View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Nottingham Trent Institutional Repository (IRep)

Abstract

Coefficient alpha is the most popular measure of reliability (and certainly of internal consistency reliability) reported in psychological research. This is noteworthy given the numerous deficiencies of coefficient alpha documented in the psychometric literature. This mismatch between theory and practice appears to arise partly because users of psychological scales are unfamiliar with the psychometric literature on coefficient alpha and partly because alternatives to alpha are not widely known. We present a brief review of the psychometric literature on coefficient alpha, followed by a practical alternative in the form of coefficient omega. To facilitate the shift from alpha to omega we also present a brief guide to the calculation of point and interval estimates of omega using a free, open source software environment.

The construction and application of psychometric scales has become accepted best practice when attempting to measure human performance and behaviour. The implications of test ‘quality’ for the individual and society are unquestioned. Statistical procedures that attempt to assess reliability have acquired the status of ingrained conventions, with certain types of analyses being routinely adopted. The predominant framework under which most such procedures fall is Classical Test Theory (CTT) (e.g., see Lord & Novick, 1968). This is the most popular way of conceptualising how a scale should perform and function. In recent years improved approaches to reliability estimation have been advocated by psychometricians. Yet, despite a widespread dissemination and publication of alternatives, there remains a staunch resistance to advancements in the interpretation, application, and reporting of a scale’s reliability, particularly when it comes to internal consistency.

The *APA Task Force on Statistical Inference* (Wilkinson and APA Task Force on Statistical Inference, 1999), placed emphasis on the correct use and treatment of reliability estimates. The most common type of reliability estimate reported in articles published by the *American Psychological Association* were internal consistency estimates (as opposed to test-retest or parallel forms). These accounted for 75% of all reported reliabilities (Hogan et al., 2000). The most common means of assessing internal consistency in the social sciences is that of coefficient alpha – also termed Cronbach’s alpha (alpha) (following Cronbach’s influential 1951 paper). This has become a routinely relied upon statistic for estimating a scale’s internal consistency. A recent search by the current authors (via Google Scholar®, 2012) confirms its prevalence – showing it to have been cited some 17,608 times since its original publication. However, as Cronbach himself stated, “The numerous citations to my paper by no means indicate that the person who cited it had read it, and does not even demonstrate that he had looked at it." (Cronbach & Shavelson, 2004, p.392).

Reflective of Cronbach's comment, researchers' understanding of reliability analysis is generally low. Aiken et al. (1990) reported only 27% of postgraduate courses judged that a clear majority of their students (>75%) could use 'methods of reliability measurement' correctly, while 38% of courses judged that fewer than 25% of their students were capable of this. Aiken et al. (2008) recorded only a modest increase, with 46% of courses judged to have a clear majority of their students (>75%) capable of reliability assessment. This percentage thus remains relatively low given alpha's prevalence.

Despite alpha coming under heavy scrutiny in numerous articles (e.g., Green et al., 1977; Green & Yang, 2009; Green & Hershberger, 2000; Raykov, 1998; Zimmerman et al., 1993; Huysamen, 2006; Zinbarg et al., 2005; Sijtsma, 2009), it remains one of the most pervasive statistics in work that uses psychometric scales. This begs the question of why alpha is so prevalent and why alternatives to alpha are not applied. We argue that this is the result of a number of factors. First, it remains likely that fewer than half of all postgraduate courses in psychology offer in-depth coverage of methods of reliability analysis. This perhaps accounts for a portion of alpha's misuse. However, Aiken et al. (2008) show there has been a steady increase in knowledge of reliability measurement. This suggests one would expect to witness a reduction in the use of alpha (for which there is little evidence). There is an important difference between one being aware of a statistical weakness and being capable of actually addressing those weaknesses. Furthermore, articles that condemn alpha tend to be very technical, and thus not only inaccessible to many psychologists but also fail to offer 'realistic' alternatives (Revelle & Zinbarg, 2009), though a recent paper by Kelley and Cheng (2012) provides a clear introduction to the technical challenges of internal consistency reliability estimation. Elsewhere in the literature, if implementation of an alternative is offered, it is either treated in passing or presented in a manner too complex for non-experts to implement easily. For example, many researchers advocate the use of Structural Equation Modeling (SEM) as the most robust tool to assess a test's reliability, mainly, because it allows one to specify and compare different models of reliability (e.g., Yang & Green, 2011; Graham, 2006; Miller, 1995). The current authors are aware that such methods as latent variable modeling outperform current analyses of reliability; however, this will rarely be the most appealing approach for the majority of researchers and users of psychometric scales. SEM for example demands large sample sizes, and it may also be impractical in the sense that it requires considerable expertise to employ correctly. This, combined with the observation that, as of 2008, only 12% or so of postgraduate psychology programs offer in-depth coverage of structural equation modeling (SEM) (Aiken et al. 2008), makes it unlikely that SEM will replace alpha in the near future. It appears that the crux of the problem is a confluence of the awareness of alpha's limitations and the lack of suitable alternatives.

What is coefficient alpha?

In order to consider alternatives to alpha it is necessary to be clear on the basic principles underpinning the calculation of alpha. In *classical test theory* (CTT) a portion of the variability in participants' responses is thought to be due to a genuine underlying difference - the true score (*T*) - in the trait being measured (*X*). The remaining portion of variability is considered to be composed of random measurement errors (*E*). Hence, $X = T + E$. Reliability is defined as the ratio of variability in T (σ_T^2) to the variability in X (σ_X^2). In other words, it is the proportion of variability in the scale score (e.g., score on a self-esteem measure) that is attributable to the trait being measured (e.g., self-esteem), in relation to the total variability contained within the participants’ responses. This allows for the quantification of reliability by encapsulating the degree of systematic responding to a scale (hypothesized to reflect the trait itself) in relation to the total amount of systematic and non-systematic responding (the trait plus all other error). This can be formulated as σ_T^2/σ_X^2 where $\sigma_X^2 = \sigma_E^2 + \sigma_T^2$. The consequence of adhering to CTT’s conceptualisation of reliability means that anything that influences error will inevitably change the reliability. This signifies that reliability is not a property of a test per se, but rather a property of a scale applied in a given context to a particular population (Miller, 1995; Thompson & Vacha-Haase, 2000).

There is also a fundamental limitation in estimating the degree of error of a scale: a researcher will never know with certainty the exact value of a test’s reliability in any given situation. This is because the variability of *T* and the variability of *E* (or the variability of X which is the sum of the variability of T and the variability of E) cannot be known perfectly, owing to the sampling error that inevitably arises. The value of the model in the population is therefore unknown (Miller, 1995). Thus, all models under CTT are ‘under-identified’; that is, there are more parameters than there are statistics (Nunnally & Bernstein, 1994). An under-identified model has too many degrees of freedom and would therefore allow the reliability estimate to take any possible value. Thus any model employed to estimate reliability must contain within it simplifying assumptions that permit the key parameters to be estimated. The assumptions inherent in the model will ultimately determine how restrictive it is in informing the user of a scale’s reliability. The more stringent the assumptions, the less likely the data are to conform to them. There have been a number of models put forward that one can apply to obtain a reliability ratio of true score variance to error score variance. There are currently three dominant methods for doing so: *essentially tau-equivalent*, *congeneric*, and *parallel*.

Reliability models

It is crucial to make sure assumptions of the models and statistical methods in use are met by the data at hand. Just as routine tests to check for any violations of normality should be carried out, so equally should tests for assumptions in reliability estimation be applied. After all, one can only be guided to assess such assumptions when one is fully aware of the optimal criteria by which the model being utilised performs. The model that defines alpha is the *essentially tau-equivalent* model, sometimes termed the 'true-score equivalent' model (McDonald, 1999). For present purposes it is useful to discuss the essentially tau-equivalent model in relation to the alternative *parallel* and *congeneric* models from the standpoint of CTT. Despite the seemingly obscure labels given to the models, all are connected by four underlying and easily-described properties of a scale (e.g., see Graham, 2006). These properties are as follows:

> i) the extent to which each item measures the same underlying personality trait (unidimensionality);
>
> ii) whether the true scores for different items have the same mean (sensitivity)[1];
>
> iii) whether the true scores for different items have the same variance; and
>
> iv) whether the error variance is the same for each item.

All three models are suitable for unidimensional traits. Thus the degree to which one assumes either constancy or variability of properties ii) to iv) is what distinguishes the *essentially tau-equivalent* from *parallel* or *congeneric* models.

The *parallel model* is the most restrictive model for measuring internal consistency. It assumes constant item means, item variances and error variances. It thus assumes all items

[1] Although Graham (2006) refers to this as 'precision', we specifically use the term 'sensitivity' to avoid any confusion with wider uses of the term which generally refer to the inverse of variance and thus absence of measurement error (e.g., narrowness of a CI).

are tapping the same personality trait, on the same scale, with identical precision and error (Kirstof, 1964; Lord & Novick, 1968; Raykov, 1997a, 1997b). The *essentially tau-equivalent model* is considerably less restrictive than the parallel model. It assumes constant item variances for the true scores but allows both the true score means and the error variances of the items to vary (Raykov, 1997a)[2]. The *congeneric model* is the least restrictive of the three. Means and variances of the true scores and the error variances are allowed to vary. The congeneric model remains suitable for a unidimensional construct, but avoids inconvenient and unrealistic assumptions about constant means and variances (Jöreskog, 1972). Problems with alpha stem largely from the restrictive nature of the underlying reliability model it assumes.

A brief review of problems with coefficient alpha

Alpha is biased

Alpha has been shown only to be representative of a measure's internal consistency when the assumptions of the essentially tau-equivalent model are met (Green et al, 1977; Revelle & Zinbarg, 2009; Shevlin et al., 2000; McDonald, 1981). However, these requirements are seldom met in practice for psychological scales (Green & Yang, 2009). Assuming the true score variance is constant across all items is exactly where alpha runs into problems. The possibility of a scale resulting in equal sensitivity across all items is unrealistic. Hence alpha is regarded by methodologists as an inappropriate measure of internal consistency reliability.

The lack of perfect inter-correlations between items' true scores (i.e., unidimensionality) is another aspect of a test that also causes problems for alpha. Perfect unidimensionality may avoid violating the equal-item variance (i.e., constant 'scale') property of the essentially tau-

[2] An important distinction should be made between a *tau-equivalent* model and an *essentially tau-equivalent* model in that a *tau-equivalent* model would allow error variances but not true-score means or variances to vary.

equivalent model. However, Sočan (2000) argues that few scales, particularly those in the personality domain, are ever truly unidimensional and instead nearly always possess some degree of multidimensionality. This element of multidimensionality inhibits perfect inter-correlations of items' true scores (Sočan, 2000), something alpha must meet if it to be considered a measure of internal consistency reliability. Additionally, if a test uses multiple response formats across items it is again more likely to violate essentially tau-equivalence (equal item variance) than those that do not (Graham, 2006). Cronbach himself concluded that the alpha formula is not appropriate for scales where questions are designed to target different areas or processes (Cronbach & Shavelson, 2004). This is ultimately the aim of many, if not all, items contained within a psychometric scale.

A key point to note is that the extent of the bias resulting from the violation of the essentially tau-equivalence model is far from trivial. Raykov (1997b) showed that a scale with all but one item assumed to be essentially tau-equivalent has substantial effects on alpha's use as a reliability measure. Lord and Novick (1968) showed that for a set of scale items where essentially tau-equivalence assumptions had been violated, alpha's estimation of reliability is lower than the population (i.e., true) level of reliability. This means alpha tends to underestimate the degree of internal consistency of a scale when errors are uncorrelated. Generally, the larger the violation of tau-equivalence the more coefficient alpha underestimates score reliability (Graham, 2006). Other authors have shown that alpha's estimation of reliability can also be inflated when the errors for each item are correlated or the number of items is significantly increased (Yuan & Bentler, 2002; Cortina, 1993). The degree of alpha's bias has also been shown to depend on the consistency of the samples from which scores are drawn. Waller (2008) demonstrated that when samples are 'commingled' (i.e., when scores are derived from multiple populations) alpha can be severely biased and the direction of this bias is difficult to predict. Waller (2008) points out that while in the majority of instances commingling inflates alpha (overestimation), in others it attenuates it (underestimation). This means comparisons across studies are difficult to make because levels of sample consistency are usually unknown. Furthermore, many researchers would find it extremely difficult to justify their definition of what constitutes distinct or similar populations. This highlights one of the chief problems in using alpha; it is difficult to gauge the magnitude, direction, or even the source of any bias.

Alpha if item deleted

During scale development psychologists will often cite ‘alpha when items deleted’ as a means of determining a preference for variants of the initial measure. This method allows one to observe any changes in alpha (i.e., reliability for remaining items) when certain items are excluded. It is also employed in the reevaluation of a measure for the purpose of shortening it. Raykov (1997) criticises the sample specificity of alpha and argues that any changes in alpha resulting from the ‘item deleted’ process are really only a consequence of the characteristics of the sample at hand and thus any implied population inferences cannot be carried over to uses with alternative samples. Secondly, the population estimate of alpha can easily be overestimated or underestimated due to the deletion of an item (see Raykov, 2007). That is, ‘alpha if item deleted’ (which is a sample statistic) may go up with the removal of an item whereas the level of true score remaining in the test has gone down (or vice versa). Hence, any reported gains in the reliability of alpha by deleting an item, are not representative of the effect this will have on the ‘true’ or population reliability of a scale. This can be explained simply by the fact that the process of ‘alpha when item deleted’ assumes equal error variance across all items. This is in contrast to the assumptions of the *essentially tau-equivalent* model (see assumptions above) that allow error variance to vary across items.

The item suggested for deletion could contain less error than the rest of the items. As a result, the deleted item with the smaller error variance would have been more representative of the population value (i.e., possess more ‘true score’ and less error) making the scale a more reliable measure of the personality trait.

Point estimation

In statistics, point estimation is the process of summarizing information about a population parameter with a single number (a ‘statistic’). This statistic is considered an estimate or ‘best guess’ of an unknown population parameter. In alpha’s case the unknown population parameter is the ‘true’ reliability of a scale. Many authors agree that one of the major flaws in the current application of reliability measures such as alpha is that they are nearly always reported as point estimates (Terry & Kelley, 2012; Raykov, 2002). That is, a researcher will frequently offer only one value as an indicator for a scale’s degree of reliability based on the sample data (e.g., “Cronbach’s alpha = .77”). What must also be taken into account is the level of reliability of the reliability estimate itself. As we have seen, alpha can vary according to a number of different factors and it can be biased in different directions. Reporting the level of certainty that a reliability estimate offers, whilst considering the characteristics of the data set at hand (i.e., factors that influence its precision), has the potential to drastically improve the interpretation of point estimates in psychometric applications. Interval estimates such as confidence intervals (CIs) are the natural way of incorporating precision of an estimate into a statistical summary. They are conceptually straightforward to understand and are considered to be a benchmark for rigorous statistical reporting in psychology and other disciplines (e.g., see Kelley & Preacher, 2012; Baguley, 2009).

Intrinsically linked with the reporting of alpha as a point estimate, is the use of a cut-off heuristic, thought to reflect the crucial stage at which a scale possesses good or poor internal consistency. The heuristic in reliability reporting is based on Nunnally and Bernstein's (1994) recommendation of .70. As pointed out by several authors (e.g., Baguley, 2008; Iacobucci & Duhachek, 2003), Nunnally's cut off of .70 was never intended as a gold standard for acceptable reliability. Iacobucci and Duhachek (2003) point out that for a simple two-item scale with an item inter-correlation of .60 and a sample of 30, coefficient alpha is .75 with a 95% confidence interval from .64 to .86. This lower bound of .64 implies that the scale could plausibly consist of 36% or more noise.

The main difficulties with the use of alpha as a measure of internal consistency can be summarized as follows:

1) alpha relies on assumptions that are hardly ever met,
2) violation of these assumptions causes alpha to inflate and attenuate its internal consistency estimations of a measure,
3) 'alpha if item deleted' in a sample does not reflect the impact that item deletion has on population reliability, and
4) a point estimate of alpha does not reflect the variability present in the estimation process, providing false confidence in the consistency of the administration of a scale.

A practical alternative to alpha

Some issues with the use of alpha, such as point estimation, can be addressed relatively straightforwardly by obtaining confidence intervals. One means of achieving this is by a method known as *bootstrapping*. For example, Iacobucci and Duhachek (2003) provide a simple method of bootstrapping to obtain CIs for alpha[3]. Bootstrapping involves repeated resampling with replacement from a sample to obtain an ‘empirical’ distribution of an estimator such as alpha (DiCiccio & Efron, 1996). Once this distribution has been obtained, confidence intervals or other quantities can be constructed from it (e.g., taken directly from the empirical distribution in the simplest case). Bootstrapping is attractive as a method when obtaining an analytic solution is difficult or known to perform badly (e.g., because it relies on distributional or other assumptions that rarely hold in practice). In the case of reliability coefficients, the assumption of multivariate normality assumed by the usual analytic methods rarely holds (Yuan & Bentler, 2002; Kelley & Cheng, 2012).[4]

Calculating a CI allows researchers to report a range of plausible values for the internal consistency of the administration of the scale in the population with some specified degree of confidence. However, as is now established, alpha relies on assumptions that will only rarely be met. Although this is also true of many statistical analyses in psychology, one should always seek to find alternatives where assumptions are more regularly met. Inevitably, any alternative should not be based on the overly restrictive essentially tau-equivalent model. Many researchers have sought to compare the performance of alpha, as an estimate of internal consistency, with either parallel or congeneric models. Some authors have reported discrepancies of up to tenths (i.e., .10) (Sijtsma, 2009), meaning where one model might estimate internal consistency as .70 another may estimate it as .60. Predominantly the larger discrepancies were a consequence of gross violation of the essentially tau-equivalent model. In terms of alternative models under CTT, numerous authors have illustrated the benefits of using a congeneric model over the essentially tau-equivalent model, for the simple reason that the congeneric model is less restrictive. This makes the congeneric model a more appropriate

[3] Iacobucci and Duhachek (2003) explain how to bootstrap a CI for alpha in SPSS, while R users can use the cronbach.alpha() function in the package ltm (Rizopoulos, 2012).

[4] A relatively safe option in the case of a correlation or reliability coefficient is to use a bias-corrected and accelerated (BC_a) bootstrap approach (Kelley & Cheng, 2012). This is an adaptation of a basic bootstrap approach that corrects for skew (bias) and kurtosis in the empirical sampling distribution.

approach for the majority of psychometric research. As it is less restrictive there is less probability of a researcher violating its assumptions. As the congeneric model allows item variances to vary (i.e., they are not assumed to be constant) it will not result in the lower bound estimations of reliability characteristic of alpha (Soĉan, 2000).

One measure that adheres to the congeneric model is that of *omega* (McDonald, 1999). Omega has been shown by many researchers to be a more sensible index of internal consistency – both in relation to alpha and also when compared to other alternatives (Zinbarg et al, 2005, 2006, 2007; Graham, 2006; Revelle & Zinbarg, 2005; Raykov, 1997). Zinbarg et al. (2005) report that even when the assumptions of the essentially tau-equivalent model are met, omega performs at least as well as alpha. However, under violations of tau-equivalence – conditions likely to be the norm in psychology – omega outperforms alpha and is clearly the preferred choice. Additionally, factors such a commingling populations or the use of incongruent response formats as highlighted by Waller (2008) and Graham (2006), would not have such divergent effects on omega as they do alpha. The assumption of scale constancy that commingling or invariant response formats would potentially violate when using alpha does not underpin omega. Omega is less risk of overestimation or underestimation of reliability.

Raykov (1998) strongly recommends the use of congeneric models along with bootstrapping to obtain confidence intervals for omega in preference to alpha. In line with this, Kelley and Preacher (2012) highlight that CIs are an important requirement for *any* good effect size metric (of which a reliability estimate is an exemplar). A further benefit of CIs is that they are generally well understood not only by psychologists but also by users of psychological scales who are not in the academic or research domain (Cronbach &

Shavelson, 2004). For these reasons we provide a practical walk-through to obtaining omega along with CIs below. One major benefit of the method illustrated below is that while some familiarity with the rationale behind bootstrapping is required, researchers no longer need to set up the bootstrapping process themselves.

With regards to the development or intended shortening of a scale, an alternative to 'alpha if item deleted' is offered. This method further exploits the benefits of CIs as already advocated. As highlighted above, when an item is deleted and alpha as a point estimate recalculated, it is quite possible that items containing less error have been disposed of incorrectly. After CIs for omega have been established it is then possible to omit items using an iterative process and recalculate the corresponding standard error (i.e., CIs) of each new omega value. This way, each item's contribution in terms of error can be estimated and a more appropriate means of establishing omega's predictive power of reliability can be employed. From a practical point of view, a researcher would need to run the procedure set out below exchanging one test item at a time and then make comparisons between omega values along with the dispersion of error (CIs); selecting the highest omega value with the narrowest CIs.

One point of caution should be noted. Unidimensionality is a requisite for all the models considered here. Therefore if a scale is known to be multidimensional (or factor analysis demonstrates some divergence from unidimensionality) then it is recommended that the scale be split into subscales. Omega and CIs for omega can then calculated on each subscale separately (Soĉan, 2000).

Overall, omega's main advantages over alpha can be summarized as follows:

1) Omega makes fewer and more realistic assumptions than alpha
2) problems associated with inflation and attenuation of internal consistency estimation are far less likely
3) employing ‘omega if item deleted’ in a sample is more likely to reflect the true population estimates of reliability through the removal of a certain scale item, and
4) the calculation of Omega alongside a confidence interval reflects much closer the variability in the estimation process, providing a more accurate degree of confidence in the consistency of the administration of a scale.

Omega as a point estimate overcomes some of the fundamental problems intrinsic to the calculation of internal consistency evident with alpha. However, limitations such as point estimation need to be considered for omega also. The next section provides a guide that will enable the calculation of omega as well as the estimation of confidence intervals under one approach. This approach remedies both the pitfalls of an essentially tau equivalent model (i.e., alpha) and also that of quantifying internal consistency using only point estimation.

Obtaining point and interval estimates for omega

Until recently, bootstrapping to obtain CIs has been considered tedious and time consuming for researchers (Raykov, 1998). However, it is relatively easy given the availability of free, open source statistical software (notably R, used here) to calculate omega for typical sample sizes used in psychology. Here we present a brief worked example of how to calculate point and interval estimates of McDonald’s omega using the MBESS package (Kelley & Lai, 2012) in R (R Development Core Team, 2012) (see Fig. 2). Included is a method of reading data into ‘R’ from two common formats. For further information on statistical analysis in R for psychologists we suggest consulting recent texts such as Field, Miles and Field (2012) or Baguley (2012).

The first step is to download and install a recent version of R appropriate for your platform (PC, Mac OS or linux). R can easily be obtained at no financial cost from the 'R' website (http://www.r-project.org/). R employs a command-line (interpreter) whose job it is to communicate with the computer's operating system (OS). Thus, one can calculate statistics such as omega by loading the appropriate add-on packages and typing in a series of commands.

Once installed, opening R will call up the 'R Console' window. In the R console you will see the symbol '>'. This is the R prompt. Commands typed after the prompt can be executed by hitting the return key. The first step in obtaining omega is to load your data into R. R requires either a full path name to be specified or for data files to be in its working directory. The working directory can be set via menus in the R console or by typing a call to the `setwd()` function. For instance, if the data were in a folder on called "omega example" on the desktop of your computer you could set this as your working directory by entering the following command after the prompt `>`.

```
> setwd(file.path(Sys.getenv("HOME"), "Desktop", "omega example"))
```

A common way to load data into R is to use comma separated variable (.csv) files (e.g., created and saved from spreadsheet programs such as Excel). In this format variables are arranged in columns, each column is separated by a comma in the saved file. Often (as is assumed here) the files also have a header row containing the variable names. The read.csv() function in R can read data in this format. The supplementary files for this paper include a file called SES.csv. The following command loads this file into R (assuming it is in the working directory):

```
> SES <- read.csv("SES.csv")
```

The arrow formed by the symbols <- is an assignment operator. Here it is used to take the output of the read.csv() function (the data from the .csv file) and place it in a newly created R object named SES. SES is a type of R object known as a data frame. For present purposes it is useful to think of this as a set of named variables arranged in a series of ordered columns. You can check that the data have loaded in correctly by looking at the first few rows of the columns using the head() function:

```
> head(SES)
```

The R console should now show the first half dozen cases for each variable in the data frame. R can also load data from formats used by common statistics packages such as SPSS, Stata or Minitab. This requires commands from the R add-on package foreign. This is part of the base R installation and can be loaded by entering:

```
> library(foreign)
```

If the data are in an SPSS .sav file the read.spss() function can be used to load the data from the working directory, though the data frame format must be explicitly requested:

```
> SES <- read.spss("SES.sav", to.data.frame=TRUE)
```

The SES data set contains responses to subscales of a technology affinity scale employed by Dunn and Castro (2012). As the overall scale is probably not unidimensional it makes sense to calculate omega separately for the subscales of interest. The following commands extract that variables representing items on just one of the subscales and removes cases with missing item values on this subscale. These variables are in columns 1 to 7 of the data frame. To preserve the original data set, the subscale scores are assigned to a new data frame for the analysis that follows.

```
> subscale1 <- SES[1:7]
> subscale1 <- na.omit(subscale1)
```

It is now possible to calculate omega for this subscale. This is achieved by installing and loading a package specifically designed for this purpose. The package is known as MBESS (Kelley & Lai, 2012). Installing the package requires a live internet connection (and you may be prompted to select a local 'mirror' website to download the package from). The installation step is required the first time you use the package after installing or updating R, whereas the library command is required in every new R session in which you want to compute omega.

```
> install.packages("MBESS", dependencies = TRUE)
> library(MBESS)
```

The `ci.reliability()` function should now be loaded as part of the package and can obtain omega an a bootstrap CI for omega. In the following example the `set.seed()`

function is not required – but using it should allow you to duplicate the output below exactly (by fixing the seed to the random number generator in R). Note that even with the default number of simulations (`B=1000`), bootstrapping the CI may take a few minutes with a large data set or a slow desktop machine.[5]

```
> set.seed(1)

> ci.reliability(data=subscale1, type="omega", conf.level = 0.95,

interval.type="bca", B=1000)
```

The first few lines of the output in the R console window should look like this:

```
$est
[1] 0.9375534

$se
NULL

$ci.lower
[1] 0.9188487

$ci.upper
[1] 0.951895
```

This could be reported as coefficient omega = .94, 95% CI [.92, .95].[6] As the preceding command used the default arguments for type of statistic, number of simulations and

[5] It is also worth noting that other interval types are available (in addition to BC_a) as well as other levels of confidence (e.g., 99%) (see Kelley & Lai (2012) for a comprehensive options list related to these options within the MBESS package).

[6] Note that the standard error of the estimate is not returned because an analytic estimate is not available (though a bootstrap estimate is returned later in the output). The authors of the MBESS package have indicated that they plan to return the bootstrap estimate of the *SE* (rather than `NULL`) in the next update of the package.

confidence level (but not type of confidence interval), the following command should have identical output:

```
> ci.reliability(subscale1, interval.type="bca")
```

It is good practice to check the bootstrapping results with a larger number of simulations (e.g., 10,000). This is done by changing the relevant argument:

```
> ci.reliability(subscale1, interval.type="bca", B=10000)
```

On this occasion increasing the number of bootstrap simulations does not alter the width of the CI if reporting to two decimal places (but takes considerably longer to compute).[7]

Conclusion and recommendations

In an ideal world it would be prudent to ensure that the assumptions about the populations we are sampling are plausible given the data at hand. However, it is more realistic to concede that one must be careful not to violate assumptions that have a material impact on the results of importance (e.g., see Baguley, 2012). In consideration of alpha, the evidence is clear that when certain assumptions are not met there is a substantial adverse impact on its ability to estimate internal consistency reliability of a scale. The impact of poor measurement reliability can compromise a researcher's ability to make inferences or establish practical or clinically significant results (Thompson, 2002).

[7] In many cases 1000 simulations will be sufficient to give a reasonable estimate of the 95% CI. If the lower and upper limits of the interval change across repeated runs of the `ci.reliability()` function it is advisable to increase the number of simulations until the limits are stable to two or more decimal places across repeated runs.

If one is confident that a scale is unidimensional and also that the assumptions of the essentially tau-equivalent model have not been violated then alpha could be applied (albeit with caution). It is clear however that such circumstances are likely to be rare for psychological scales. Furthermore, apparently modest violations of the essentially tau-equivalent model can dramatically bias alpha. Even if alpha is considered appropriate, it is essential to report an interval estimate such as a CI alongside alpha.

Checking that the assumptions of alpha are met is a challenging task. However, there are some simple considerations researchers should review when assessing the likelihood of violating these assumptions. If the *SD* of item scores composing a test are markedly different from one another they are likely to have different true score or error variances (therefore violating one or more assumptions of alpha) (Graham, 2006). Also, if a test uses multiple response formats across items it is also more likely to violate the assumptions of tau-equivalence (Raykov, 1997; Iacobucci & Duhachek, 2003; Graham, 2006). Finally, if samples are commingled (i.e., taken from different populations) then the tau-equivalence assumption is also likely to be untenable.

Given the consensus in the psychometric literature that alpha is rarely appropriate and given the good performance of omega when the assumptions of alpha are not met, it is recommended that psychologists change to the routine reporting of omega in place of alpha. If unidimensionality is uncertain or if any form of multidimensionality is suspected then it is recommended that omega be calculated along with CIs for each subscale comprising the test (Soĉan, 2000).

In this article we have set out the case against routine reporting of coefficient alpha in psychology – a position that we believe reflects a broad consensus among experts. Given the availability of suitable alternatives – notably coefficient omega – and easy access to software capable of calculating accurate point and interval estimates, there is an overwhelming case for a change in practice. Researchers should switch from alpha to omega.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology - replication and extension of Aiken, West, Sechrest and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*, 32-50. doi:10.1037/0003-066X.63.1.32

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., Scarr, S., Sherman, S. J. (1990). Graduate training in statistics, methodology, and - a survey of PhD programs in North America. *American Psychologist*, *45*, 721-734. doi:10.1037/0003-066X.45.6.721

Baguley, T. (2008). The perils of statistics by numbers. *The Psychologist*, *21*, 224-224.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603-617. doi:10.1348/000712608X377117

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi: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. doi:10.1177/0013164404266386

DiCiccio, T. J, & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*, 189-228. doi:10.1002/sim.4134

Dunn, J. T., & Castro A. (2012). Postmodern society and the individual: The structural characteristics of postmodern society and how they shape who we think we are. *The Social Science Journal, 49*, 352-358. doi:10.1016/j.soscij.2012.02.001

Field, A., Miles J. N. V., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.

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. doi:10.1177/0013164406288165

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

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. doi:10.1207/S15328007SEM0702_6

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

Huysamen, G. (2007). Coefficient alpha: Unnecessarily ambiguous; unduly ubiquitous. *SA Journal of Industrial Psychology, 32*, 34-40. Retrieved from http://sajip.co.za/index.php/sajip/article/view/242

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology*, *13*, 478–487. doi:http://dx.doi.org/10.1207/S15327663JCP1304_14

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Kelley, K., & Lai, K. (2012). MBESS: MBESS. R package version 3.3.2. http://CRAN.R-project.org/package=MBESS

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137-152. doi:10.1037/a0028086

Lord, F. I., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Cambridge, MA: Addison-Wesley.

McDonald, R. (1981). The dimensionality of tests and items. *British Journal of Mathematical & Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates. doi:10.1111/j.2044-8317.1981.tb00621.x

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. doi:10.1080/10705519509540013

Nunnally, B. H., J.C. (1994). *Psychometric theory* (3rd Ed.). London: McGraw-Hill.

R Development Core Team. (2012). R: A language and environment for statistical computing Vienna, Austria. http://www.R-project.org/.

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement,* 21, 173–184. doi:10.1177/01466216970212006

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. doi:10.1177/01466216970212006

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement*, *22*, 375–385. doi:10.1177/014662169802200407

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76. doi:10.1177/01466216010251005

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research*, *37*, 89–103. doi:10.1207/s15327906mbr3204_2

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

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8* , 350-353. doi:10.1037/1040-3590.8.4.350

Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences*, *28,* 229–237. doi:http://dx.doi.org/10.1016/S0191-8869(99)00093-8

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

Soĉan, G. (2000). Assessment of reliability when test items are not essentially tau-equivalent. *Advances in Methodology and statistics*, *15*, 23-35. Retrieved from: http://ams.sisplet.org/uploadi/editor/mz15socan.pdf

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. doi:10.1111/j.2044-8317.2011.02030.x

Thompson, B. (2002). “Statistical,” “practical,” and “clinical”: How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71. doi:10.1177/0013164400602002

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195. doi:10.1002/j.1556-6678.2002.tb00167.x

Waller, N. G. (2008). Commingled samples: A neglected source of bias in reliability analysis. *Applied Psychological Measurement*, *32*, 211-223. doi:10.1177/0146621607300860

Woodward, J., & Bentler, P. (1978). A statistical lower bound to population reliability. *Psychological Bulletin*, *85*, 1323-1326. doi:10.1037/0033-2909.85.6.1323

Yuan, K., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika, 67,* 251-259. doi:10.1007/BF02294845

Zimmerman, D., Zumbo, B., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53,* 33-49. doi:10.1177/0013164493053001003

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

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

Zinbarg, R. E., Revelle, W., & Yovel, I. (2007). Estimating ω□ for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement, 31*, 135-157. doi:10.1177/0146621606291558