British Journal of Mathematical and Statistical Psychology (2012), 65, 371–401
© 2011 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals

Leann Terry[1] and Ken Kelley[2]*

[1]Pennsylvania State University, University Park, Pennsylvania, USA
[2]University of Notre Dame, Indiana, USA

Composite measures play an important role in psychology and related disciplines. Composite measures almost always have error. Correspondingly, it is important to understand the reliability of the scores from any particular composite measure. However, the point estimates of the reliability of composite measures are fallible and thus all such point estimates should be accompanied by a confidence interval. When confidence intervals are wide, there is much uncertainty in the population value of the reliability coefficient. Given the importance of reporting confidence intervals for estimates of reliability, coupled with the undesirability of wide confidence intervals, we develop methods that allow researchers to plan sample size in order to obtain narrow confidence intervals for population reliability coefficients. We first discuss composite reliability coefficients and then provide a discussion on confidence interval formation for the corresponding population value. Using the accuracy in parameter estimation approach, we develop two methods to obtain accurate estimates of reliability by planning sample size. The first method provides a way to plan sample size so that the expected confidence interval width for the population reliability coefficient is sufficiently narrow. The second method ensures that the confidence interval width will be sufficiently narrow with some desired degree of assurance (e.g., 99% assurance that the 95% confidence interval for the population reliability coefficient will be less than *W* units wide). The effectiveness of our methods was verified with Monte Carlo simulation studies. We demonstrate how to easily implement the methods with easy-to-use and freely available software.

1. Introduction

In research and practice, the importance of understanding, reporting, and critically evaluating issues of reliability in testing and measurement situations cannot be overstated. Standards call for estimates of relevant reliabilities to be reported for all tests used in research (e.g., Committee on Reviewing Evidence to Identify Highly Effective Clinical

*Correspondence should be addressed to Ken Kelley, Department of Management, University of Notre Dame, Notre Dame, IN 46556, USA. (email: kkelley@nd.edu).

DOI:10.1111/j.2044-8317.2011.02030.x

Services, 2008; Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association (AERA), the American Psychological Association (APA), & the National Council on Measurement in Education (NCME), 1999; Wilkinson & the APA Task Force on Statistical Inference, 1999). As Wilkinson and the APA Task Force on Statistical Inference, 1999, p. 596) pointed out:

> It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees (Feldt & Brennan, 1989). Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. Interpreting the size of observed effects requires an assessment of the reliability of the scores.

It is the test scores on a particular administration of a test, not the test itself, that are reliable – this is a key point often overlooked. It is a widespread misconception that a test itself is reliable. A test, when administered, say, 100 different times, will likely result in 100 different reliability coefficients (Thompson & Vacha-Haase, 2000). The differences among estimated reliability coefficients could be due to different populations among the 100 administrations, sampling error, or both. A meta-analytic method known as reliability generalization was developed by Vacha-Haase (1998) to explore the error variance across studies and helps to evaluate the different estimates from within the same or across different populations. Thus, it is not sufficient to cite the reliability for the scores from the sample from which the test was normed that is provided in the test manual. All researchers should report an estimate of reliability for the scores obtained on their particular administration(s) of the test(s). Indeed, even the passage of time within the same population may lead to a different value of the population reliability coefficient at different time points.

In addition to the recommendation that confidence intervals (CIs) be reported for estimates of population quantities on a variety of parameters (AERA, 2006; APA, 2009; Cohen, 1994; Kline, 2004; Meehl, 1997; Wilkinson & the APA Task Force on Statistical Inference, 1999), calls have recently been made to report the CI specifically for the population value of the reliability coefficient (Duhachek & Iacobucci, 2004; Fan & Thompson, 2001; Zinbarg, Yovel, Revelle, & McDonald, 2006). Yet the question arises why CIs are not often reported for the population value of a reliability coefficient. One answer might lie in a problem plaguing the use of CIs in general: they can be largely uninformative for gauging the population value. Cohen (1994, p. 1002) suggested that the lack of CIs in the literature might be because their widths are often 'embarrassingly large'. These wide CIs illustrate that the reliability estimates obtained may not accurately reflect their corresponding population values. It is the population values of a reliability coefficients that are ultimately of interest, not the value obtained based on an idiosyncratic sample. Another reason why CIs for the population reliability coefficient are not often reported is that they are not discussed much in texts that deal with psychometric issues or general statistical issues in psychology and related disciplines. Because of the importance of CIs and the undesirability of wide CIs, it would be ideal if a method existed so that a researcher could plan an appropriate sample size for a study, such that the CI for the population reliability coefficient were sufficiently narrow.

Using the accuracy in parameter estimation (AIPE) approach to sample size planning (see Maxwell, Kelley, & Rausch, 2008, for a review) this paper provides methods to plan sample size so that: (a) the expected CI width for the population reliability coefficient is sufficiently narrow; and (b) the CI width will be sufficiently narrow with some desired degree of assurance. The desired degree of assurance is the probability of achieving a CI no wider than desired. An application of method (a) would provide the sample size necessary so that the expected CI width for a reliability coefficient would be no wider than specified by the researcher. However, due to the fact that the CI width is a continuous random variable, essentially any computed interval will be less than or greater than the expected width approximately half of the time. This necessitates a method of planning sample size so the CI will be sufficiently narrow with a desired degree of assurance. Method (b) would provide a modified sample size that is larger so that the CI is no wider than specified with any desired degree of assurance (e.g., 99% assurance that the 95% CI for the population reliability coefficient will be sufficiently narrow). These methods are developed for two types of reliability coefficients in the context of homogeneous (i.e., single-construct) tests: coefficient alpha, which assumes true-score equivalence; and coefficient omega, which assumes only a congeneric structure.

First, we review reliability for homogeneous tests and discuss coefficient alpha and coefficient omega from a confirmatory factor-analytic perspective. Second, we discuss CIs for coefficient alpha and coefficient omega. Third, we explain sample size planning from the AIPE perspective and provide sample size tables which are useful under specified conditions for planning studies where narrow CIs for the population reliability coefficient are desired. Fourth, we discuss several possible applications of this approach for reliability coefficients. Fifth, an example is illustrated with the freely available MBESS package (Kelley, 2007a, 2007b; Kelley & Lai, 2011a) for use in the open source program R (R Development Core Team, 2011).[1] The MBESS R package allows the methods discussed to be immediately and easily implemented by researchers.

2. Homogeneous tests as a confirmatory factor model

Homogeneous tests are tests that measure only one attribute (e.g., McDonald, 1999). From a classical test theory (CTT) perspective (e.g., Lord & Novick, 1968) there are three common measurement models for homogeneous tests: (a) parallel; (b) true-score equivalent; and (c) congeneric. McDonald (1999) reframed the classical true-score theory in a confirmatory factor-analytic framework for homogeneous tests. McDonald demonstrated how the confirmatory factor model can be used to represent each of the three models in a 'unified framework'. Following McDonald (1999), throughout this paper the confirmatory factor-analytic approach will be used as a general approach for to conceptualizing reliability. The reliability of a set of scores can be estimated for each type of model, but there are assumptions governing the reliability coefficient for each of the models. The appropriate estimate of reliability thus depends on the assumptions specified and properties of the test as it applies to the population of interest. Figure 1(a) shows a homogeneous factor model for a hypothetical five-item test. As can be seen, the underlying construct (η) has been measured by five items (X_j),

[1] For this paper, all statistical computations were conducted in R (R Core Development Team, 2011). R is a statistics programming language that is open source and extremely flexible (Kelley, Lai, & Wu, 2008). Because of this flexibility, researchers can perform many statistical procedures which are not possible in more mainstream statistical packages. Using R, researchers are able to download more than 2,000 add-on packages that implement advanced statistical techniques.

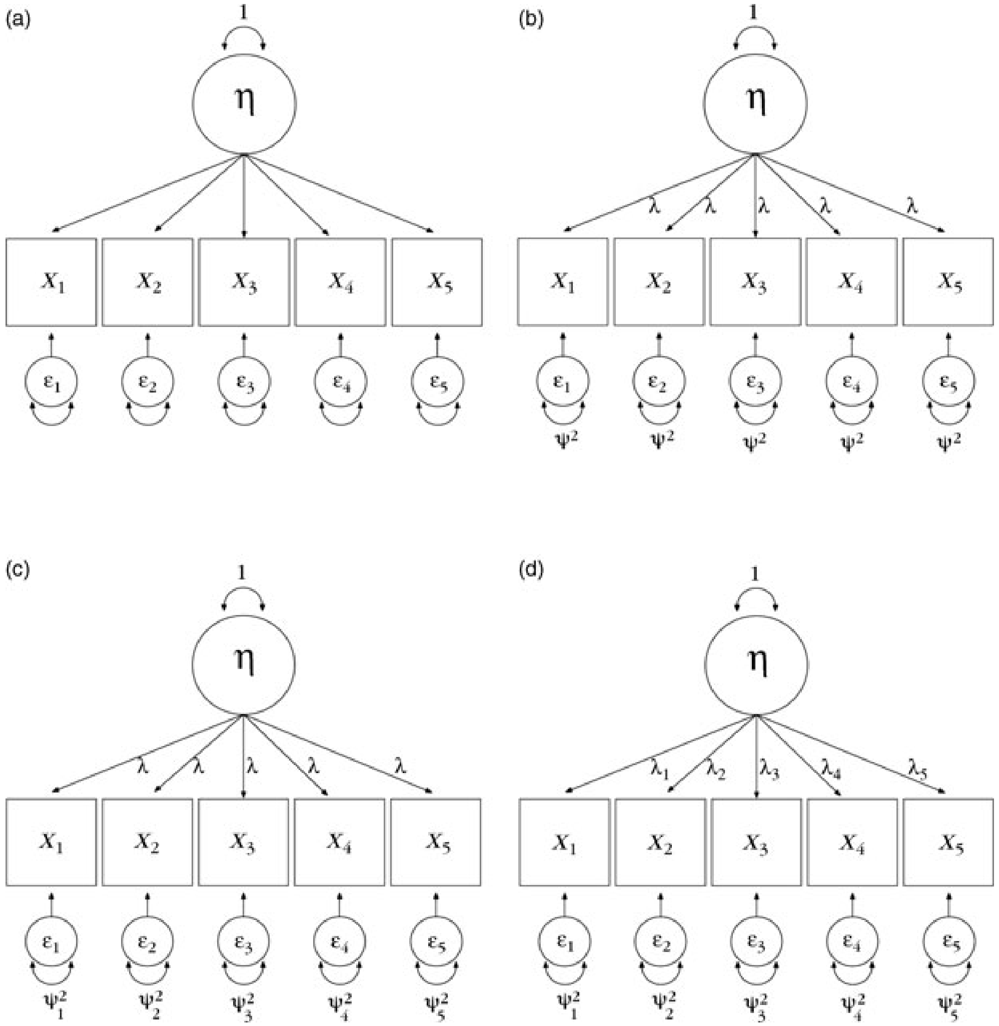

Figure 1. (a) General factor model where η is the latent trait being measured, X is the observed score, and ε is the error. (b) Parallel items model where test items (X) have common error variance (ψ^2) and factor loadings (λ). (c) True-score model where error variances are allowed to differ, but factor loadings are the same across test items. (d) Congeneric model where factor loadings and error variances are allowed to differ across test items.

with each score having a true component ($\eta\lambda$) and an error term (ε_j). In all the models discussed, the errors for the test items are uncorrelated. Although the assumption of uncorrelated errors can be relaxed, we do not consider such a model here. If the errors are correlated, coefficient alpha may overestimate the value of the population reliability (Green & Hershberger, 2000; Komaroff, 1997; Zimmerman, Zumbo, & Lalonde, 1993).

The parallel items model (see Figure 1(b)) is the most restrictive model, which assumes that the test item scores share a common error variance (ψ^2) and a common factor loading (λ), implying there are only two parameters, regardless of the number of test items (see Hoyt, 1941; McDonald, 1999; Raykov, 1997; Spearman, 1910, for a discussion). Because all the test item scores have a common factor loading, this means that no test item has more or less discriminating power than any other test item (McDonald, 1999). Additionally, as the test items share a common error variance, all the covariances among the test items are equal to one another. In practice, it is highly unlikely that the assumptions of the parallel model are met in psychology and related disciplines (e.g., Green & Yang, 2009). In most situations, we believe that it is likely that the population factor loadings and the population error variances differ for the J different test items. Correspondingly, we do not discuss the parallel model further in this paper.

In the true-score equivalence model (also called essentially tau-equivalent; see Figure 1(c)) the error variances are allowed to differ (noted with each error variance having a corresponding subscript) but the factor loadings are restricted to be the same (Lord & Novick, 1968; McDonald, 1999). Because the factor loadings do not vary, this indicates that each test item measures the common factor with equal discriminating power (McDonald, 1999).

The congeneric model (see Figure 1(d)) is the least restrictive, and we believe generally most reasonable, model we discuss, in which the error variances and factor loadings can be unique to each test item (depicted by the corresponding subscripts). This implies that some test items measure the attribute of interest more sensitively and discriminate more clearly between higher and lower values of the common factor (McDonald, 1999).

3. Reliability

In the CTT framework, each test item score on a particular test is composed of a true-score component and an error component. In CTT, the score on the jth ($j = 1, \ldots, J$) test item is given by

$$X_{ij} = \tau_i + \varepsilon_{ij}, \tag{1}$$

where τ_i is the true-score for the ith individual and ε_{ij} the error for the ith individual on the jth item (Lord & Novick, 1968; McDonald, 1999); throughout we assume centred scores for simplicity (i.e., the mean of the item scores across the individuals is zero). Within the CTT framework, the reliability of the scores on a homogeneous test is the ratio of the true variance to the sum of the true variance and error variance, which is given as

$$\frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}, \tag{2}$$

where σ_τ^2 is the population true-score variance and σ_ε^2 is the population error variance. Given this, a low reliability estimate indicates that a relatively large proportion of the variance is due to error.

3.1. Reliability coefficients

One of the most common measures of reliability is coefficient alpha (e.g., Cortina, 1993; Cronbach, 1951; Guttman, 1945; Hogan, Benjamin, & Brezinski, 2000). Coefficient alpha is defined in the population as

$$\alpha_c \equiv \frac{J}{J-1}\left(1-\frac{\sum_{j=1}^{J}\sigma_j^2}{\sigma_Y^2}\right), \tag{3}$$

where α_c is coefficient alpha, J is the number of test items, σ_j^2 is the variance of each test item's scores, and σ_Y^2 is the variance of the total scores on the test. The common estimate for alpha in a sample is

$$\hat{\alpha}_c \equiv \frac{J}{J-1}\left(1-\frac{\sum_{j=1}^{J}s_j^2}{s_Y^2}\right), \tag{4}$$

where $\hat{\alpha}_c$ is the estimate of coefficient alpha, s_j^2 is the unbiased estimate of the variance of the scores of each test item, and s_Y^2 is the unbiased estimate of the variance of the entire test. The population value of coefficient alpha equals the true population reliability only when the test items fulfil the requirements of the true-score equivalence model; otherwise it is necessarily a lower bound on the population reliability (see Green & Yang, 2009; Revelle & Zinbarg, 2009; and Sijtsma, 2009, for a review of coefficient alpha). Despite the possibility of underestimating the population reliability when true-score equivalence does not hold, coefficient alpha is the most widely used statistic to report reliability for a set of scores. Its continued and widespread use is probably due to several of its properties, namely, that it (a) often represents a lower bound on reliability and is thus 'conservative' (Lord & Novick, 1968; McDonald, 1999), (b) does not require more than one rater or multiple administrations of the test (Streiner, 2003), (c) is computationally easy to compute, (d) has a long history in psychology and related disciplines, and (e) is included in widely available statistical software packages.

Coefficient omega is a generalization of coefficient alpha in that it does not require true-score equivalence, only a congeneric structure (i.e., a homogeneous factor structure with potentially unique path coefficients and error variances). For a homogenous test, coefficient omega is given by

$$\omega = \frac{\left(\sum_{j=1}^{J}\lambda_j\right)^2}{\left(\sum_{j=1}^{J}\lambda_j\right)^2+\sum_{j=1}^{J}\psi_j^2}, \tag{5}$$

where λ_j is the factor loading for the *j*th test item and ψ_j^2 is the error variance for the *j*th test item (McDonald, 1999). With a homogeneous measurement test (i.e., with a single common factor) this can be rewritten as

$$\omega = 1 - \frac{\sum_{j=1}^{J} \psi_j^2}{\sigma_Y^2}, \tag{6}$$

where σ_Y^2 is the variance of the total test scores (McDonald, 1999, equation (6.21)). Interested readers are referred to Graham (2006), Lucke (2005), and McDonald (1999, Chapter 6), for a thorough discussion of coefficient omega, and Zinbarg, Revelle, Yovel, and Li (2005) and Revelle and Zinbarg (2009) for a discussion of ω_h, which is a generalization of omega to the hierarchical factor model.

3.2. Confidence intervals for reliability coefficients

Many methodologists, professional societies, and scientific organizations emphasize the importance of reporting and discussing CIs for population values. Correspondingly, CIs for population reliability coefficients should always be reported. We provide an example to highlight the importance of reporting CIs for the population reliability coefficients. Suppose a researcher calculates coefficient alpha as .74 for a set of scores on one administration of a test with a sample of 500 participants from some population of interest. A second researcher calculates coefficient alpha of .74 for a set of scores, with the same test, administered with a sample of 50 participants from a different population.[2] When comparing these findings, one might assume at first that the population reliability values in the two samples are the same. However, if the researchers reported CIs for the corresponding population reliability estimates, a different picture emerges with regard to the population values. The first estimated reliability coefficient has a corresponding 95% CI that ranged from .71 to .77. The second researcher might be surprised to find a corresponding 95% CI from .63 to .85. Such a wide interval in the second population illustrates that the range of plausible parameter values is large and demonstrates the uncertainty in the estimate of the population value of alpha. Thus, although the point values for the reliability coefficients were identical, the wide CIs reveal that the estimates were estimated with different degrees of accuracy.[3] Consequently, despite the point estimate being identical, the reliability estimates should be interpreted and compared differently. Holding everything else constant, a narrower CI yields a smaller range of plausible parameter values, which corresponds to a more accurate estimate of the population value (Kelley & Maxwell, 2003).

[2] A sample size of 50 falls within the range of median sample sizes found in the *American Educational Research Journal* from 1988 to 1997 (which ranged from a median sample size of 43 to 169; Kieffer, Reese, & Thompson, 2001). A sample size of 500 is more likely to be found in a test manual, rather than in published research.

[3] In a statistical sense, accuracy is defined as the square root of the mean square error (RMSE), which is a function of both precision and bias. RMSE is defined as $\sqrt{E[(\hat{\theta} - \theta)^2]} = \sqrt{E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta} - \theta])^2} = \sqrt{\sigma_{\hat{\theta}}^2 + B_{\hat{\theta}}^2}$, where $E[\cdot]$ represents the expectation of the bracketed quantity, θ is the parameter of interest, $\hat{\theta}$ is an estimate of θ, $\sigma_{\hat{\theta}}^2$ is the population variance of the estimator, and $B_{\hat{\theta}}^2$ is the squared bias of the estimator.

3.2.1. True-score model

Advances in forming CIs for coefficient alpha and the true-score model were possible with van Zyl, Neudecker and Nel's (2000) derivation of the asymptotic normal distribution of the maximum likelihood estimator (MLE) of coefficient alpha. Maydeu-Olivares, Coffman and Hartmann (2007) compared asymptotically distribution-free (ADF) versus normal-theory (NT) CI estimation (Maydeu-Olivares *et al.*, 2007) and concluded that ADF CIs performed better empirically than did NT CIs when coefficient alpha was the appropriate measure of reliability. However, they pointed out that the methods were impossible to differentiate when the sample size was smaller than 400 and item skewness was less than 0.5 (Maydeu-Olivares *et al.*, 2007). They concluded that NT CIs can be safely used when test items are normally distributed and with relatively small sample sizes (i.e., less than 100). Otherwise, ADF CIs should be used when item scores are not normal or with larger sample sizes (Maydeu-Olivares *et al.*, 2007). The NT approach is used in this paper for coefficient alpha in the true-score model.

Van Zyl *et al.* (2000) provided the asymptotic distribution of the MLE of coefficient alpha, such that as $n \rightarrow \infty$ the variance is

$$\frac{J^2}{(J-1)^2}\zeta, \tag{7}$$

where J is the number of responses, and ζ is defined as

$$\zeta = \frac{2}{(\mathbf{g}'\boldsymbol{\Phi}\mathbf{g})^3}[(\mathbf{g}'\boldsymbol{\Phi}\mathbf{g})(\mathrm{tr}\boldsymbol{\Phi}^2 + \mathrm{tr}^2\boldsymbol{\Phi}) - 2(\mathrm{tr}\boldsymbol{\Phi})(\mathbf{g}'\boldsymbol{\Phi}^2\mathbf{g})], \tag{8}$$

in which $\mathbf{g}$ is a $J \times 1$ column vector of ones, and $\boldsymbol{\Phi}$ is the population covariance matrix of the J responses (van Zyl *et al.*, 2000; notation changed to reflect current usage). A two-sided CI for coefficient alpha under true-score equivalent assumptions is given by

$$\alpha_c \pm z_{1-\alpha_e/2}\sqrt{\frac{\left(J^2/(J-1)^2\right)\zeta}{N-1}}, \tag{9}$$

where α_e is the Type I error rate, $z_{1-\alpha_e/2}$ is the quantile from the standard normal distribution, and N is the total sample size. Thus, the full width of the $(1-\alpha_e)100\%$ CI for coefficient alpha can be shown to equal

$$w = 2z_{1-\alpha_e/2}\sqrt{\frac{\left(J^2/(J-1)^2\right)\zeta}{N-1}}. \tag{10}$$

3.2.2. Congeneric model

In addition, several methods exist for estimating CIs for the population value of coefficient omega and the congeneric model, including an analytic formula from Raykov (2002), which uses the delta method for multivariate normally distributed item scores. The analytic version of forming a CI for the population value of coefficient omega used in this paper utilizes Raykov's (2002) approach (see also Kelley & Cheng, in press, for a discussion of the approach and methods of implementation). Using the delta method, Raykov (2002) discussed an analytic CI for ω that is asymptotically correct for multivariate normally distributed test item scores, as sample size grows towards infinity. The delta method is a way to obtain a standard error for a function of one or more parameter estimates, which can then be used for CIs (e.g., Casella & Berger, 2002; Oehlert, 1992). The method produces asymptotically correct CIs, where 'asymptotically correct' refers to the CI procedure actually producing $(1 - \alpha_e)100\%$ CIs as sample size approaches infinity. This implies that for infinite sample size the procedure is 'approximately correct'. This issue, however, is not unique to CIs for the population reliability coefficients, but rather is generally the case for CIs and null hypothesis probability values in the context of structural equation modelling, which is based on asymptotic estimation theory but implemented in situations of finite sample size (e.g., Bollen, 1989).

Raykov (2002) provided the technical details of the delta method as it applies to CIs for the population reliability coefficients, but we give an overview and explanation here. Using the parameters from the homogeneous congeneric factor model, let

$$\upsilon = \sum_{j=1}^{J} \lambda_j \tag{11}$$

and

$$\nu = \sum_{j=1}^{J} \psi_j^2. \tag{12}$$

Coefficient omega (equation (5)) can be written as

$$\omega = \frac{\upsilon^2}{\upsilon^2 + \nu}. \tag{13}$$

Let

$$\Delta_1 = \frac{2\upsilon\nu}{(\upsilon^2 + \nu)^2} \tag{14}$$

and

$$\Delta_2 = -\frac{\upsilon^2}{(\upsilon^2 + \nu)^2}. \tag{15}$$

The standard error of the reliability coefficient, Raykov's (2002) equation 12, can be written as

$$SE(\hat{\rho}(Y)) = \left[\hat{\Delta}_1^2 \mathrm{Var}(\hat{\upsilon}) + \hat{\Delta}_2^2 \mathrm{Var}(\hat{\nu}) + 2\hat{\Delta}_1\hat{\Delta}_2 \mathrm{Cov}(\hat{\upsilon}\,\hat{\nu})\right]^{1/2}, \tag{16}$$

where $\hat{\upsilon}$, $\hat{\nu}$, $\hat{\Delta}_1$, and $\hat{\Delta}_2$ are the estimates for υ, ν, Δ_1, and Δ_2, respectively. Thus, a $(1 - \alpha_e)100\%$ CI for an estimate of coefficient omega is given by

$$p\left[\hat{\omega} - z_{1-\alpha_e/2} SE(\hat{\omega}) \leq \rho(Y) \leq \hat{\omega} + z_{1-\alpha_e/2} SE(\hat{\omega})\right] = 1 - \alpha_e/2, \tag{17}$$

where $SE(\hat{\omega})$ is obtained from equation (16) and is calculated using the estimates from the homogeneous factor model. Although equation (17) seems straightforward, estimation of $\mathrm{Var}(\hat{\upsilon})$, $\mathrm{Var}(\hat{\nu})$, and $\mathrm{Cov}(\hat{\upsilon}, \hat{\nu})$ is somewhat involved. Estimation details are provided in Kelley and Cheng (in press). The CI methods have been implemented in the MBESS R package (Kelley & Lai, 2011a) for calculating a CI for the population value of coefficient omega. We later provide an example demonstrating the ease of implementing the CI procedure with MBESS.

4. AIPE for sample size planning

AIPE is an approach to sample size planning where the goal is to achieve parameter estimates that are accurate, which is operationalized by obtaining CIs that are sufficiently narrow (Kelley & Maxwell, 2003; Kelley & Rausch, 2006; Maxwell *et al.*, 2008). Because there is now an expectation that CIs be reported for empirical research, researchers need methods to help plan the necessary sample size for sufficiently narrow CIs, where the estimates of reliability will have a high degree of expected accuracy. In a statistical sense, accuracy is defined as the square root of the mean square error, which is a function of both precision and bias (see footnote 4 for its formal definition). Precision is inversely related to the variance of the estimator, and bias is the systematic discrepancy between an estimator and the parameter it estimates. For unbiased estimates, precision and accuracy are equivalent concepts because the bias is zero. The expected accuracy of the estimate improves (i.e., the square root of the mean square error is reduced) as the CI width decreases and the estimate is contained within a narrower set of plausible parameter values. Thus, holding everything else constant, when the width of the $(1 - \alpha_e)100\%$ CI decreases, the expected accuracy of the estimate improves (see Maxwell *et al.*, 2008).

To decrease the width of the CI for the population value of coefficient alpha, test developers can increase the number of items on the test or increase the correlations among the test items (Duhachek & Iacobucci, 2004). However, an applied researcher cannot control the number of test items or the correlations with established scales. Further, it can become difficult to add more and more test items to a homogeneous scale such that the scale remains homogeneous with the additional items (and reasonably short so as not to overburden the test takers). To decrease the width of the CI for the population coefficient alpha, researchers may increase the sample size on a particular administration of a test. As sample size increases, the covariance among the item scores becomes more stable and the expected width of the CI decreases (Iacobucci & Duhachek, 2003). However, there are contradictory recommendations regarding the number of participants necessary for sufficiently accurate reliability estimates, with ranges from a minimum of 30 (provided inter-item score reliability is high; Iacobucci & Duhachek, 2003) to 400 (Charter, 1999). Rules of thumb for planning sample size are not useful in this case, because the sample size depends on several factors and specific goals that are not considered in the rules of thumb. Thus, the best approach for an applied researcher is to plan the size of the sample in a way that explicitly considers the width for the CI of the population reliability coefficient.

Several authors have provided formulas to plan sample size for reliability studies (Bonett, 2002; Iacobucci & Duhachek, 2003). Bonett (2002) provided formulas for estimating coefficient alpha with a desired CI width as a function of sample size. Bonett's (2002) work is for tests that assume compound symmetry (i.e., have equal variances and equal covariances), or are under the parallel test items model. As shown in Figure 1(b), this is the most restrictive model for reliability. Nevertheless, Bonett's (2002) method should be used in a situation where the assumptions of the parallel test items model fit the test and a sufficiently narrow CI is desired. Iacobucci and Duhachek (2003) also provided a method to solve for the necessary sample size to obtain an expected CI width (their equation 6, p. 483) under the assumptions of multivariate normality and large sample size. Their work is based on the assumptions discussed in van Zyl *et al.* (2000) and applies to the true-score equivalent model. However, neither of these methods utilizes an assurance parameter that allows sample size to be planned such that a sufficiently narrow CI will be observed with a researcher-specified degree of assurance. Bonett's (2002) approach for the parallel model functionally incorporates a near 100% assurance parameter. This may be too large for some instances, analogously to the way that sample size for near 100% power may be too large for some purposes.

As of yet, no known sample size planning methods for narrow CIs have been developed for coefficient omega, despite the recommendations and findings that coefficient omega is often more appropriate than coefficient alpha (Lucke, 2005; Zinbarg *et al.*, 2005). We address these existing limitations for both coefficient alpha and coefficient omega by providing sample size planning methods from the AIPE perspective that apply to true-score equivalence (when alpha is appropriate) and congeneric models (when omega is appropriate). The methods we develop are for an expected CI width that is sufficiently narrow as well as for the CI to be sufficiently narrow with some desired degree of assurance. In line with the recommendation of Revelle and Zinbarg (2009) for psychometric contributions to be available in open source software, we implement our methods in the program R, which is detailed in the illustrative example at the end of this paper. Additionally, the methods we develop can be implemented in other software packages/programs.

5. Sample size planning for reliability coefficients

Two AIPE procedures have been developed for sample size planning for the true-score and congeneric models. The goal of the first procedure for each type of reliability coefficient is to plan *a priori* an appropriate sample size so that the *expected* CI width is sufficiently narrow. In this procedure the researcher specifies a desired CI width (W) and our methods provides the minimum sample size necessary so that the expected CI width ($\mathrm{E}[w]$) will be no larger than the desired CI width (W).[4]

However, the mere fact that the expected CI width is sufficiently narrow does not imply that the observed CI will be sufficiently narrow. In (hypothetical or actual) repeated samplings, approximately half of the time the computed CI will be wider than desired and approximately half of the time narrower than desired. This is the case because the CI width, like the estimate of reliability itself, is a random variable and the expected width is the mean width. We develop a second procedure so that there will be a desired degree of assurance that the CI width is sufficiently narrow (e.g., 99% assurance that the

[4] However, because the theoretical sample size where $\mathrm{E}[w] = W$ is almost always a fractional value, $\mathrm{E}[w]$ is almost always just less than W in order for the necessary sample size to be a whole number. This is the case because sample size must increase following a step-function, whereas confidence interval width theoretically increases following a continuous function.

95% CI will be no wider than *W*). This second procedure uses an *a priori* Monte Carlo method to plan the appropriate sample size.

5.1. True-score equivalence model

This section uses the AIPE approach to provide a method for sample size planning for sufficiently narrow CIs for the population coefficient alpha under the true-score model.

5.1.1. Expected width

The full width of the CI for coefficient alpha is

$$w = 2z_{1-\alpha_e/2}\sqrt{\frac{\left(J^2/(J-1)^2\right)\hat{\zeta}}{N-1}}. \tag{18}$$

This equation can be solved for N. Suppose we replace w in equation (18) with W, the desired CI width, and also replace $\hat{\zeta}$ with its population value, ζ. Solving that equation for the sample size results in the necessary N so that, if that sample size were used in a population as described (i.e., where ζ is the true value) the CI width would be expected to be approximately W:

$$N_{\text{nec}} = \frac{\left(J^2/(J-1)^2\right)\zeta}{\left(W/2z_{1-\alpha_e/2}\right)^2} + 1. \tag{19}$$

The result of solving equation (19) is N_{nec} – the procedure-implied (i.e., necessary) sample size such that the CI for the population coefficient alpha will be expected to be sufficiently narrow (i.e., have a width of approximately *W*). As can be seen, the Type I error rate, the desired CI width, the number of test items, and the covariance matrix (as a function of the factor loadings and error variances) all influence the necessary sample size for a desired CI width. Equation (19) can thus be used to determine the sample size necessary so that the expected width of the CI for the population coefficient alpha is sufficiently narrow.

5.1.2. Evaluation of the sample size method for the expected width procedure

A Monte Carlo simulation, commonly used in methodological works to assess the robustness of a procedure, was conducted to ensure that equation (19) appropriately and consistently yields accurate estimates of the necessary sample size to achieve the desired width.

In the Monte Carlo simulation study, a total of 123 different conditions were evaluated to assess the effectiveness of equation (19). Across the different conditions a variety of factors were specified: the numbers of test items (*J*) were 3, 5, 7, 9, 11, 13, 15; the estimated inter-item score correlation (ρ) ranged from .1 to .6 in increments of .1; and a desired CI width (*W*) of .05, .1, .15 and .2. These conditions were chosen because of their reasonableness with what is used and anticipated to be used in the applied literature. Our upper bound for the number of test items is slightly larger than has been recommended in other studies. For instance, Iacobucci and Duhachek (2003) recommended using up to 10 test items (due to the lack of longer scales in use and the difficulty creating a unidimensional measure with more test items). However, in order to simulate the full range of potential number of test items that researchers might use, we decided to, if anything, err on the side of using a greater number of conditions. Each condition in the simulation was based on 10,000 replications. Conditions were not included in the simulation if the determinant of the population covariance was less than 0.000001 or if the necessary sample size was less than 30.[5] This resulted in 45 conditions not being applicable, leaving 123 conditions in the simulation from the original total for the fully crossed design of 168.

Values for the population factor loading (λ) and population error variances (ψ^2) were needed to simulate covariance matrices fitting a specific true-score model that conforms to our conditions. Recall Figure 1(c) in which the error variances could differ across test item scores but the factor loadings were the same. To provide for a common measurement unit, the variance among the test items was standardized to equal one for the population quantities. Consequently, ψ_j^2 was calculated as $1 - \lambda^2$ for the jth test item, and then increased or decreased systematically for the remaining J - 1 test items in the condition. For example, in the conditions with $J = 15$, the population error variances were modified from the standardized variance $(1 - \lambda^2)$ by $\pm.025$, $\pm.05$, $\pm.075$, $\pm.1$, $\pm.125$, $\pm.15$, and $\pm.20$. The population error variances in conditions with $J > 5$ did not extend beyond the ±.2 upper and lower bounds, which were established in the conditions with fewer test items, so as not to have a major discrepancy between population error variances across the number of test items. Producing population covariance matrices in this way allowed us to generate sample data where the true-score model assumptions were met.

The factor loading was calculated from the estimated mean population inter-item score correlation, such that $\lambda = \sqrt{\rho}$. The range of inter-item score correlations varies across tests. Osburn (2000) identified tests such as cognitive ability subtests, supervisor ratings, and personality dimensions as types of tests that have relatively large correlations (e.g., .5). Tests that have lower correlations (e.g., .25) include those measuring personality facets, perception, organizational commitment, and role ambiguity. Thus, in our study a range of inter-item score correlations were included to be consistent with the variety of tests commonly used in psychology and related disciplines.

In the Monte Carlo simulation study, a population covariance matrix was formed for each of the conditions. Using the known population values as input parameters for the sample size planning procedure, the necessary sample size was calculated from equation (19). Random multivariate normal data were generated that conformed to the specific conditions with the number of cases generated based on the method-implied sample size. The `mvrnorm()` function from the R package MASS (Venables & Ripley, 2002) was used to generate the multivariate normal data. The CI was calculated using equation (9) by way of the `ci.reliability()` function from the R package MBESS (Kelley & Lai, 2011a) on the simulated data. This process, generating multivariate normal data meeting the specific conditions, was repeated 10,000 times for each of 123 conditions we examined. This then enabled us to calculate the mean and median of the observed widths, and other descriptive statistics, to compare them with the desired

[5] Covariance matrices with a permissible structure have a determinant that is greater than zero. This implies that one row/column cannot be written as a linear function of other rows/columns. Because we are working with arbitrary matrices that are square and symmetric, it is possible to define one such that it represents a non-full-rank matrix.

width (W) to evaluate if equation (19) appropriately and consistently yields accurate estimates of the necessary sample size to achieve the desired width.

The results of the simulation indicate that the sample size formula worked very well. Table 1 gives the necessary sample size, mean, and median observed CI width from each condition of the analytic process verifying equation (19). The percentage of error in each condition was calculated to demonstrate the effectiveness of the procedure. The percentage of error is the difference between the observed mean w for each condition and W, divided by W and then multiplied by 100. Across all conditions, the mean percentage of error was 1.00%, the median was 0.55%, and the standard deviation of the percentage of error was 1.12%. The condition with $W = .1$, $\rho = .5$, $J = 9$, and $N_{nec} = 34$ performed the worst (as calculated by the condition with the largest percentage of error), with a percentage of error of 4.61%. This comes from the difference between the desired width ($W = .1$) and the mean of the observed widths ($w = .1046$) of .0046. The relative lack of success of the method in this condition is due to the small sample size resulting from the procedure.

5.1.3. Incorporating an assurance parameter through an a priori Monte Carlo simulation procedure

To provide a desired degree of assurance that the observed CI for the population coefficient alpha will be no wider than desired, a modified sample size planning procedure is developed here using an *a priori* Monte Carlo simulation.[6] An *a priori* Monte Carlo study uses the simulation-based approach of a traditional Monte Carlo simulation, as in the previous section, but can be used to assess the effects of sample size on properties of statistical outcomes. That is, it generates conditions believed to be true and evaluates various statistical properties in those conditions, which differs from using analytic methods on those same supposed conditions.

As Maxwell *et al.* (2008, p. 553) stated as a general rule, 'sample size can be planned for any research goal, on any statistical technique, in any situation with an a priori Monte Carlo simulation study' (see Muthén & Muthén, 2002, for an application of this method in the context of structural equation models). The basic steps in this process include: (a) generating random data with the appropriate assumptions satisfied at a particular sample size; (b) performing a statistical technique of interest (CIs for reliability coefficients under the true-score equivalence assumptions in this case); (c) repeating it a large number of times (e.g., 10,000); (d) evaluating whether the outcome of interest has been satisfied to determine if a particular value of sample size is appropriate; and (e) systematically adjusting the sample size of the randomly generated data. Steps (a)–(e) are repeated until the specified goal has been reached (Maxwell *et al.*, 2008). The idea is to perform a simulation study to discern empirical properties of the statistic of interest under the specified conditions.

Using a simulation study in this way allows us to develop a method to incorporate a desired degree of assurance (γ), as no formal analytic method is known to exist or reasonably developed. Using this method, the CI obtained will be no wider than $W$with no less than γ100% assurance. This assurance parameter provides a probabilistic component to the sample size planning procedure that satisfies the following inequality:$p(w \leq W) \geq \gamma$. For example, a researcher may want to have 85% assurance that the obtained 95% CI for

[6] An analytic based approach was tested, but it was not effective across the spectrum of possible conditions. Thus, a computer-intensive simulation approach was used.

Table 1. Sample sizes necessary for a specified width for 95% CIs for coefficient alpha in the true-score model simulation with mean and median observed CI widths

		J						
ρ		3	5	7	9	11	13	15
					$W = 0.05$			
0.1	Mean	0.0500	0.0500	0.0500	0.0500	0.0501	0.0500	0.0500
	Median	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
	$N_{\text{Necessary}}$	10275	6280	4491	3411	2705	2202	1828
0.2	Mean	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
	Median	0.0500	0.0500	0.0500	0.0501	0.0500	0.0501	0.0501
	$N_{\text{Necessary}}$	5913	2977	1864	1282	945	725	574
0.3	Mean	0.0500	0.0501	0.0501	0.0501	0.0501	0.0502	0.0503
	Median	0.0500	0.0500	0.0500	0.0500	0.0500	0.0499	0.0500
	$N_{\text{Necessary}}$	3436	1514	876	571	405	303	234
0.4	Mean	0.0500	0.0501	0.0502	0.0502	0.0504	0.0503	0.0505
	Median	0.0500	0.0500	0.0500	0.0499	0.0500	0.0499	0.0499
	$N_{\text{Necessary}}$	1976	790	434	273	190	140	107
0.5	Mean	0.0501	0.0501	0.0504	0.0505	0.0504	0.0510	0.0513
	Median	0.0500	0.0499	0.0500	0.0500	0.0496	0.0499	0.0498
	$N_{\text{Necessary}}$	1100	409	217	133	92	67	51
0.6	Mean	0.0501	0.0503	0.0503	0.0510	0.0515	0.0518	–
	Median	0.0500	0.0499	0.0496	0.0499	0.0498	0.0494	–
	$N_{\text{Necessary}}$	575	202	105	63	43	32	–
					$W = 0.10$			
0.1	Mean	0.1001	0.1000	0.1001	0.1002	0.1002	0.1003	0.1003
	Median	0.1000	0.0999	0.1000	0.1000	0.0999	0.1000	0.1000
	$N_{\text{Necessary}}$	2570	1571	1124	854	677	552	458
0.2	Mean	0.1001	0.1003	0.1003	0.1004	0.1005	0.1006	0.1006
	Median	0.1000	0.1002	0.1000	0.1000	0.0999	0.0998	0.0996
	$N_{\text{Necessary}}$	1479	745	467	322	237	182	145
0.3	Mean	0.1002	0.1004	0.1006	0.1010	0.1017	0.1018	0.1023
	Median	0.1000	0.0999	0.0999	0.0999	0.1002	0.0997	0.0998
	$N_{\text{Necessary}}$	860	380	220	144	102	77	60
0.4	Mean	0.1002	0.1004	0.1011	0.1020	0.1025	0.1041	–
	Median	0.0998	0.0996	0.0999	0.0998	0.0997	0.1001	–
	$N_{\text{Necessary}}$	495	199	110	69	49	36	–
0.5	Mean	0.1004	0.1010	0.1024	0.1046	–	–	–
	Median	0.0998	0.0998	0.0996	0.1001	–	–	–
	$N_{\text{Necessary}}$	276	103	55	34	–	–	–
0.6	Mean	0.1005	0.1020	–	–	–	–	–
	Median	0.0993	0.0989	–	–	–	–	–
	$N_{\text{Necessary}}$	145	52	–	–	–	–	–
					$W = 0.15$			
0.1	Mean	0.1499	0.1504	0.1505	0.1505	0.1503	0.1506	0.1511
	Median	0.1498	0.1501	0.1502	0.1499	0.1496	0.1498	0.1501
	$N_{\text{Necessary}}$	1143	699	500	380	302	246	204
0.2	Mean	0.1500	0.1504	0.1510	0.1515	0.1520	0.1521	0.1535
	Median	0.1496	0.1496	0.1499	0.1496	0.1500	0.1493	0.1500
	$N_{\text{Necessary}}$	658	332	208	144	106	82	65

(continued)

Table 1. Continued

		J						
ρ		3	5	7	9	11	13	15
					$W = 0.15$			
0.3	Mean	0.1505	0.1508	0.1519	0.1528	0.1550	0.1562	–
	Median	0.1499	0.1492	0.1500	0.1494	0.1500	0.1497	–
	$N_{\text{Necessary}}$	383	170	99	65	46	35	–
0.4	Mean	0.1506	0.1520	0.1534	0.1554	–	–	–
	Median	0.1494	0.1494	0.1492	0.1481	–	–	–
	$N_{\text{Necessary}}$	221	89	50	32	–	–	–
0.5	Mean	0.1509	0.1531	–	–	–	–	–
	Median	0.1490	0.1481	–	–	–	–	–
	$N_{\text{Necessary}}$	124	47	–	–	–	–	–
0.6	Mean	0.1530	–	–	–	–	–	–
	Median	0.1488	–	–	–	–	–	–
	$N_{\text{Necessary}}$	65	–	–	–	–	–	–
					$W = 0.20$			
0.1	Mean	0.2004	0.2007	0.2011	0.2008	0.2014	0.2019	0.2012
	Median	0.1997	0.2002	0.2001	0.1996	0.1998	0.2002	0.1989
	$N_{\text{Necessary}}$	644	394	282	215	170	139	116
0.2	Mean	0.2006	0.2017	0.2025	0.2021	0.2047	0.2043	0.2083
	Median	0.1997	0.2001	0.2001	0.1986	0.1994	0.1980	0.2002
	$N_{\text{Necessary}}$	371	187	118	82	60	47	37
0.3	Mean	0.2014	0.2027	0.2056	0.2090	–	–	–
	Median	0.1999	0.1988	0.2001	0.2008	–	–	–
	$N_{\text{Necessary}}$	216	96	56	37	–	–	–
0.4	Mean	0.2011	0.2048	–	–	–	–	–
	Median	0.1984	0.1989	–	–	–	–	–
	$N_{\text{Necessary}}$	125	51	–	–	–	–	–
0.5	Mean	0.2038	–	–	–	–	–	–
	Median	0.1990	–	–	–	–	–	–
	$N_{\text{Necessary}}$	70	–	–	–	–	–	–
0.6	Mean	0.2058	–	–	–	–	–	–
	Median	0.1960	–	–	–	–	–	–
	$N_{\text{Necessary}}$	37	–	–	–	–	–	–

Note. J is the number of items on the test, ρ is the estimated inter-item correlation, W is the desired CI width, the means and medians refer to the observed CI widths from the 10,000 simulations in each condition, $N_{\text{Necessary}}$ is the necessary sample size calculated from equation (19), and – indicates the conditions did not meet the necessary criteria. Multiple error variances were simulated to obtain these results – see the text for how the variances were calculated.

the population coefficient alpha will be no wider than .10. Thus, γ is .85, and W is .10. The idea is that the observed CI width will be wider than .10 no more than 15% of the time.

The same 123 conditions that were used in the study evaluating the expected width were used in the simulation to plan sample size for coefficient alpha in the true-score model, providing assurance that the CI will be no wider than desired a certain percentage of the time. For an example of this procedure, if the parameters entered in one condition are ($W = .15$, $\rho = .16$, $J = 5$, $\gamma = .85$) with population error variances of .20, .25, .30,

.35, and .40, then equation (19) is used to provide an initial sample size estimate and results in an N_{nec} of 123. Then multivariate normal data for this condition is generated 10,000 times based on these parameters entered (from the population data and the initial sample size of 123). Suppose that, out of 10,000 replications, in 4,890 of the replications the CI was no wider than .15. As seen, 48.9% of the time the CI was appropriately narrow, which we define as the empirical assurance. However, because the desired degree of assurance was .85 the sample size of 123 did not lead to enough of the CIs being appropriately narrow (as only 48.9% were appropriately narrow), and the process is repeated by increasing the sample size by 1.[7] Then, 10,000 more replications are evaluated at the increased sample size. This iterative process continues until the modified sample size, N_{mod} is found where the observed CI width is no wider than desired $\gamma 100\%$ of the time. In this example, N_{mod} was computed as 162, which leads to an empirical assurance of .863. A sample size of 161 was too small, as it led to an empirical assurance of .8454, and thus the process was repeated after increasing the sample size by 1 so that the smallest sample size that satisfied the assurance parameter was satisfied.[8]

In this context, the *a priori* Monte Carlo simulation study actually determines the modified sample size; whereas the previously discussed Monte Carlo simulation study, which was not *a priori*, conveyed how effective our formula-based method was. To within sample error, which is minimized in our study due to the 10,000 replications used, the sample size determined from the *a priori* Monte Carlo procedure provides the modified sample size.

Table 2 presents the modified sample size (N_{mod}) from the *a priori* Monte Carlo procedure for each condition in the true-score equivalence model that incorporates a desired degree of assurance of 85%. An examination of the empirical values demonstrates that this method also worked very well. All of the empirical values were .85 or greater. An examination of the empirical assurance values across all the conditions, with mean .857, median .856, and standard deviation .006, reveals that the simulation provides an excellent method to plan sample size for true score models that incorporate a specific assurance parameter.

With the assumptions of the true-score model and sample size planning for coefficient alpha, in many cases, a relatively small increase from the necessary sample size provides an 85% assurance that the observed CI will be no wider than desired. For instance, in the true-score model condition where $W = .1$, $\rho = .4$, with $J = 7$, the necessary sample size is 110. To have 85% assurance the observed CI width will be no wider than .1, the modified sample size is 144.

[7] A sandwich-type algorithm could be used, where low and high values are used and the distance is halved until the appropriate sample size is found. However, we found this improved algorithm was unnecessary, given the speed of modern computing facilities. Although a 'plus 1' approach may not be optimal from a computer science perspective of algorithm development, that is not the emphasis of the current method. Our current method may take slightly longer, but ultimately arrives at the same solution. The correctness of the solution, of course, is what is of importance here.

[8] Setting γ to .85 in our study allows us to examine whether the procedure returns a sample size that is consistently too small (e.g., 60% of the CIs were appropriately narrow) or too large (e.g., 99% of the CIs were appropriately narrow). If γ is .99 then we would not be able to know if the procedure returns the optimal sample size, or one that overestimates it by a large amount, as both situations could have 99% or more of the CIs be appropriately narrow. We want the empirical assurance to be just slightly larger than specified so that the conditions are met.

Table 2. Sample sizes necessary for a specified width for 95% CIs for coefficient alpha in the true-score model *a priori* Monte Carlo simulation with 85% desired assurance

		J						
ρ		3	5	7	9	11	13	15
					$W = 0.05$			
0.1	Population α_c	0.250	0.357	0.438	0.500	0.550	0.591	0.625
	N_{Modified}	10620	6532	4692	3582	2864	2343	1951
	γ_{E}	0.862	0.854	0.85	0.851	0.86	0.86	0.858
0.2	Population α_c	0.429	0.556	0.636	0.692	0.733	0.765	0.789
	N_{Modified}	6174	3144	1990	1391	1041	801	647
	γ_{E}	0.854	0.85	0.85	0.857	0.876	0.858	0.856
0.3	Population α_c	0.563	0.682	0.750	0.794	0.825	0.848	0.865
	N_{Modified}	3635	1632	968	644	467	356	281
	γ_{E}	0.853	0.864	0.852	0.852	0.864	0.853	0.865
0.4	Population α_c	0.667	0.769	0.824	0.857	0.880	0.897	0.909
	N_{Modified}	2131	874	500	325	231	176	139
	γ_{E}	0.851	0.852	0.851	0.859	0.85	0.864	0.862
0.5	Population α_c	0.750	0.833	0.875	0.900	0.917	0.929	0.938
	N_{Modified}	1211	474	264	168	121	91	73
	γ_{E}	0.859	0.851	0.85	0.852	0.85	0.856	0.857
0.6	Population α_c	0.818	0.882	0.913	0.931	0.943	0.951	–
	N_{Modified}	1211	474	264	168	121	91	–
	γ_{E}	0.859	0.851	0.85	0.852	0.85	0.856	–
					$W = 0.10$			
0.1	N_{Modified}	2739	1694	1227	942	755	623	520
	γ_{E}	0.851	0.856	0.863	0.856	0.852	0.851	0.852
0.2	N_{Modified}	1606	833	534	378	286	225	182
	γ_{E}	0.85	0.855	0.859	0.863	0.852	0.867	0.85
0.3	N_{Modified}	964	441	268	179	134	104	85
	γ_{E}	0.85	0.853	0.853	0.864	0.85	0.85	0.852
0.4	N_{Modified}	570	247	143	98	70	55	–
	γ_{E}	0.851	0.855	0.863	0.86	0.855	0.854	–
0.5	N_{Modified}	337	139	78	55	–	–	–
	γ_{E}	0.853	0.865	0.851	0.878	–	–	–
0.6	N_{Modified}	186	76	–	–	–	–	–
	γ_{E}	0.856	0.86	–	–	–	–	–
					$W = 0.15$			
0.1	N_{Modified}	1254	784	570	443	351	296	249
	γ_{E}	0.851	0.85	0.85	0.863	0.853	0.86	0.857
0.2	N_{Modified}	749	390	252	181	137	110	91
	γ_{E}	0.85	0.853	0.857	0.859	0.864	0.856	0.858
0.3	N_{Modified}	452	213	131	91	67	54	–
	γ_{E}	0.864	0.853	0.86	0.86	0.851	0.855	–
0.4	N_{Modified}	274	122	71	50	–	–	–
	γ_{E}	0.851	0.855	0.854	0.853	–	–	–
0.5	N_{Modified}	163	69	–	–	–	–	–
	γ_{E}	0.856	0.858	–	–	–	–	–
0.6	N_{Modified}	94	–	–	–	–	–	–
	γ_{E}	0.856	–	–	–	–	–	–

(continued)

Table 2. Continued

		J						
ρ		3	5	7	9	11	13	15
					$W = 0.20$			
0.1	N_{Modified}	732	457	336	261	210	175	148
	γ_{E}	0.858	0.862	0.864	0.85	0.851	0.862	0.861
0.2	N_{Modified}	437	230	154	110	85	70	56
	γ_{E}	0.862	0.851	0.862	0.866	0.855	0.87	0.86
0.3	N_{Modified}	267	130	81	58	-	-	-
	γ_{E}	0.852	0.863	0.86	0.855	-	-	-
0.4	N_{Modified}	166	74	-	-	-	-	-
	γ_{E}	0.873	0.863	-	-	-	-	-
0.5	N_{Modified}	101	-	-	-	-	-	-
	γ_{E}	0.86	-	-	-	-	-	-
0.6	N_{Modified}	59	-	-	-	-	-	-
	γ_{E}	0.858	-	-	-	-	-	-

Note. J is the number of items on the test, ρ is the estimated inter-item correlation, *W* is the desired CI width, *Population*α_c is the population coefficient alpha, N_{Modified} is the modified sample size using an 85% assurance parameter in an *a priori* Monte Carlo simulation, γ_{E} is the empirical assurance or the percentage of CIs which were no wider than desired with a specified assurance level set at .85, and – indicates the conditions did not meet the necessary criteria. Across all conditions the desired degree of assurance was .85. Multiple error variances were simulated to obtain these results; see the text for how the variances were calculated. Note that the population alpha coefficients do not vary across the multiple CI widths, thus they are only reported for the first *W*.

The MBESS package for the R program contains a function that automates this computationally intense procedure and is easy for researchers to use. An example will be given later, which demonstrates this procedure in the R software program.

5.2. Congeneric model

5.2.1. Expected width

The techniques in the true-score model were not appropriate for use with the congeneric model with coefficient omega. Unlike the true-score model, sample size planning for the congeneric model and coefficient omega cannot be calculated with a closed-form solution due to the iterative maximum likelihood estimation procedure used to obtain the necessary estimates (e.g., Kelley & Change, in press). More specifically, iteration is required because the variance of $\hat{\upsilon}$, $\hat{\nu}$, and the covariance of the two must be estimated using the Fisher information matrix from a maximum likelihood context. Because the CI for the population coefficient omega requires an estimation of the standard error, a formula like equation (19) cannot be developed for the congeneric case. Thus, the necessary sample size cannot be solved in the coefficient omega case as it was for the expected width case for coefficient alpha. Another *a priori* Monte Carlo simulation provides a sample size at which the expected width of the CI for the population coefficient omega is appropriately narrow. Such *a priori* Monte Carlo simulation studies are useful when closed-form analytic expressions are unavailable.

The same conditions used in the true-score model Monte Carlo simulations were used again for this *a priori* simulation. However, the factor structure of a congeneric model, rather than a true-score equivalent model, needed to be simulated. Recall from Figure 1(d) that in the congeneric model the population error variances and the factor loadings can differ.[9] In each condition the error variance (ψ^2) was calculated the same as it was in the true-score simulation. For instance, in the conditions with $J = 15$, the population error variances were modified from the standardized variance, $1 - \lambda^2$, by $\pm .025$, $\pm .05$, $\pm .075$, $\pm .1$, $\pm .125$, $\pm .15$, and $\pm .20$. Additionally, the factor loadings needed to be calculated. As before, the first factor loading (λ) was calculated as $\sqrt{\rho}$, however, the values of ρ were not constant. In the conditions with $J = 3$, the factor loadings were $\sqrt{\rho}$ and $\sqrt{\rho} \pm .1$. The conditions with $J = 5$ included the same factor loadings as $J = 3$, and two additional ones, $\sqrt{\rho} \pm .2$. Each subsequent set of factor loadings included the previous ones, plus an additional two: $\sqrt{\rho} \pm .05$ (for $J = 7$), $\sqrt{\rho} \pm .15$ (for $J = 9$), $\sqrt{\rho} \pm .025$ (for $J = 11$), $\sqrt{\rho} \pm .075$ (for $J = 13$), and $\sqrt{\rho} \pm .125$ (for $J = 15$). Thus, the factor loadings for the conditions with $J = 15$ were: $\sqrt{\rho} \pm .025$, $\sqrt{\rho} \pm .05$, $\sqrt{\rho} \pm .075$, $\sqrt{\rho} \pm .1$, $\sqrt{\rho} \pm .125$, $\sqrt{\rho} \pm .15$ and $\sqrt{\rho} \pm .2$. The same conditions as in the previous simulations were used for ρ (.1 to .6, in steps of .1), J (3–15; increasing by 2), and W (.05, .1, .15, and .2). After eliminating the conditions where the determinant of the covariance matrix was less than 0.000001 or if the necessary sample size was less than 30, 123 conditions remained.

The population covariance matrix for the congeneric model was used in equation (19) as if true-score equivalence held so as to obtain a starting value for the sample size procedure. Use of equation (19) for a congeneric structure does not provide the correct sample size; however, it provides a useful starting point for the necessary sample size that ultimately will be found with the use of the *a priori* Monte Carlo simulation. The simulation for calculating the necessary sample size for the congeneric model uses the same steps as describe before in the *a priori* Monte Carlo simulation that provided true-score sample sizes with specified assurance. Randomly generated multivariate normal data are obtained, as discussed before, that are consistent with the congeneric model so that the properties of the procedures in the condition at the particular sample size can be obtained. The corresponding CI for the population coefficient omega is calculated from this data across the 10,000 replications at each condition. The *a priori* Monte Carlo simulation increments the sample size of the generated data until it finds the minimum necessary sample size such that the mean CI width (i.e., the expected width) is less than or equal to the desired width.

The necessary sample size for the expected width for coefficient omega is not reported in tables for brevity's sake, as in general it is ideal to incorporate a high degree of assurance that the CI will be sufficiently narrow.

5.2.2. Incorporating an assurance parameter

As was the case for the incorporation of the assurance parameter in the true-score model, an *a priori* Monte Carlo simulation provides a given degree of assurance that the observed

[9] Should a distribution other than multivariate normal be presumed (e.g., positively skewed items, Likert-type items, or items with floor and/or ceiling effects), the *a priori* Monte Carlo simulation study we implemented could be reimplemented under the presumed item distributions.

CI for coefficient omega will be no wider than desired. This simulation procedure follows the same procedure as was used in the true-score model for incorporating the assurance parameter. However, the CIs were calculated using equation (17), and the standard error was calculated using equation (16). The population covariance matrices (with multiple factor loadings and multiple error variances) were specified as described above in Section 5.2.1. Across all conditions, γ was set to .85. However, γ can be set to any value without affecting the implementation of the *a priori* simulation procedure. Of course, the necessary sample size would be different if, for example, γ was set to .99, but the implementation is exactly the same.

Table 3 gives the modified sample size for each condition of the simulation that incorporates a desired degree of assurance for the congeneric model with coefficient omega. Table 3 also reports the empirical assurance that was observed across the 10,000 conditions at the modified sample size. Examination of the empirical assurance values across all the conditions shows that the procedure worked very well, having a mean of .855 (standard deviation .004) and a median of .854. The closeness of the mean and the median to the desired assurance parameter illustrates the successfulness of the method. None of the empirical assurance values were less than the desired .85 level of assurance, and the biggest discrepancy was when the empirical assurance parameter was .878. Thus, in the worst case across all conditions, the empirical assurance was only .028 too large. Although this appears to be a large discrepancy, any sample size smaller would have resulted in a less than desired assurance parameter. Thus, this result is the correct one that is the smallest to satisfy the goal of achieving an assurance no less than .85.

6. Applications of AIPE for reliability coefficients

Sample size planning methods for appropriately narrow CIs for population reliability coefficients can be used by both test developers and test users. Test developers have a number of factors to consider when planning sample size for narrow CIs of population reliability coefficients. To decrease the width of the CI, developers can increase the number of test items or the correlations among the items on a test, in addition to increasing the sample size of the norming sample. However, one risk when increasing the number of test items is that the measure may no longer be unidimensional. Eventually, the correlations between the item scores chosen to be included are no longer increasing substantially upon modifications, and the test becomes too long to be 'user friendly'. Our methods allow the developer to also consider sample size to achieve an appropriately narrow CI. We recommend that these procedures be used in an iterative process which allows the test developer to control the width of the CI from a variety of perspectives; each refining the others in an ongoing process during test development, pilot testing, obtaining norms, etc. Depending on the resources and needs of the test developer, each method may be feasible and appropriate at different stages of the process. We encourage test developers to use the AIPE approach as one consideration for controlling the width of the CI on the population reliability coefficient in an effort to home in on the population value with a high degree of confidence.

Users of established tests do not have the opportunity of modifying the number of test items or increasing the correlations among the test items as a way to reduce the width of the CI. When considering sample size, test users need to consider both the need for accurate parameter estimates for a variety of point estimates, as well as how to achieve adequate statistical power. We advocate using both AIPE and power-analytic methods for

Table 3. Sample sizes necessary for a specified width for 95% CIs for coefficient omega in the congeneric model *a priori* Monte Carlo simulation with 85% desired assurance

		J						
ρ		3	5	7	9	11	13	15
					$W = 0.05$			
0.1	Population ω	0.257	0.364	0.443	0.506	0.555	0.595	0.629
	N_{Modified}	10644	6375	4617	3535	2827	2318	1932
	γ_{E}	0.851	0.852	0.854	0.85	0.851	0.851	0.852
0.2	Population ω	0.437	0.562	0.641	0.697	0.737	0.767	0.792
	N_{Modified}	6063	3098	1976	1380	1032	803	642
	γ_{E}	0.854	0.856	0.851	0.855	0.864	0.857	0.857
0.3	Population ω	0.571	0.687	0.753	0.797	0.827	0.850	0.867
	N_{Modified}	3552	1619	965	641	466	355	280
	γ_{E}	0.8527	0.854	0.853	0.853	0.857	0.853	0.857
0.4	Population ω	0.674	0.774	0.826	0.860	0.882	0.898	0.910
	N_{Modified}	2074	872	500	325	232	175	139
	γ_{E}	0.859	0.855	0.851	0.854	0.852	0.852	0.86
0.5	Population ω	0.757	0.837	0.877	0.902	0.918	0.930	0.939
	N_{Modified}	1178	473	265	170	122	92	73
	γ_{E}	0.851	0.85	0.85	0.855	0.858	0.852	0.853
0.6	Population ω	0.825	0.886	0.915	0.933	0.944	0.952	0.958
	N_{Modified}	636	250	138	90	64	49	–
	γ_{E}	0.852	0.858	0.852	0.868	0.854	0.86	–
					$W = 0.10$			
0.1	N_{Modified}	2873	1646	1202	928	747	614	516
	γ_{E}	0.851	0.85	0.852	0.86	0.854	0.852	0.85
0.2	N_{Modified}	1589	814	528	373	282	222	180
	γ_{E}	0.851	0.852	0.856	0.852	0.857	0.861	0.857
0.3	N_{Modified}	940	436	264	180	132	104	83
	γ_{E}	0.857	0.853	0.851	0.851	0.85	0.866	0.859
0.4	N_{Modified}	556	241	141	94	70	54	–
	γ_{E}	0.854	0.854	0.854	0.851	0.855	0.859	–
0.5	N_{Modified}	321	136	78	54	–	–	–
	γ_{E}	0.855	0.858	0.851	0.876	–	–	–
0.6	N_{Modified}	180	75	–	–	–	–	–
	γ_{E}	0.854	0.858	–	–	–	–	–
					$W = 0.15$			
0.1	N_{Modified}	1483	751	554	431	348	289	245
	γ_{E}	0.854	0.852	0.852	0.854	0.856	0.852	0.853
0.2	N_{Modified}	744	379	250	178	136	108	89
	γ_{E}	0.854	0.851	0.854	0.858	0.854	0.855	0.854
0.3	N_{Modified}	440	208	129	90	67	53	–
	γ_{E}	0.856	0.858	0.86	0.87	0.854	0.861	–
0.4	N_{Modified}	263	117	72	49	–	–	–

(continued)

Table 3. Continued

		J						
ρ		3	5	7	9	11	13	15
					$W = 0.15$			
	γ_{E}	0.854	0.855	0.861	0.859	–	–	–
0.5	N_{Modified}	155	67	–	–	–	–	–
	γ_{E}	0.853	0.852	–	–	–	–	–
0.6	N_{Modified}	88	–	–	–	–	–	–
	γ_{E}	0.863	–	–	–	–	–	–
					$W = .20$			
0.1	N_{Modified}	1046	433	324	255	207	173	147
	γ_{E}	0.851	0.858	0.858	0.854	0.85	0.852	0.858
0.2	N_{Modified}	450	224	149	108	83	68	55
	γ_{E}	0.855	0.865	0.86	0.858	0.852	0.866	0.853
0.3	N_{Modified}	261	124	78	55	–	–	–
	γ_{E}	0.85	0.857	0.854	0.851	–	–	–
0.4	N_{Modified}	158	72	–	–	–	–	–
	γ_{E}	0.861	0.856	–	–	–	–	–
0.5	N_{Modified}	94	–	–	–	–	–	–
	γ_{E}	.853	–	–	–	–	–	–
0.6	N_{Modified}	55	–	–	–	–	–	–
	γ_{E}	.859	–	–	–	–	–	–

Note. J is the number of items on the test, ρ is the estimated inter-item correlation, *W* is thedesired confidence interval width, *Populationω* is the population coefficient omega, N_{Modified} is the necessary sample size using an 85% assurance parameter in an *a priori* Monte Carlo simulation, γ_{E} is the empirical assurance or the percentage of confidence intervals which were no wider than desired with a specified assurance level set at .85, and – indicates the conditions did not meet the necessary criteria. Multiple error variances and factor loadings were simulated to obtain these results; see the text for how the variances were calculated. Note that the population reliability coefficients do not vary across the multiple confidence interval widths, thus they are only reported for the first *W*.

sample size planning when research questions are based on having an accurate estimate that is also statistically significant. AIPE and power analyses should not be regarded as being independent. In fact, increasing the sample size will improve both the accuracy of an estimate as well as the statistical power of a null hypothesis statistical test that tests a false null hypothesis, but not necessarily equally. Thus, both power analysis and AIPE analysis should be conducted according to the needs and goals of the researcher. If the power analysis yields the larger sample size, then that sample size should be used. Conversely, if the AIPE analysis yields the larger size for an appropriately narrow CI for a reliability coefficient, then that should be the sample size goal. Ideally, an adequate sample size from both perspectives would be used (e.g., Jiroutek, Muller, Kupper, & Stewart, 2003). However, these *a priori* sample sizes may be quite different. Given the researcher's goals, resources, and needs, the researcher should decide which method yields the most appropriate sample size. At times, a researcher may find a sample size to be unreasonably large and impractical, given the available resources. As Maxwell *et al.* (2008) described, meta-analyses can provide one method to help address the wide CIs in individual studies when researchers may have limited resources. Another approach is to consider multisite studies, with the idea being to 'spread the burden but reap the benefits of estimates that are accurate and/or statistically significant' (Kelley & Rausch, 2006, p. 375). Regardless of which sample size is chosen, the researcher has further information regarding levels of accuracy for the reliability coefficient and power for the desired key effects and can anticipate findings for after the study has been completed.

To plan sample size for appropriately narrow CIs for population reliability coefficients, test users and developers need any one of the following sets of information to enter into our function in MBESS (Kelley & Lai, 2011a): (a) a covariance matrix; (b) the number of test items, estimated or observed inter-item score correlations, and a vector of observed or estimated error variances; or (c) the number of test items, factor loadings, and a vector of observed or estimated error variances.

6.1. Specifying input values

Lai and Kelley (2011) propose several methods to aid in the specification of input values for a covariance matrix in the context of AIPE for targeted effects in structure equation models, which are applicable in the present context (see also Kelley & Lai, 2011b). First and foremost, these input values should be based on the existing literature. A thorough literature review provides the foundation for specifying the relationships between the variables. Existing data sets, previous analyses, and similar constructs should be evaluated and reasonable estimates can be extracted from these.

Lai and Kelley (2011) discuss how a covariance matrix can be estimated based on the relationship between the correlation matrix and the covariance matrix of relevant variables. The correlation coefficient of each pair of variables can be estimated and arranged into a matrix. Specifying the correlations between the variables can be aided by Cohen's (1988) widely used suggestions of 'small' (.10), 'medium' (.30), and 'large' (.50) correlation coefficients. Lai and Kelley (2011) also outline how specifying the covariance matrix can be made easier by using coefficient H, proposed by Hancock and Mueller (2001). For further details on this process, and description of a function in MBESS that creates a model-implied covariance matrix given a model and model parameters, see Lai and Kelley (2011).

7. Illustrative example

Suppose a researcher wants to use some of the test items from the Americans' Changing Lives study (House, 2002). The Americans' Changing Lives study is a longitudinal study investigating multiple facets of Americans' lives, including: how people are 'productive'; how they adapt to stressors that impact their health and functioning; and the consequences of their activities and relationships (House, 2002). The study specifically focused on differences between White and Black Americans who were in their middle and later life phases. The 2002 data consist of three waves of data collection, from 1986, 1989, and 1994, with over 3,000 participants. One of the indices used in the 2002 report is a marital satisfaction and harmony index, consisting of eight test items. Using only the 1,555 complete cases from the 1994 wave of data collection, the reliability estimate using coefficient omega and the congeneric model is .821, with 95% CI [.807, .834].

Suppose a researcher wants to use the marital satisfaction and harmony index in a different study, include only Black and White Americans in their third marriage. This new study will have a smaller number of participants. Consequently, the researcher knows that with a smaller sample size, all other things being equal, the width of the CI will increase. Thus, the researcher wants to plan an appropriate sample size in advance so the CI for the population value of the reliability for the marital satisfaction and harmony index is sufficiently narrow. It would be undesirable to report the results with a reliability coefficient that is accompanied by a wide CI.

The MBESS (Kelley, 2007a, 2007b; Kelley & Lai, 2011a) package for R can be used to plan sample size for an appropriately narrow CI with a desired degree of assurance by implementing the methods we have developed here. To use the computationally intense *a priori* Monte Carlo method in MBESS, the researcher would substitute the necessary information into the `ss.aipe.reliability()` function:

```
ss.aipe.reliability(model = "Congeneric", width = W, S = S, conf.level =
1 − α_e, assurance = γ, initial.iter = 500, final.iter = 5000)
```

where `model` is used to identify the model of interest, either `true-score equivalent` or `congeneric;` `S` is the supposed population covariance matrix of the item scores measuring a particular factor; and `initial.iter` and `final.iter` represent the number of initial iterations and final iterations the simulation performs.[10] Instead of using a covariance matrix, users can also enter more specific data, such as a data set from a pilot test, or the inter-item score correlations, estimated factor loadings, and population error variances. More details on the options for using the `ss.aipe.reliability()` function are available in the MBESS help files.

If the researcher had the covariance matrix for the eight test items in the marital satisfaction and harmony index, from the Americans' Changing Lives data set, the `ss.aipe.reliability()` function would be implemented as follows:

```
ss.aipe.reliability(model = "Congeneric", width = .10,
S = Cov.Marit.Satisfaction, conf.level = .95,
assurance = .85, initial.iter = 500, final.iter = 5000)
```

where `Cov.Marit.Satisfaction` is the covariance matrix of the six test items used to measure the index in the original data set.

After submitting the code above, the `ss.aipe.reliability()` function returns the following:

```
$Required.Sample.Size
[1] 110
$width
[1] 0.1
$specified.assurance
[1] 0.85
```

[10] The number of iterations indicates how many times in the *a priori* simulation the data are generated and the CIs are calculated. This is a computationally intense procedure, thus the option of entering the initial iterations (e.g., 100 or 200 times) allows the user to specify a smaller number of iterations that can be used to arrive at an approximate sample size. Once the percentage of CIs reaches the desired degree of assurance (e.g., 85% of the CIs are appropriately narrow), then the simulation uses the number of final iterations. The final number of iterations homes in on the true sample size value due to the large number of replications (e.g., 10,000 times) and the correspondingly small degree of sample error in the *a priori* Monte Carlo simulation results. The process continues to modify the sample size until the desired accuracy is reached.

```
$empirical.assurance
[1] 0.856
$final.iter
[1] 5000
```

As can be seen, the sample size that is necessary for the researcher to have a CI not wider than .1 more than 85% of the time for the population value of coefficient omega is 110. The output also gives the observed empirical assurance across the 5,000 final iterations of the simulation. Recall that the empirical assurance will tend to be slightly more than the desired assurance so the sample size is required to be a whole number. Although this is a computationally intense approach to sample size planning, it is easy to do with the MBESS `ss.aipe.reliability()` function.

Suppose the researcher conducts a study with 110 participants. Now the researcher wants to calculate a CI along with a point estimate for the population coefficient omega. The `ci.reliability()` function in MBESS can easily calculate this, with the code

```
ci.reliability(model = "Congeneric", S = Cov.Mat, N = 110)
```

where `Cov.Mat` is the covariance matrix from the researcher's sample with 110 participants. The options for model and type are the same as for the `ss.aipe.reliability()` option, with the additional required field of the observed sample size (N).

This code returns the following:

```
$CI.lower
[1] 0.7923504
$CI.upper
[1] 0.8864628
$Estimated.reliability
[1] 0.8394066
$SE.reliability
[1] 0.02400871
$Conf.Level
[1] 0.95
```

The lower and upper limits of the CI are .792 and .886, respectively, with a point estimate of coefficient omega of .839. The function also returns the standard error and a reminder of the CI coverage that was used. Although the width of the CI was specified to be .10 in the sample size planning procedure, the researcher found with this particular sample that the CI was smaller, .094. This slight decrease is not unexpected, as the sample size planning procedure returns a value that is not wider than specified (and the desired width was .10).

8. Discussion

This paper has covered several key areas. First, the foundation was provided with a brief description of homogeneous tests and three common measurement models. This was followed by a review of reliability with a specific focus on coefficient alpha and coefficient omega. Second, methods were given for forming CIs for the population reliability coefficients in the true-score and congeneric models. Third, methods were provided for sample size planning for accurate estimation of coefficient alpha and coefficient omega via narrow CIs. A modification to these methods provided a level of certainty such that the CIs will be sufficiently narrow with no less than the desired degree of assurance. Descriptions of the Monte Carlo simulation studies that were used to evaluate and plan sample size were provided, and the results from these studies given. Applications of the AIPE approach were provided for test users and test developers. Finally, an example was given that illustrates how the easy-to-use and freely available program, R, can be used with the MBESS package to implement the procedures described in this paper.

Reporting reliability coefficients for a set of scores coming from each administration of an instrument is crucial (Wilkinson & APA Task Force on Statistical Inference, 1999). In general, reporting CIs is seen as a 'best reporting strategy' (APA, 2001, p. 22) and is endorsed by APA (2009, p. 34), AERA (2006) and Wilkinson and the APA Task Force on Statistical Inference (1999). Support for reporting CIs for population reliability coefficients comes from a variety of factors: it allows for the comparison of estimates across studies (e.g., Thompson, 2002), assists and encourages meta-analytic thinking (e.g., Hunter & Schmidt, 2004; Rodriguez & Maeda, 2006), and allows for the assessment of the uncertainty in the population value estimate (e.g., Hahn & Meeker, 1991). Substantial information is lost if only the reliability point estimate is reported instead of also reporting the CI. The AIPE approach depicted here provides an approach to planning sample size to achieve reliability coefficient estimates that are sufficiently accurate. We encourage test developers and test users to plan sample size for CIs for the population value of coefficient omega rather than coefficient alpha in almost all cases of a homogeneous test.

A common 'urban legend' is that a reliability estimate should be no less than .70 (Lance, Butts, & Michels, 2006). This guideline seems to have grown out of a passage from Nunnally (1978) and is widely referenced (Peterson, 1994). However, as Lance *et al*. (2006) described, this is an overly simplistic representation of the original source, as Nunnally did not recommend a lower limit for reliability estimates as universally being .70; rather he stressed that 'what a satisfactory level of reliability is depends on how a measure is being used' (Nunnally, 1978, p. 245). The lower limit of .70 was for tests in the 'early stages of research' (Nunnally, 1978, p. 245). However, for basic research, he endorsed a reliability estimate of .80; for settings where 'important decisions' are made, the minimum should be .90 (Nunnally, 1978). We argue that, if such a rule of thumb is to be used, the lower limit of a CI is the value that should be compared to Nunnally's benchmarks, rather than the point estimate itself. A point estimate does not demonstrate the range of plausible parameter values. We propose that researchers use a lower limit of a CI of .70 for early stages of research, .80 for basic research, and .90 for applied settings with decisions based on the results. We advocate that the lower limit of the CI should be the focus for guidelines for 'acceptable' reliability estimates, rather than the point estimate itself. We hope these methods will allow researchers to plan sample size as a way to avoid potentially 'embarrassingly large' CIs.

When planning sample size for reliability coefficients we recommend incorporating a large (e.g., .99) assurance parameter. Such an assurance parameter provides probabilistic assurance that the CIs that will be computed will be no wider than desired with a specified degree of assurance. Some may argue that it unnecessarily increases the sample size without enough 'gains'. We argue that the benefit of a probabilistic assurance that the CIs will be no wider than specified is an important reason to increase sample size. An examination of the increase in sample sizes from the necessary sample size (with no assurance parameter) to the modified sample size (with an assurance parameter) demonstrates that the increase in necessary sample size tends to be quite small relative to the necessary size in order for the expected CI width to be sufficiently narrow.

It is the test scores on a particular administration of a test that are reliable, not the test itself. Thus, given the need for researchers to report an estimate of reliability for the scores obtained on their particular administration of a test, we hope that this paper helps researchers gain a better understanding of reliability and CIs for population reliability coefficients, facilitates sample size planning when interest is in the value of the population reliability coefficient, and makes easy-to-use functions available so that researchers can easily apply the methods we discussed to facilitate their research.

References

American Educational Research Association (AERA). (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33–40.

American Psychological Association (APA). (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

American Psychological Association (APA). (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335–340.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559–566.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p <$ .05). *American Psychologist*, *49*(12), 997–1003.

Committee on Reviewing Evidence to Identify Highly Effective Clinical Services. (2008). *Knowing what works in health care: A roadmap for the nation*, J. Eden, B. Wheatley, B. McNeil, & H. Sox (Eds.). Washington, DC: National Academies Press.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

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

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61*(4), 517–531.

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*(6), 930–944.

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.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135.

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

Hahn, G. J., & Meeker, W. Q. (1991). *Statistical intervals: A guide for practitioners*. Hoboken, NJ: Wiley.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. H. C. du Toit & D. Sörbom (Eds.), *Structural equation modeling:*

Past and present. A Festschrift in honor of Karl G. Jöreskog (pp. 195–261). Chicago: Scientific Software International.

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*(4), 523–531.

House, J. S. (2002). *Americans' changing lives: Waves I, II, and III 1986, 1989, and 1994 [Data file and code book]*. Ann Arbor: University of Michigan, Institute for Social Research, Survey Research Center [producer], Inter-university Consortium for Political and Social Research [distributor].

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*(3), 153–160.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology*, *13*(4), 478–487.

Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P.W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics*, *59*, 580–590.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association (AERA), the American Psychological Association (APA), & the National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8), 1–24.

Kelley, K. (2007b). Methods for the Behavioral, Educational, and Social Sciences: An R package. *Behavior Research Methods*, *39*(4), 979–984.

Kelley, K., & Cheng, Y. (in press). Estimation and confidence interval formation for reliability coefficients of homogeneous measurement instruments. *Methodology*.

Kelley, K., & Lai, K. (2011a). MBESS: Methods for the Behavioral, Educational, and Social Sciences, Version 3.0.0 or higher. [Computer software and manual]. Retrieved from http://www.cran.r-project.org/

Kelley, K., & Lai, K. (2011b). Accuracy in parameter estimation for the root mean square error of approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research*, *46*, 1–32.

Kelley, K., Lai, K., & Wu, P. (2008). Using R for data analysis: A best practice for research. In J. Osbourne (Ed.), *Best practices in quantitative methods* (pp. 535–572). Newbury Park, CA: Sage.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(3), 305–321.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in "AERJ" and "JCP" articles from 1988 to 1997: A methodological review. *The Journal of Experimental Education*, *69*(3), 280–309.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kline, R. B. (2005). *Principles and practices of structural equation modeling* (2nd ed.). New York: Guilford Press.

Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α. *Applied Psychological Measurement*, *21*(4), 337–348.

Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, *16*, 127–148.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220.

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

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

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537-563.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*(2), 157-176.

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

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393-426). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *4*, 599-620.

Nunnally, J. C. (1978). *Psychometric theory*, (2nd ed.) New York: McGraw-Hill.

Oehlert, G. W. (1992). A note on the delta method. *American Statistician*, *46*, 27-29.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*(3), 343-355.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*, 381-391.

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essentially tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*(4), 329-353.

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research*, *37*(1), 89-103.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*(3), 306-322.

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

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

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25-32.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174-195.

Vacha-Haase, T. (1998). Reliablity generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*(3), 271-280.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. (4th ed.). New York: Springer.

Wilkinson, L. & the American Psychological Association (APA) Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

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

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

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

Received 18 October 2010; revised version received 30 July 2011