as factors and hence part of the common factor space, or as residual covariances and hence as part of the unique space, should be left up to the goals of the investigator" (p. 139).

and k_f is the number of items that load on the fth group factor. V_X is the total variance after rotation which is equal to the sum of each element of the sample Pearson (or polychoric) correlation matrix (in matrix notation, this can be succinctly written as $\mathbf{1}^{\mathrm{T}}\,\mathbf{R1}$ where $\mathbf{R}$ is the sample correlation matrix).

Omega hierarchical is based on the exact same Schmid-Leiman transformation except that it only considers contributions of the general factor and disregards the loadings of both the group factors in addition to the item-specific factors,

$$\omega_H = \frac{\left(\sum_{i=1}^{k} \lambda_{gi}\right)^2}{V_X} \tag{5}$$

For interested readers, Kelley and Pornprasertmanit (2016) provide a highly readable description of omega hierarchical and when it should be used. Readers looking for complete details on omega hierarchical are referred to Zinbarg et al. (2005).

Though the formulas may look intimidating, the idea is quite straightforward because software will handle the rotation and complexities of the formula. Explanations of how these values are extracted from the data are provided in the software appendix.

Coefficient *H* and Maximal Reliability

Should researchers want to use the information present from the factor loadings to create a scale that is *optimally weighted* where each item contributes different amounts of information to the overall scale score (instead of each item being given the same weight with unit-weighting), then *maximal reliability* is a more appropriate measure of the scale's reliability (Bentler, 2007; Hancock & Mueller, 2001; Raykov, 2004).[4] Hancock and Mueller (2001) derived Coefficient *H* as a measure of maximal reliability for an optimally weighted scale. Similar to the form of omega total presented in Equation 2, Coefficient *H* requires the (standardized) factor loadings from a unidimensional factor analysis of the scale (or from unidimensional subscales). Coefficient *H* is calculated by.

$$H = \left(1 + \left(\sum_{i=1}^{k} \frac{\ell_i^2}{1 - \ell_i^2}\right)^{-1}\right)^{-1} \tag{6}$$

where k is again the number of items on the scale and ℓ_i is the standardized factor loadings for the ith item. Unlike Equation 2, notice that the squaring of the factor loadings occurs prior to summing over the each of the items. Both Cronbach's alpha and omega (all versions) are adversely affected by items with negative loadings, whereas Coefficient *H* squares the loadings first so that magnitude (and not sign) is the only important feature. This means that negatively worded items do not need to be reverse coded with Coefficient *H*.

There are several other features of Coefficient *H* that differentiate it from omega total. First, error variances are not included in the denominator of the equation. This means that items with weak factor loadings do not negatively affect Coefficient *H* as they do in the computation of omega total. In Equation 2, an item with a weak loading will necessarily have a large error variance (i.e., the underlying construct accounts for a small percentage of the variance, so the remaining variance must be attributable to error). In Coefficient *H*, the scale is not penalized for featuring weaker items because its intended use is for optimally weighted scales. For example, whereas adding an item completely unrelated to the construct of interest to a scale reduces reliability for Cronbach's alpha and omega (which are appropriate for unit-weighted scales), with optimal-weighted scales, an unrelated item's factor loading will essentially be 0 and the information from this item would not affect the scale scores. Put another way, in unit-weighted scales, every item receives equal treatment so an unrelated item hurts the scale; in optimally weighted scales, items are differentially weighted so an unrelated item does not hurt reliability because the item simply receives very little or zero consideration when scoring the scale. Another property exclusive to Coefficient *H* is that the reliability of the scale cannot be less than the squared loading (the definition of reliability in factor analytic models) of the single best item (Geldhof et al., 2014).

Greatest Lower Bound

The greatest lower bound (GLB) is a class of methods for assessing reliability which are all based on the same conceptual idea. First introduced by Jackson and Agunwamba (1977), the GLB is based on the classical test theory approach to reliability. First, the GLB extends the classical test theory formula from $X = T + E$ to $Cov(\mathbf{X}) = Cov(\mathbf{T}) + Cov(\mathbf{E})$—the covariance matrix of all observed scores $\mathbf{X}$ is equal to the covariance matrix of all true scores $\mathbf{T}$ plus the covariance matrix of all the errors $\mathbf{E}$ (Shapiro & ten Berge, 2000; ten Berge & Sočan, 2004). Conceptually, Jackson and Agunwamba (1977) argued that the greatest lower bound for reliability could be calculated from the estimate of the covariance matrix of $\mathbf{E}$ with the largest trace that is consistent with the data (provided that $Cov(\mathbf{T})$ and $Cov(\mathbf{E})$ are non-negative definite).[5] Once the estimated covariance matrix for $\mathbf{E}$ with the largest trace is found, GLB reliability is calculated by

$$GLB = 1 - \frac{trace[Cov(\mathbf{E})]}{s_X^2} \tag{7}$$

where s_x^2 is the variance of the observed items. More simply, the goal is to determine the maximal values for the error component of the observed scores that is consistent with the data because reliability calculated with these maximum errors will yield the lowest possible value for reliability (Sočan, 2000). Jackson and Agunwamba (1977) showed that Cronbach's alpha and other single administration measures like split-half reliability can be shown to be based on the same principle as the GLB with the exception that they inefficiently estimate $Cov(\mathbf{E})$ and therefore do not exceed the theoretical GLB value.

Though appealing theoretically, a major challenge for GLB reliability is its computation. The difficulty stems from finding the estimate of $Cov(\mathbf{E})$ that maximizes the trace. In fact, a simple analytical solution is generally impossible, so several iterative methods have been proposed to determine this matrix with leading

[4] When using optimal weighting, the contribution of each item to the scale score is based on the magnitude of its standardized factor loading. For example, an item with a standardized loading of 0.90 would have a much larger impact on the scale score than an item with a standardized loading of 0.50.

[5] The trace of a matrix is computed by adding up all of the diagonal elements and non-negative definite means that the diagonal elements of the matrix are 0 or larger.

candidates being the minimum rank factor analysis (MRFA) approach of ten Berge and Kiers (1991) and the GLB algebraic solution from Moltner and Revelle (2015; both of which can be implemented in R). An additional limitation of GLB reliability is that it tends to overestimate reliability with smaller sample sizes (e.g., bias is rather large with a sample size of 100 but is reasonable with a sample size of 500; Shapiro & ten Berge, 2000; Trizano-Hermosilla & Alvarado, 2016).

Practical Comparison of Methods

Table 1 compares the six aforementioned methods (Cronbach's alpha, omega total, Revelle's omega total, omega hierarchical, Coefficient *H*, and the GLB) based on practical considerations. That is, because adopting new statistical approaches often entails a steep learning curve, Table 1 does not compare strict statistical properties or asymptotic behavior but rather overviews which software can compute each method, whether the method is calculable by hand, notable conceptual advantages, and notable conceptual disadvantages. Alternatives to Cronbach's alpha tend to have very little support in general software, so the easiest measures to report are omega total or Coefficient *H* because they can be calculated using a simple spreadsheet. More computationally intensive measures are only currently

Table 1
Comparison of Practical Considerations for Six Different Methods

Measure	Ease of implementation in general statistical software	Notable advantages	Notable disadvantages
Cronbach's alpha	Ubiquitous in general software (e.g., SPSS, SAS, Stata, R).	Familiar to readers and reviewers.	Underestimates reliability, requires tau equivalence.
Omega total	Available in the `MBESS` R package via the `ci.reliability` function or via the `scaleStructure` function in the `userfriendlyscience` package. Also calculable with a spreadsheet (provided in the Appendix). No built-in option for computing a polychoric covariance matrix, though factor analysis procedures do, which does not affect manual ease of calculation.	Most conceptually related to Cronbach's alpha (Cronbach's alpha is a special case). Formula can be extended to take design-driven error covariances into account.	Tends to be yield conservative estimates compared with other alternative methods.
Revelle's omega total	Available in the `psych` R package via `omega` function or via the `scaleStructure` function in the `userfriendlyscience` package. Not calculable manually.	Tends to justifiably exceed Omega total and often exceeds the GLB.	Proportionality assumption of Schmid-Leiman must be met.
Omega hierarchical	Available in the `psych` R package via `omega` function or via the `scaleStructure` function in the `userfriendlyscience` package. Not calculable manually. Includes a built-in option for internally computing and using a polychoric covariance matrix.	Accounts for and excludes effects of minor dimensions.	Most conceptually distant from traditional Cronbach's alpha. Also dependent on Schmid-Leiman assumptions.
GLB	Available in the `psych` R package via `glb.fa` or `glb.algebraic` function. Available via the `scaleStructure` function in the `userfriendlyscience` package. Not calculable manually. No built-in option for computing polychoric covariance matrix.	Exceeds Cronbach's alpha, even if all assumptions are met.	No analytic solution, current software does not offer polychoric option.
Coefficient *H*	Very simple to calculate in a spreadsheet (provided in in the Appendix), calculated by default in the `scaleStructure` function in the `userfriendlyscience` package for continuous items.	Designed for optimal-weighted scales, not affected by addition of poor items.	Misleading if the scale is scored with unit-weighting.

supported in R. We realize that R is not the first-choice software for many psychologists, so extensive annotated R code is provided in an appendix to assist in calculating measures that require more computational resources (e.g., Schmid-Leiman transformation, MRFA).

Empirical Examples

In this section, we provide example analyses to demonstrate the shortcomings of Cronbach's alpha. The first example dataset is based on a subsample of the Early Childhood Longitudinal Study—Kindergarten (ECLS-K) from the United States' National Center for Educational Statistics. The data include 21,054 students and thousands of variables such as direct cognitive assessments of students, teacher reports of students, parental reports of students, and detailed information about demographic information and students' home life at seven time-points. The data are publicly available from the United States' National Center for Educational Statistics (https://nces.ed.gov/ecls) and are intended to allow researchers to answer research questions pertaining to child development, school readiness, and experiences in schools. We used a subsample consisting of 1977 students who had complete math and reading scores at all seven waves of the study. Socioeconomic status is not captured by a single variable in ECLS-K, therefore researchers have argued and demonstrated that it is more fruitful to form a scale for socioeconomic status using variables that capture different aspects of socioeconomic status (Curran & Kellogg, 2016; Lubienski & Crane, 2010). In this example, we use nine variables: mother's education, father's education, household income (in dollars), parents' expectation of child's eventual education level, the number of books the child has, whether the child qualifies for free or reduced lunch, whether the parent volunteers at school, whether there is a computer in the house (these data were collected in the late 1990s when home computers were not ubiquitous), and whether the child is enrolled in music lessons. These variables were collected during the fall semester of the child's kindergarten year. The first example primarily demonstrates how the assumption of tau equivalence adversely affects Cronbach's alpha in ways that do not affect other measures. Differences between reliability for optimally weighted and unit-weighted scales are also shown.

The second example contains responses to 25 Likert items from the Big Five Inventory for personality traits. The data contain responses from 2,800 people and were collected as part of the Synthetic Aperture Personality Assessment (SAPA) project (Revelle, Wilt, & Rosenthal, 2010). The data are freely available in the psych R package as the "bfi" data. This example shows how the various measures are similar when tau equivalence is approximately met and how the measures diverge when scales are congeneric. The data in this example are based on Likert items, so the example also shows how reliability is attenuated if discrete responses are treated as continuous and how discrete items similarly affect alternatives measures as well.

Although we previously listed other assumptions earlier in the text, these examples primarily focus on violations of the tau equivalence and continuous item assumptions. This is intentional because theses assumptions of Cronbach's alpha are frequently violated and are the simplest assumptions to relax.

ECLS-K Example

To demonstrate the large violation of tau-equivalence in these data, we first perform a likelihood ratio test comparing a model with constrained standardized loadings across all items to a model with standardized loadings freely estimated for all items. We reverse coded the free or reduced lunch variable because its loading was negative, which would adversely affect fit. With all loadings constrained, $\chi^2(35) = 625.33$, SRMR = .12, McDonald Centrality = .83[6] and the standardized loading for all items was estimated to be 0.48. When loadings were allowed to be unconstrained, $\chi^2(27) = 160.52$, SRMR = .05, McDonald Centrality = .96. A likelihood ratio test of these two models results in a value of $\Delta\chi^2(8) = 464.81$ which is clearly significant (the 0.05 cut-off is 15.51) and indicates that the model with constrained loadings fits significantly worse. The standardized loadings for the unconstrained model are presented in Table 2, which clearly show a wide range of standardized factor loadings (Range: 0.21 to 0.76). The fit indices also provide evidence that the scale is unidimensional because a one factor solution fits the data reasonably well. Table 2 provides the reliability estimates using Cronbach's alpha, omega total, Revelle's omega total, the GLB, and Coefficient *H*. If Cronbach's alpha is used, the value is in the mid .70s which would result in the scale being seen as "acceptable" using common guidelines from Kline (1986) and DeVellis (1991). However, recall that the loadings in this example are highly discrepant and that this negatively biases Cronbach's alpha estimates. Using an alternative measure of reliability results in noticeable increases in reliability estimates, as high as 10% with Coefficient *H*.

Although many researchers would consider removing the music lessons variable due to its low loading, we have retained it to demonstrate the difference in reliability estimates for unit-weighted and optimally weighted scales. For Cronbach's alpha, both omega totals, and the GLB, a weakly related item decreases reliability because each item receives equal consideration when computing scale scores. However, optimally weighted scales (for which Coefficient *H* is appropriate) differentially weight each item based on its factor loading. As a result, Coefficient *H* in this case is higher (5% higher than Revelle's omega total) because the music lessons variable is heavily down-weighted and the other, more reliable items would be weighted much more heavily when scale scores are computed. As a reminder, even though it may be appealing to report Coefficient *H* in such a case because it is higher, it is only appropriate if the scale score is calculated using optimal weights.

[6] Hu and Bentler (1999) recommend McDonald's Centrality > .90 and SRMR < .09 as a combinational rule that minimizes the sum of Type-I and Type-II errors (p. 26) while McDonald's Centrality > .93 and SRMR < .06 also worked fairly well but tended to overreject true models. We use this criteria to establish goodness-of-fit throughout these examples because factor models for scales with few items tend to have few degrees of freedom, for which RMSEA vastly overrejects well-fitting models (Kenny, Kaniskan, & McCoach, 2015) and because the sample size in both models is rather large, which may render the chi-square test overpowered (e.g., Hu & Bentler, 1998). Note that there has been a steady wave of criticism against generalizing the Hu and Bentler cut-offs (e.g., Hancock & Mueller, 2011; Marsh, Hau, & Wen, 2004) although our examples fall fairly closely to their original simulation design (factor model with five items per factor and standardized loadings near 0.70).

Table 2
ECLS-K Example Standardized Factor Loadings, Estimated Reliability Using Different Methods, and Model Fit Indices

Variable	Std. loading	Measure	Estimate	% Increase
FR lunch	−.52	Cronbach's alpha	.74	—
Mom education	.73	Omega total	.75	1.4%
Dad education	.76	Revelle's omega total	.77	4.1%
Household income	.60	Greatest lower bound	.80	8.1%
Expect education	.35	Coefficient *H*	.81	9.5%
Number of books	.40			
Music lessons	.21	Fit		
Computer at home	.44	SRMR	.05	
Parent volunteers	.39	McDonald's Centrality	.96	

Note. SRMR = standardized root mean squared residual; % Increase = the percent relative increase of reliability compared with Cronbach's alpha. The free or reduced lunch variable was reverse coded when calculating Cronbach's alpha and both Omega totals so that all covariances would be positive.

Big Five Inventory Example

Unlike the previous example where tau equivalence was badly violated, this example features five subscales with various gradations of (possible) violations to tau equivalence. Table 3 shows the standardized factor loadings based on the Pearson and polychoric correlation matrices. Both sets of results were obtained in R using the psych package and the scaleStructure wrapper from the userfriendlyscience package (details are provided in the appendix). Each subscale in this dataset contains five items that are intended to be unidimensional (i.e., each item only measures a single construct). To assess the unidimensionality of these subscales, SRMR and McDonald's Centrality are provided for each subscale; the values for each subscale meet the suggested guidelines and we continue under the assumption that unidimensionality for each subscale is preserved.

Upon initial inspection of Table 3, the various subscales adhere to tau equivalence to varying degrees. The loadings for the *conscientiousness* subscale are rather close to one another (magnitude range: 0.55 to 0.67 using a Pearson covariance matrix, 0.58 to 0.72 using a polychoric covariance matrix). On the other hand, the loadings for the *agreeableness* subscale are quite variable (Range: 0.37 to 0.76 using a Pearson covariance matrix, 0.43 to 0.80 using a polychoric covariance matrix). To more rigorously demonstrate the similarity of the loadings on the *conscientiousness* subscale, we constrained the standardized loadings to be equal and compared the fit to a model where all loadings are freely estimated. The likelihood ratio test was significant $\chi^2(4) = 28.17$, $p < .01$ but the changes in the SRMR (ΔSRMR = .0125) and McDonald's Centrality (ΔMcDonald = −.0048) were rather small.[7] We proceed by allowing the loadings to be freely estimated, but we treat the *conscientiousness* subscale as an exemplar of the behavior of the various reliability measures when tau equivalence is roughly appropriate.

Table 4 shows the estimated reliability using Cronbach's alpha, omega total, Revelle's omega total, the GLB (using the MRFA approach), and Coefficient *H* using both a Pearson covariance matrix and a polychoric covariance matrix. First, notice that when the subscale is very closely tau equivalent (as in the *conscientiousness* subscale), there are small differences between the various reliability measures.[8] However, the difference between the estimates grows larger the as the subscales deviate from tau equivalence with relative percentage increases over Cronbach's alpha ranging from 5% to 12% across subscales.

This example also shows the effect of treating truly discrete items as continuous when calculating reliability, which is an assumption of all methods because each use the interitem covariance matrix in some form in their calculation. Even though item responses are on a 6-point Likert scale, the reliability estimates using the polychoric covariance matrix are noticeably larger because treating the items as continuous attenuates the covariances. Across each subscale, the estimates based on the polychoric covariance matrix are between .02 to .11 points higher for the same measure than if the Pearson covariance matrix is used. Regardless of which method is used to calculate reliability, when assessing reliability, it is important to consider the scale of the responses.

Among the various alternatives to Cronbach's alpha, the expected trends can be seen in this example. First, Cronbach's alpha consistently yields the lowest estimate of reliability. This is expected because Cronbach's alpha is the only method making the tau equivalence assumption which is rarely tenable and inappropriate for at least four of the five subscales in this example. Second, when subscales have an item that has a noticeably poor item relative to the other items (e.g., Item 1 on *agreeableness*, Item 4 on *openness*), Coefficient *H* tends to provide larger reliability estimates than omega total, the GLB, and sometimes than Revelle's omega total because the scale would be better scored using optimal weighting (to down-weight the impact of the poor item). When subscales have factor loadings in the same general vicinity (but not necessarily close enough to be considered approximately tau equivalent), the GLB and Revelle's omega total yield higher estimates than Coefficient *H*. In the case of approximate tau

[7] When sample size is large, some studies have recommended using change in fit indices instead of likelihood ratio test (e.g., Cheung & Rensvold, 2002; Chen, 2007). Although the field has not uniformly accepted this approach (e.g., Barrett, 2007), these changes in fit indices between models are below the recommend cut-offs (less than .025 for SRMR when testing loadings, greater than −.005 for McDonald's (1999) Centrality; Chen, 2007).

[8] When a scale is perfectly tau equivalent, omega total and Coefficient *H* will be identical to Cronbach's alpha, provided that all other assumptions are met. With tau equivalence, there is no difference between unit weighting and optimal weighting because, with optimal weighting and tau equivalence, each item receives the same weight. The GLB will not necessarily be equal to Cronbach's alpha, even if a scale is tau equivalent (Sočan, 2000).

Table 3
Standardized Factor Loadings for Big Five Example, Treating the Items as Continuous With a Pearson Covariance Matrix and Discrete With a Polychoric Covariance Matrix

Subscale	Item 1	Item 2	Item 3	Item 4	Item 5	SRMR	MC
Pearson covariance matrix							
Agreeableness	−.37	.66	.76	.48	.63	.04	.98
Conscientiousness	.55	.61	.55	−.67	−.59	.05	.97
Extraversion	−.61	−.73	.58	.69	.52	.04	.99
Neuroticism	.82	.80	.72	.55	.50	.07	.93
Openness	.55	−.44	.65	.30	−.51	.04	.99
Polychoric covariance matrix							
Agreeableness	−.43	.71	.80	.52	.67	.04	.99
Conscientiousness	.59	.64	.58	−.72	−.62	.06	.97
Extraversion	−.64	−.77	.60	.74	.54	.04	.99
Neuroticism	.86	.84	.74	.57	.52	.08	.93
Openness	.60	−.48	.69	.37	−.58	.05	.99

Note. SRMR = standardized root mean squared residual; MC = McDonald Centrality.

equivalence, Coefficient *H* converges to Cronbach's alpha whereas the GLB is known to exceed Cronbach's alpha in such instances (e.g., Sočan, 2000). When there is moderate separation between the loadings of the various items (as on the *neuroticism* subscale), Coefficient *H* and the GLB are approximately equal.

Take-Home Message

The take-home message of these examples is that there is a vast discrepancy in the reliability estimates when applying the conventional Cronbach's alpha compared to employing alternative methods. In the Big Five Inventory example, Cronbach's alpha for the *openness* subscale using a Pearson covariance matrix is .61 which would be classified as borderline poor (DeVellis, 1991 and Kline, 1986 designate the "poor" classification at <.60) and would likely need to be defended if an article were submitted for publication. However, by appropriately accounting for the discreteness of the responses and using a method that does not mandate tau equivalence, Revelle's omega total, the GLB, and Coefficient *H* estimate the reliability to be well above .70. The GLB yields the highest estimate at .76, 25% higher than the Cronbach's alpha estimate based on the Pearson covariance matrix.

Discussion

Although Cronbach's alpha is familiar, commonly reported, and easy to obtain in software, it is rarely an appropriate measure of reliability—its assumptions are overly rigid and almost always violated. Worse yet, under the near ubiquitous violation of tau equivalence, Cronbach's alpha estimates make scales appear much less reliable than they are in actuality. Moreover, even if all assumptions are met, Cronbach's alpha is a special case of the alternative measures overviewed in this article meaning that, even if Cronbach's alpha is appropriate, other methods will yield the exact same values and others (Revelle's omega total and the GLB) have been shown to routinely exceed Cronbach's alpha. Quite plainly, there is no situation where Cronbach's alpha is the optimal method for assessing reliability.

Table 4
Comparison of Subscale Reliabilities for Model in Big Five Inventory Example Using Cronbach's Alpha, Both Versions of Omega Total, the GLB, and Coefficient H

Subscale	Cronbach's alpha	Omega total	Omega Revelle	Greatest lower bound	Coefficient *H*
Pearson covariance matrix					
Agreeableness	.71	.71	.77	.75	.77
Conscientiousness	.73	.73	.77	.77	.74
Extraversion	.76	.77	.80	.82	.78
Neuroticism	.81	.82	.88	.85	.85
Openness	.61	.62	.68	.65	.65
Polychoric covariance matrix					
Agreeableness	.76	.77	.83	.79	.81
Conscientiousness	.77	.77	.81	.81	.78
Extraversion	.79	.80	.83	.84	.81
Neuroticism	.84	.84	.90	.87	.88
Openness	.67	.68	.73	.76	.71

Note. Omega Revelle = Revelle's omega total from psych R package. Items with negative loadings were recoded when calculating Cronbach's alpha and both omega totals so that all covariances would be positive.

Despite a steady stream of criticism against Cronbach's alpha, researchers continue to report it in flagship APA journals, as reviewed in the introduction. A common tactic when reporting unfavorable values of Cronbach's alpha is to appeal to the weakness of the method. This approach, while well-intended, is highly problematic for the scientific process because it impedes the ability to identify scales with less desirable properties. That is, if a scale has a Cronbach's alpha value of 0.40, the value could be low because (a) the scale is not reliable or (b) the scale is sufficiently reliable but assumption violations led to downwardly biased estimates of Cronbach's alpha. This uncertainty leads toward a dichotomy where either (a) the use of the scale is supported because reliability is sufficiently high (e.g., 0.70 or greater) or (b) Cronbach's alpha should be higher but was underestimated because assumptions were violated and the scale is still usable. Such a dichotomy hides a third option which is simply that the scale is not reliable. In the long run, it does the field little good to use faulty methods whose results may subsequently be disregarded; the process of scale validation at such a point becomes highly subjective and not readily falsifiable, eroding the credibility of psychometric analysis.

Given that many psychologists employ latent variable methods (item response theory, confirmatory factor analysis, or exploratory factor analysis) to explore their scales rather than classical test theory, it is difficult to excuse the continued use of Cronbach's alpha. Specifically, the vital assumption of tau equivalence is quite easy to inspect by examining the similarity of the factor loadings. Even the classic eyeball test can be an effective approximation in many cases. For instance, in the ECLS-K example, formal tests are not likely necessary to determine that standardized loadings of 0.21 and 0.76 are not approximately equal. If the factor loadings are not equivalent for all items on the scale, then Cronbach's alpha is not appropriate and its use will adversely affect results by making reliability appear lower than it actually is. Other measures are susceptible to other assumption violations, but we remind readers that there are ways in which these could be addressed such as omega hierarchical for the presence of minor dimensions, including error covariances between items for design-driven reasons, or basing estimates on a polychoric rather than Pearson covariance matrix if item responses are discrete rather than continuous. We would like to note that Likert items, even with many categories, attenuate the item covariances that are used in all methods we discuss in this article, which results is downwardly biased estimates of reliability. Therefore, it tends to be in researchers' best interest to acknowledge potential discreteness of items.

Although there have been previous calls to abandon Cronbach's alpha, Revelle and Zinbarg (2009) noted that software for other methods was somewhat limited and that empirical researchers may be hesitant because of the undoubted attraction to methods that have simple software applications. Although the GLB and Revelle's omega total are best estimated in R because of some computational complexities, omega total and Coefficient *H* are fairly straightforward to compute manually or with spreadsheets and do not require sophisticated or iterative processes. In the Appendix, we provide annotated R code that can be used to estimate these alternative measures. Some of the functionality included in these packages may require additional analyses in R, which we realize may not be helpful to users who are unfamiliar with or who dislike using R (though the scaleStructure function can eliminate the need for these additional analyses for most of the alternative measures). In an attempt to make these measures more accessible, we provide an Excel spreadsheet on the first author's personal web site and on the *Open Science Framework* that allows researchers to compute Coefficient *H* and omega total using only the standardized factor loadings. Guidance for using this spreadsheet is also provided in the Appendix.

This article is not intended to fully cover all the nuances and issues associated with Cronbach's alpha or calculating and reporting scale reliability as this literature is rather extensive. Other researchers have provided more technical information on this topic for those seeking a deeper understanding of the issues surrounding reliability. Geldhof et al. (2014) provide further guidance on calculating reliability with Cronbach's alpha, omega total, and Coefficient *H* when data come from a multilevel structure. Kelley and colleagues have several recent articles discussing the importance of confidence intervals around reliability estimates and discuss how to compute such intervals for many measures which have been included in their MBESS R package (e.g., Kelley & Cheng, 2012; Kelley & Pornprasertmanit, 2016; Terry & Kelley, 2012). Zhang and Yuan (2016) discuss robust methods to compute Cronbach's alpha and omega total with non-normal or missing data and also provide the R package coefficientalpha. We presented only a few of the possible alternatives to Cronbach's alpha. Bentler's rho (Bentler, 1968) has also been recommended and is easy to compute in the EQS software while Sijtsma (2009) has vouched for the explained common variance (ECV) method. We focused on unidimensional scales, although there is a growing trend in the literature to assess the reliability of multidimensional scales. Bifactor and hierarchical models (where there is a single general factor and several subscale factors) are more appropriate for these types of scales and there are alternative measures (Reise, 2012; Reise, Bonifay, & Haviland, 2013; Reise, Morizot, & Hays, 2007).

In conclusion, we hope that we have sufficiently demonstrated why Cronbach's alpha is obsolete and that it is time for the field to move on to better, more general alternatives. As seen in the empirical examples, the practical differences among the competing alternatives tends to be rather small—the example showed that the GLB, Revelle's omega total, and Coefficient *H* tend to provide the highest estimates of reliability. We realize that readers may be hoping for guidance on which of the aforementioned methods should be the "successor" to Cronbach's alpha.[9] Although some of these comparisons have been noted in the literature and some general relations are known (such as those presented in Table 1), these results should not be taken as rigorous and comprehensive since they are anecdotal and not based on analytic derivations or simulation results (though such comparisons would undoubtedly be a fruitful avenue of future research).The common theme we hope to espouse is that Cronbach's alpha is outperformed by *all* of these methods. We believe that the most important message empirical researchers receive from this article is that using *any* of the alternatives is preferable to continued use of Cronbach's alpha. Cronbach's alpha had a good run and was able to hold down the fort for the field for over 50 years, but methodological reinforcements have indeed arrived.

[9] This phrase was used by a reviewer, which we adopted because we thought it very aptly described the current state of affairs.

References

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84,* 191–215. http://dx.doi.org/10.1037/0033-295X.84.2.191

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182. http://dx.doi.org/10.1037/0022-3514.51.6.1173

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42,* 815–824. http://dx.doi.org/10.1016/j.paid.2006.09.018

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5,* 370–379. http://dx.doi.org/10.1037/1082-989X.5.3.370

Bentler, P. M. (1968). Alpha-maximized factor analysis (alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika, 33,* 335–345. http://dx.doi.org/10.1007/BF02289328

Bentler, P. M. (2007). Covariance structure models for maximal reliability of unit-weighted composites. In S. Lee (Ed.), *Handbook of computing and statistics with applications:* Vol. 1. (pp. 1–19). New York, NY: Elsevier.

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

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347–372. http://dx.doi.org/10.1007/BF02289768

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464–504. http://dx.doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods, 12,* 381–398. http://dx.doi.org/10.1037/1082-989X.12.4.381

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104. http://dx.doi.org/10.1037/0021-9010.78.1.98

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. New York, NY: Holt.

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

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

Crutzen, R. (2007). Time is a jailer: What do alpha and its alternatives tell us about reliability? *European Health Psychologist, 16,* 70–74.

Crutzen, R., & Peters, G. J. Y. (2015). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*. Advance online publication. http://dx.doi.org/10.1080/17437199.2015.1124240

Curran, F. C., & Kellogg, A. T. (2016). Understanding science achievement gaps by race/ethnicity and gender in kindergarten and first grade. *Educational Researcher, 45,* 273–282. http://dx.doi.org/10.3102/0013189X16656611

DeVellis, R. F. (1991). *Scale development*. London, UK: Sage.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399–412. http://dx.doi.org/10.1111/bjop.12046

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491. http://dx.doi.org/10.1037/1082-989X.9.4.466

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,* 1–13.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19,* 72–91. http://dx.doi.org/10.1037/a0032138

Gessaroli, M. E., & Folske, J. C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing, 2,* 277–295. http://dx.doi.org/10.1080/15305058.2002.9669496

Gignac, G. E., Bates, T. C., & Jang, K. (2007). Implications relevant to CFA model misfit, reliability, and the five factor model as measured by the NEO–FFI. *Personality and Individual Differences, 43,* 1051–1062. http://dx.doi.org/10.1016/j.paid.2007.02.024

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement, 66,* 930–944. http://dx.doi.org/10.1177/0013164406288165

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8,* 88–101. http://dx.doi.org/10.1037/1082-989X.8.1.88

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7,* 251–270. http://dx.doi.org/10.1207/S15328007SEM0702_6

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37,* 827–838. http://dx.doi.org/10.1177/001316447703700403

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

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74,* 155–167. http://dx.doi.org/10.1007/s11336-008-9099-3

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

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement, 71,* 306–324. http://dx.doi.org/10.1177/0013164410384856

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164. http://dx.doi.org/10.1177/014662168500900204

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60,* 523–531. http://dx.doi.org/10.1177/00131640021970691

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453. http://dx.doi.org/10.1037/1082-989X.3.4.424

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous

items: I: Algebraic lower bounds. *Psychometrika, 42,* 567–578. http://dx.doi.org/10.1007/BF02295979

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39,* 979–984. http://dx.doi.org/10.3758/BF03192993

Kelley, K., & Cheng, Y. (2012). Estimation of and confidence interval formation for reliability coefficients of homogeneous measurement instruments. *Methodology, 8,* 39–50. http://dx.doi.org/10.1027/1614-2241/a000036

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21,* 69–92. http://dx.doi.org/10.1037/a0040086

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44,* 486–507. http://dx.doi.org/10.1177/0049124114543236

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. London, UK: Methuen.

Lubienski, S., & Crane, C. C. (2010). Beyond free lunch: Which family background measures matter? *Education Policy Analysis Archives, 18,* 11. http://dx.doi.org/10.14507/epaa.v18n11.2010

Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research, 51,* 698–717. http://dx.doi.org/10.1080/00273171.2016.1215898

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical & Statistical Psychology, 23,* 1–21. http://dx.doi.org/10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2,* 255–273. http://dx.doi.org/10.1080/10705519509540013

Moltner, A., & Revelle, W. (2015). *Find the greatest lower bound to reliability*. Retrieved from http://personality-project.org/r/psych/help/glb.algebraic

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1–13. http://dx.doi.org/10.1007/BF02289400

Nunnally, J. C., & Bernstein, R. H. (Eds.). (1994). The assessment of reliability. *Psychometric theory* (pp. 248–292). New York, NY: McGraw-Hill.

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

Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist, 16,* 56–69.

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology, 98,* 194–198.

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173–184.

Raykov, T. (1997b). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329–353.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22,* 375–385.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35,* 299–331.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9,* 195–212.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95,* 129–140.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16,* 19–31.

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

Revelle, W. (2008). *psych: Procedures for personality and psychological research* (R package version 1.0–51).

Revelle, W. (2016, May). *Using R and the psych package to find* ω. Retrieved from http://personality-project.org/r/psych/HowTo/omega.pdf

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

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74,* 145–154.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthese, 16,* 170–233.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61.

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

Shapiro, A., & ten Berge, J. M. (2000). The asymptotic bias of minimum trace factor analysis, with applications to the greatest lower bound to reliability. *Psychometrika, 65,* 413–425.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods, 18,* 572–582.

Sheng, Y., & Sheng, Z. (2012). Is coefficient alpha robust to non-normal data? *Frontiers in Psychology, 3,* 34.

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

Sočan, G. (2000). Assessment of reliability when test items are not essentially τ-equivalent. *Developments in Survey Methodology, 15,* 23–35.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1,* 81–97.

ten Berge, J. M., & Kiers, H. A. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika, 56,* 309–315.

ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613–625.

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher, 22,* 209–213.

Terry, L., & Kelley, K. (2012). Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology, 65,* 371–401.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7,* 769.

van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature, 514,* 550–553.

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65,* 271–280.

Wolff, H. G., & Preising, K. (2005). Exploring item and higher order factor structure with the Schmid-Leiman solution: Syntax codes for SPSS and SAS. *Behavior Research Methods, 37,* 48–58.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29,* 377–392.

Zhang, Z., & Yuan, K. H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement, 76,* 387–411.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53,* 33–49.

Zinbarg, R. E., Revelle, W., & Yovel, I. (2007). Estimating ω h for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement, 31,* 135–157.

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,* 123–133.

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

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage.

Appendix

Software Code and Associated Screenshots

Using R

Basics and Installing Packages

Because R is open source, new statistical packages are being added almost daily. In R, a "package" is a set of procedures that can be used to perform certain statistical analyses. This is equivalent to the "Proc" commands in SAS, procedures in SPSS, or commands in Stata. For example, to fit a linear multilevel model, SAS uses the Proc Mixed procedure, SPSS uses the MIXED procedure, Stata uses the xtmixed command, and R would use the lme4 package.

In R, not all packages are available by default upon opening the program (in fact, only very basic packages are available). The packages needed to calculate scale reliability (`psych`, `MBESS`, and `userfriendlyscience`) are not included and must be installed. This is done with the following code:

```
install.packages("psych")
install.packages("MBESS")
install.packages("userfriendlyscience")
```

Note that code in R is case-sensitive so capitalization is important. After running this code, you will likely be prompted to select a "mirror site" which is the location from where these packages are downloaded. A list of geographic locations may appear; it makes little difference which is selected and they all contain the same information. These packages may take a few minutes to install. Installing packages only needs to be done once per machine. Once the packages are installed, they do not need to be installed again.

Loading the Data

Undoubtedly, one of the most difficult tasks when working with a new software is to successfully load the desired dataset. In this appendix, we use the data from the Big Five Inventory example because it is included as an internal example without the `psych` package. After installing the `psych` package, the Big Five Inventory dataset can be loaded with the following code,

```
data(bfi, package = "psych")
```

(*Appendix continues*)

In general, there are multiple ways to load data into R. Although the pathway to the file can be explicitly stated, it is often easier to find the desired file from a dialog menu. The following code shows how to input datafiles into R that are saved in either the .csv, .sav (SPSS), .dta (Stata), or permanent SAS data set formats.

```
install.packages("foreign")
require(foreign) # after installing a package, the require command tells
R to use the package
dat<-read.csv(file.choose()) # CSV
dat<-read.spss(file.choose())#SPSS
dat<-read.dta(file.choose()) #Stata
dat<-read.ssd(file.choose()) # SAS
```

If the `userfriendlyscience` package is already installed, then one can use the `getDat()` function to import data. This function determines the appropriate format and will automatically import the data and assign it the name "dat."

To simplify the analysis, we will separately break the full data into five separate data sets such that each of the five subscales are contained within their own data set.

```
agre<-bfi[,1:5]
cons<-bfi[,6:10]
extr<-bfi[,11:15]
neur<-bfi[,16:20]
open<-bfi[,21:25]
```

The name of the left side of the arrow is the new data name. On the right side of the arrow is the old dataset (called bfi here because that is the default name for this data when loaded in from R) and a set of brackets. Within these brackets, users specify which parts of the data matrix to use. The first value is blank because we want all the rows (people). The second numbers correspond to the columns in the data. So, for the *agreeableness* dataset (agre), we want the first five columns of the bfi data. The *conscientious* dataset (cons) is composed of the sixth through tenth columns of the bfi and so on.

Reverse Scoring

As is common in psychometric scales, some items may need to be reverse scored (this is required for appropriate calculation of some reliability coefficients like Cronbach's alpha). This can be done with the `invertItems` function that is part of the `userfriendlyscience` package.

```
agreRev <- invertItems(agre, 1)
consRev <- invertItems(cons, c(4,5));
extrRev <- invertItems(extr, c(1, 2));
openRev <- invertItems(open, c(2, 5));
```

This code creates a new R object (agreRev, consRev, extrRev, openRev) from the original R data. After the `invertItems` function, the first value within the parentheses is the data set to reverse score. After the comma, the numbers listed are the columns in the data that should be reverse scored. The "c" indicates that a list will follow and is needed if multiple items are reverse scored. So, the *agreeableness* scale will reverse score Item 1, the *conscientiousness* scale will reverse score Items 4 and 5, and so on. The *neuroticism* scale does not contain any items that need to be reverse scored.

Cronbach's Alpha

Cronbach's alpha can be calculated as part of many different functions. The simplest is to use the `alpha` function from the `psych` R package. If relevant items are reverse scored as discussed previously, then the only argument of the `alpha` function is the dataset.

```
alpha(agreRev)
alpha(consRev)
alpha(extrRev)
alpha(neur)
alpha(openRev)
```

(*Appendix continues*)

The output for the *agreeableness* scale is as follows. The estimate of Cronbach's alpha can be found in the first row of the output under std.alpha.

```
> omega(agreRev)> > > alpha(agreRev)

Reliability analysis
Call: alpha(x = agreRev)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd
      0.7      0.71    0.68      0.33 2.5 0.009  4.7 0.9

 lower alpha upper     95% confidence boundaries
0.69 0.7 0.72

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r S/N alpha se
A1      0.72      0.73    0.67      0.40 2.6   0.0087
A2      0.62      0.63    0.58      0.29 1.7   0.0119
A3      0.60      0.61    0.56      0.28 1.6   0.0124
A4      0.69      0.69    0.65      0.36 2.3   0.0098
A5      0.64      0.66    0.61      0.32 1.9   0.0111

 Item statistics
      n raw.r std.r r.cor r.drop mean  sd
A1 2784  0.58  0.57  0.38   0.31  4.6 1.4
A2 2773  0.73  0.75  0.67   0.56  4.8 1.2
A3 2774  0.76  0.77  0.71   0.59  4.6 1.3
A4 2781  0.65  0.63  0.47   0.39  4.7 1.5
A5 2784  0.69  0.70  0.60   0.49  4.6 1.3
```

Omega Total

To calculate the measure that we call omega total (*not* Revelle's omega total), one must go outside of the `psych` package to the `MBESS` package.

In the `MBESS` package, the `ci.reliability` function will estimate omega total as well as its confidence interval.

```
require(MBESS)#only necessary the first time the package is used
ci.reliability(agreRev)
```

This yields the following output:

```
> require(MBESS)
> ci.reliability(agreRev)
$est
[1] 0.7104131

$se
[1] 0.01018984

$ci.lower
[1] 0.6904414

$ci.upper
[1] 0.7303848

$conf.level
[1] 0.95

$type
[1] "omega"

$interval.type
[1] "robust maximum likelihood (wald ci)"
```

The estimate of omega total is the first value which appears beneath $est. On the *agreeableness* subscale, omega total is estimated to be 0.71 with a 95% confidence interval of [.69, .73]

(*Appendix continues*)

Revelle's Omega Total

Revelle's omega total is calculated from the `omega` function in the `psych` package. The `omega` function also outputs Cronbach's alpha as well, so it can be used in lieu of the `alpha` function. Again, the only argument needed in the function to obtain Revelle's omega total using a Pearson covariance matrix is the data set.

```
omega(agreRev)
omega(consRev)
omega(extrRev)
omega(neur)
omega(openRev)
```

The output from this function for the *Agreeableness* subscale is as follows:

```
> omega(agreRev)
Omega
Call: omega(m = agreRev)
Alpha:                 0.71
G.6:                   0.68
Omega Hierarchical:    0.64
Omega H asymptotic:    0.83
Omega Total            0.77

Schmid Leiman Factor loadings greater than  0.2
      g   F1*   F2*   F3*   h2   u2   p2
A1 0.31  0.26              0.18 0.82 0.55
A2 0.62  0.57              0.71 0.29 0.55
A3 0.78               0.39 0.76 0.24 0.80
A4 0.43        0.22        0.25 0.75 0.75
A5 0.57        0.21        0.39 0.61 0.82

With eigenvalues of:
   g  F1*  F2*  F3*
1.61 0.40 0.10 0.19
```

The Alpha row shows Cronbach's alpha, which matches the output from the `alpha` function. Revelle's omega total is the last value in the first set of values which is listed as 0.77. Notice that this value is not the same as omega total because it uses a variance decomposition based on a Schmid-Leiman transformation (the details of which are provided below the output).

A convenient option in the `omega` function is that a polychoric covariance matrix can be estimated and used internally and is possible by specifying only two additional words in the code.

```
omega(agreRev, poly=TRUE)
omega(consRev, poly=TRUE)
omega(extrRev, poly=TRUE)
omega(neur, poly=TRUE)
omega(openRev, poly=TRUE)
```

(*Appendix continues*)

The output from the omega function with the polychoric option for the *agreeableness* subscale is as follows:

```
> omega(agreRev, poly=TRUE)
Omega
Call: omega(m = agreRev, poly = TRUE)
Alpha:                 0.76
G.6:                   0.74
Omega Hierarchical:    0.69
Omega H asymptotic:    0.83
Omega Total            0.83

Schmid Leiman Factor loadings greater than  0.2
      g  F1*  F2*   F3*   h2   u2   p2
A1 0.34 0.24            0.19 0.81 0.61
A2 0.70 0.71            0.99 0.01 0.49
A3 0.79      0.49       0.87 0.13 0.72
A4 0.52           0.23 0.32 0.68 0.83
A5 0.62                0.43 0.57 0.88

With eigenvalues of:
   g  F1*  F2*  F3*
1.88 0.56 0.29 0.08
```

Notice that the Alpha and (Revelle's) omega total values are much higher than in the previous output. The `alpha` function does not feature this `poly` option, so Cronbach's alpha with a polychoric covariance matrix is best run through the `omega` function.

The computation of Revelle's omega total is a little involved and there are not many sources that describe this version of the omega coefficient (outside of documentation for the `psych` R package). We outline where Revelle's omega total from where comes for the remainder of this section to elucidate what Revelle's omega total is calculating. In Equation 4 of the main text, we defined Revelle's omega total as

$$\omega_{RT} = \frac{\left(\sum_{i=1}^{k} \lambda_{gi}\right)^2 + \left(\sum_{f=1}^{F}\sum_{i=1}^{k_f} \lambda_{fi}\right)^2}{V_X}$$

Revelle (2016) notes the numerator of this formula is equal to the communality of each item, h_i^2 so the formula can be rewritten as

$$\omega_{RT} = 1 - \frac{\sum_{i=1}^{K}(1 - h_i^2)}{V_X} \tag{A1}$$

This can be simplified to

$$1 - \frac{\sum_{i=1}^{K}(u_i^2)}{V_X} \tag{A2}$$

where u_i^2 is the uniqueness of the ith item (a.k.a. the error variance).

Using the polychoric covariance analysis of the *agreement* subscale above, the communalities appear in the "h2" column and the uniquenesses appear in the "u2" column. The sum of the uniquenesses is equal to .81+.01+.13+.68+.57 = 2.20 which is the numerator of Equation A2. Unfortunately, the denominator V_X does not appear in the output. Fortunately, this value is quite simple to calculate in R. Recall, that V_X is equal to the sum of all elements of the sample correlation matrix. The polychoric correlation matrix in R can be saved as an object with the following code:

```
mat<-polychoric(agreRev)
agrepoly<-mat$rho
```

(*Appendix continues*)

The `sum` function can then be used to add all the individual elements

```
sum(agrepoly)
```

which yields

```
> sum(agrepoly)
[1] 12.72892
```

Therefore, $\omega_{RT} = 1 - \frac{2.20}{12.73} = 0.8272 \approx 0.83$, matching the output above.

Omega hierarchical is similar except that the numerator is only equal to the variance explained by *only* the common factor. This can be found by adding up all the values in the "g" column and squaring (be sure to add first and then square the sum, do not square first and then add the squares). In the polychoric *agreement* example, $(.34 + .70 + .79 + .52 + .62)^2 = 8.82$. V_X is still equal to the same value (12.73) so hierarchical omega is equal to 8.82/12.73 = 0.692.

Greatest Lower Bound

The `glb.fa` function in the `psych` package estimates the greatest lower bound. Similar to other methods in the `psych` package, the only necessary argument of the function is the data name.

```
glb.fa(agreRev)
glb.fa(consRev)
glb.fa(extrRev)
glb.fa(neur)
glb.fa(openRev)
```

The output for the *agreeableness* subscale is as follows,

```
$glb
[1] 0.7457282

$communality
      A1-        A2        A3        A4        A5
0.2089775 0.5771862 0.5666628 0.2282605 0.4613867

$numf
[1] 2

$Call
glb.fa(r = agre)
```

The greatest lower bound estimate appears as the first item in the output after `$glb`

Unfortunately, the `glb.fa` function does not offer the option to use a polychoric covariance matrix internally and therefore uses a Pearson covariance matrix. However, this can be circumvented by separately estimating a polychoric covariance or correlation matrix, and using that as the input file instead of the raw data. However, it can be a bit tricky to save a polychoric correlation matrix as a data frame in R.

First, the polychoric matrix is estimated with the polychoric function from the psych package. Rather than immediately outputting the results, the output is saved to an object (called "mat" in the code below). The output contains both the polychoric correlation matrix and thresholds; the thresholds are not needed, so we want to exclude them and only save the matrix. In doing so, we also must convert the object to a data frame. The R code for doing so for the *agreeableness* subscale is as follows:

```
mat<-polychoric(agreRev)
agre.poly<-as.data.frame(mat$rho)
```

The `glb.fa` function can accept a correlation matrix as input, so we can use the saved polychoric correlation matrix as the input of the function.

```
glb.fa(agre.poly)
```

(*Appendix continues*)

This will provide the desired output,

```
> mat<-polychoric(agreRev)
> agre.poly<-as.data.frame(mat$rho);
> glb.fa(agre.poly)
$glb
[1] 0.7933371

$communality
       A1        A2        A3        A4        A5
0.2407055 0.6996489 0.6365752 0.2623242 0.5301503

$numf
[1] 2
```

scaleStructure Function

Although the above analyses are not difficult to perform because the commands are quite straightforward, for inexperienced or reluctant R users, the `scaleStructure` package can estimate these quantities in a single pass and summarizes the output.

```
scaleStructure(dat=agreRev, ci=FALSE)
scaleStructure(dat=consRev, ci=FALSE)
scaleStructure(dat=extrRev, ci=FALSE)
scaleStructure(dat=neur, ci=FALSE)
scaleStructure(dat=openRev, ci=FALSE)
```

`ci=FALSE` indicates that we do not want the confidence interval for the estimate (although best practice suggests that this is helpful to report).

The output for the *agreeableness* subscale from this function is as shown on the further in the Appendix.

```
> scaleStructure(dat=agreRev, ci=FALSE)

Information about this analysis:

                 Dataframe: agreRev
                     Items: all
              Observations: 2709
     Positive correlations: 10 out of 10 (100%)

Estimates assuming interval level:

             Omega (total): 0.71
      Omega (hierarchical): 0.64
   Revelle's omega (total): 0.77
Greatest Lower Bound (GLB): 0.75
             Coefficient H: 0.77
          Cronbach's alpha: 0.7

Estimates assuming ordinal level:

     Ordinal Omega (total): 0.77
 Ordinal Omega (hierarch.): 0.77
  Ordinal Cronbach's alpha: 0.76
```

The function goes through the previously outlined methods, estimates reliability, saves the output, and summarizes them in one window. The first set of output shows the results assuming a Pearson covariance matrix followed by results that use a polychoric covariance matrix. It also differentiates between omega total and Revelle's omega total and is the only R package of which the author is aware that provides estimates of Coefficient *H*.

(*Appendix continues*)

Using Excel

Although R is the best available software option for estimating alternatives to Cronbach's alpha (and it is open source), we realize that some users may be hesitant to adopt a new software program, especially to use methods with which they are unfamiliar. In attempt to make these methods as broadly accessible as possible, we have included two Excel spreadsheets for calculating omega total and Coefficient *H* using only the standardized loadings from a factor analysis. These loadings can be obtained from any software program of the user's choosing and does not require learning any new software.

The provided Excel spreadsheet has two tabs, one for Coefficient *H* and one for omega total. The spreadsheet allows for up to 36 items. A factor analysis must be conducted to obtain the factor loadings. This can be done in any program of the user's choosing. Then, these loadings are placed into column B of the spreadsheet. For omega total, the spreadsheet is setup to automatically calculate the uniqueness terms based on the standardized loadings. Column G for Coefficient *H* and column F for omega total will reveal the estimate of these measures.

Using the *agreeableness* subscale example that was used in the previous section, we will first obtain the standardized factor loadings using maximum likelihood in R using the `fa` function from the `psych` package. These loadings need not be obtained from R and can be estimated from any program of the user's choice (e.g., M*plus*, SPSS, SAS, Stata)

```
fa(agre, nfactors=1, fm="ml")
```

The output of this analysis yields the following:

```
> fa(agre, nfactors=1, fm="ml")
Factor Analysis using method =  ml
Call: fa(r = agre, nfactors = 1, fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
     ML1   h2   u2 com
A1- 0.37 0.14 0.86   1
A2  0.66 0.43 0.57   1
A3  0.76 0.58 0.42   1
A4  0.48 0.23 0.77   1
A5  0.63 0.40 0.60   1
```

The "ML1" column contains the standardized factor loadings for this scale (these correspond to those provided in Table 3 of the main text). Taking these loadings and entering them into the Excel spreadsheet for Coefficient *H* and omega total gives:

	A	B	G
1	Item	Loadings	**Coefficient H**
2	1	0.370	**0.765**
3	2	0.660	
4	3	0.760	
5	4	0.480	
6	5	0.630	

	A	B	C	F
1	Item	Loadings	Uniqueness	**Omega**
2	1	0.370	0.863	**0.72297**
3	2	0.660	0.564	
4	3	0.760	0.422	
5	4	0.480	0.770	
6	5	0.630	0.603	

Received July 9, 2016
Revision received February 12, 2017
Accepted February 20, 2017 ■