Alpha's Standard Error (ASE): An Accurate and Precise Confidence Interval Estimate

Adam Duhachek and Dawn Iacobucci

Alpha's Standard Error (ASE): An Accurate and Precise Confidence Interval Estimate

In this research, we present the inferential statistics for Cronbach's Coefficient Alpha based on the standard statistical assumption of multivariate normality. The estimation of alpha's standard error (ASE) and confidence interval for alpha is described, and we conduct an analytical demonstration to illustrate the effects on these estimates of the components of the equations, including the number of items, the item intercorrelations, and sample size. We then empirically examine the consistency and efficiency of alpha and its standard error in a Monte Carlo simulation modeling the effects of alpha's components, sample size, and compound symmetry. We then demonstrate the superiority of this estimate compared to previous derivations of alpha's standard error in a separate Monte Carlo simulation. For the researcher interested in assessing the difference between alphas obtained in two independent samples, we present a sampling error and test statistic for such a comparison. We conclude with a prescription that includes the recommendation that all alpha coefficients be reported in conjunction with a standard error or confidence interval estimate. We offer SAS and SPSS programming codes for easy implementation.

Alpha's Standard Error (ASE): An Accurate and Precise Confidence Interval Estimate

Measurement development is an integral part of behavioral research. Numerous articles have appeared in the literature dedicated to the development and assessment of scales concerning a broad array of research topics. (e.g., Kinicki, McKee-Ryan, Schriesheim & Carson, 2002; Colquitt, 2001; Ghorpade, Hattrup & Lackritz, 1999; Chan, Drasgow & Sawin, 1999). The validity of empirical behavioral research hinges on ensuring the reliability of the measures upon which research conclusions are based, so measurement reliability is an issue of unquestioned importance to both the theoretical and applied researcher. In fact, extensive volumes are dedicated to documenting vast sets of reliable measures for use in future research (Robinson, Shaver & Wrightsman, 1991).

Texts and articles are also numerous that guide the researcher in the construction of scales, from the early stages of theoretical demarcation and domain sampling, through to the statistical testing, e.g., through confirmatory factor analysis, of the convergent and discriminant validity of the scale's factorial structure (e.g., Aiken, 2002; Bentler & Raykov, 2000; Gregory, 2001; Nunnally & Bernstein, 1994). An important step in this evaluation of the goodness of a scale is the estimation of its reliability. Establishing measurement reliability is of inarguable importance in both applied and theoretical research, because reliability constitutes a necessary first step towards ensuring construct validity (e.g., Allen & Yen, 1979; Anastasi and Urbina, 1996; Cronbach, 1951). Reliability is deemed so important that even when authors are not creating a scale, rather they are only using established scales, reviewers and readers nevertheless expect a reliability index to be reported. By far the most frequently reported index of internal consistency is Cronbach's Coefficient Alpha (Cortina 1993; Hogan, Benjamin & Brezinski, 2000).

Thus, coefficient alpha is important. In this paper, we investigate the ramifications of recent statistical developments regarding the distribution and standard error of coefficient alpha. While

methods for estimating confidence intervals (Feldt, 1965; Feldt & Ankenmann 1999) and standard errors (Hakstian & Whalen 1976; Barchard & Hakstian, 1997a; 1997b) for coefficient alpha have existed for some time, such advances have largely been disregarded perhaps because they were based on highly constrained, practically unobservable data assumptions (e.g., strict item equivalence). Recently, an asymptotic distribution for the maximum likelihood estimator of the variance of coefficient alpha (van Zyl, Neudecker & Nel, 2000) based on the standard statistical assumption of multivariate normality was derived. From this variance, we present the estimate of Alpha's Standard Error, hereafter referred as ASE. These developments are of great importance to empirical researchers. For the first time, applied and theoretical researchers are able to estimate the standard errors of their measures, thereby revealing precisely the magnitude and severity of the problem of measurement error with less restrictive assumptions on the data. Further, these standard errors may be used to construct confidence intervals around alpha. Such intervals have a variety of practical applications, such as comparing group differences across independent test samples, or estimating measurement error bias in observed Pearson correlations. As of yet, the implications of these developments for researchers have not been investigated in the literature.

In this paper, we first present the statistics for computing confidence intervals and standard error estimates for alpha. In study 1, we analytically demonstrate the sensitivity of ASE to the component factors of: the number of items, the number of respondents, and the level of item intercorrelations, as well as the effect of heteroscedasticity (unequal variances) on this new estimate. In study 2, we employ a Monte Carlo simulation to investigate the effect of covariance heterogeneity on alpha's standard error, that is, the multidimensionality of a scale. We conclude that, although alpha is robust to this factor, alpha's standard error is adversely affected (although this effect is small). In study 3, we comparatively assess the performance of the new derivation of ASE vis-à-vis earlier derivations, offering conclusive evidence for the robustness of the new statistic to data conditions commonly observed in empirical research, while revealing significant biases in alternative derivations. We extend

the derivation of these test statistics one step further, developing a standard error to assess differences between alphas derived from independent samples. To facilitate use in empirical research, we provide computer programs for the user to assess the size of standard error estimates for their own research scales. Finally we synthesize the findings of the studies and articulate best practices for the reporting of alpha, recommending that researchers publish standard error estimates and confidence intervals in addition to reliability point estimates. We begin by motivating our investigation into alpha's standard error by underscoring the benefits of augmenting the conventional reporting of coefficient alpha with the reporting of a standard error estimate.

The Benefits of Reporting Standard Error Estimates of Measurement Reliability

Conventional wisdom dictates the reporting of a point estimate of Cronbach's alpha as a means of assessing internal consistency of measurement scales as a necessary and sufficient component of empirical research. However, the reporting of a mere point estimate conveys no information regarding the precision of the estimate. The computation of a standard error estimate is therefore critical. In all of behavioral research, both basic and applied, more inferential statistics (e.g., F-tests, p-values, confidence intervals) are acknowledged to provide greater information beyond their descriptive counterparts (e.g., means, proportions, or in this case, coefficient alphas). Although the strongest argument in support of reporting inferential statistics in conjunction with alpha point estimates is based on statistical theoretical grounds, there are also ample practical circumstances in which the benefits of these additional statistics are particularly salient. In addition, studies 1-3 examine factors which differentially impact alpha and its standard error. Therefore, both statistics are useful and each confers unique information about the soundness of a scale.

In this section, we briefly discuss four common research scenarios which would benefit from a standard error being reported in conjunction with a reliability statistic. These examples are by no means exhaustive; it is our belief that in all research settings, theoretical or applied, the supplemental

reporting of alpha's standard error conveys valuable information and that such reporting represents the state-of-the-art and best practices for assessing measurement reliability.

First, consider the assessment of rivaling tests measuring constructs such as organizational commitment and job satisfaction. Normally, the tests deemed superior are those that experts agree possess greater construct and/or predictive validity, or are more practical to administer. If the tests possess comparable alpha reliabilities, empirical evidence of the superiority of one measure over another may prove elusive, if not for the calculation of standard errors, which would provide another point of comparison for evaluative decisions regarding test validities.

A second common research scenario involves the large random component inherent in data, particularly in laboratory responses. According to Schmidt & Hunter (1996 p.203):

> "…in most research contexts—and in laboratory studies in particular—any observed response will have a very large random component. …[T]he correlation between 'replicated' responses is rarely any higher than .25."
>
> Imagine a research setting where the construct … is measured using a single item. If we make the justifiable assumption that the reliability of this measure is no higher than .25, we observe … according to the Spearman-Brown formula, had the researcher used a seven-item scale to measure satisfaction, the reliability would climb to .70, … nearly three times more reliable...
>
> … [In a recent review of] over 300 meta-analyses on a very wide variety of psychological treatments, [t]he average effect size was d=.46 (corresponding to a point-biserial r of .22, if sample sizes are equal in experimental and control groups).
>
> …With a reliability of .25, the effects sizes in [the] study would be reduced from a potential d=.46 to an actual d=.23 (i.e., $\sqrt{.25}(.46) = .23$). The probability of

finding a significant result would then be only 12% (using a one-tailed test). That is, the probability of an error for the significance test in this study is 88%.

In such instances, to reduce the uncertainty stemming from such dramatic reductions in statistical power, the computation of standard errors of alpha would enable researchers to determine whether or not the effect would obtain at the increased level of reliability implicated by the confidence interval upper bound.

In another example, consider the researcher who needs a correlation coefficient of .50 or greater to affect a public policy. The computed correlation is .45 but the researcher diligently uses the reliabilities on the two scales to compute the estimated correlation among the true scores, using the standard correction for disattenuation, $r_{t_x t_y} = .45/\sqrt{(.85)(.95)} = .501$, thereby achieving the desired value through a completely defensible procedure. This scenario is not unlikely; these reliabilities are high, typical of much research that appears in top journals, such as *JAP*[1]. The researcher comfortably concludes that even with the correction, the correlation appears to exceed the requisite threshold. However, had each of those reliabilities been computed with its respective standard error, and a confidence interval composed about each, the story could have been different. If we assume a sample size of n=100, a scale length of p=5, and item intercorrelations averaging fairly high, r =0.6, then the standard errors of these reliabilities will be .02 and .01. The 95% confidence regions then would be .811-.889 and .930-.970. The disattenuation correction then ranges from an acceptable $.45/\sqrt{(.811)(.930)} = .518$ to a more questionable: $.45/\sqrt{(.889)(.970)} = .484$.

Finally, consider a scenario for which a testing organization is trying to refute claims of bias against a subpopulation. If the organization wished to demonstrate the strength of the relationship between the test and some criterion was no different from that relationship in another population, the reliabilities, as well as their high and low end estimates would need to be considered.

[1] In fact, Peterson (1994) found that average alpha reliability was only .77. In such cases, the potential impact of considering standard errors is enhanced even further.

These scenarios illustrate that standard error estimates might provide additional evidence that may qualitatively alter the conclusions formulated in a research context. The examples are intended to accentuate the value associated with computation of standard error statistics as a diagnostic research tool. In the subsequent studies, we examine factors that differentially impact alpha's standard error.

Cronbach's Coefficient Alpha

Cronbach's coefficient alpha is widely known and defined as follows (Cortina 1993; Cronbach, 1951; Li, 1997; Li, Rosenthal & Rubin, 1996; Osburn, 2000; Mendoza, Stafford & Stauffer, 2000; van Zyl, Neudecker & Nel, 2000; Yuan & Bentler, 2002):

$$\alpha = \frac{p}{p-1}\left[1 - \frac{\sum_{i=1}^{p}\sigma_i^2}{\sigma_T^2}\right], \qquad (1)$$

where p is the number of items in the scale (given the denominator of the first term, p must be 2 or greater); σ_i^2 is the variance of the ith item, i=1,2,…p; and σ_T^2 is the variance of the entire test, hence it is the sum of the item variances and covariances: $\sigma_T^2 = \sum_{i=1}^{p}\sigma_i^2 + \sum_{i\neq}\sum_{j}\sigma_{ij}$. Equation (1) is the formula that is familiar to researchers employing indices measuring internal consistency (cf. Cortina 1993 p.99 for multiple interpretations of alpha).

Coefficient Alpha's Statistical Distribution

The early works of Kristoff (1963) and Feldt (1965) are among the first attempts reported in the literature to account for a statistical distribution for alpha, both based on an exact *F*-distribution confidence interval procedure. Feldt (1969) applied this theoretical distribution in presenting a means

of comparing independent alpha reliability coefficients for the dichotomous variable K-R 20 case[2], and later for the case of dependent coefficients (Feldt 1980), under the strict assumption of compound symmetry (equality of test item variances and covariances).

Hakstian & Whalen (1976) first extended Feldt's findings to the K>2 group case based on asymptotic distribution theory. This normalizing procedure approach was also predicated on the restrictive assumption of compound symmetry. Subsequent work (Barchard & Hakstian, 1997a, 1997b) demonstrated that such an approach was not robust to Type I error under violations of these assumptions, thus discouraging use of this approach in many applied research settings. Although the benefits of a statistically valid confidence interval perhaps motivated this research, the restrictive assumptions imposed on the data by these derivations was not conducive to empirical settings. Although alternative statistical procedures, such as bootstrapping (Raykov, 1998a) have been employed to obtain estimates of alpha's standard error, such procedures are not without biases of their own. Only recently has this problem been addressed, making the computation of standard errors and confidence intervals for alpha possible using standard inferential statistical assumptions.

Recent research has affirmed that equation (1) is the form of the maximum likelihood estimator of alpha based on a standard assumption of multivariate normality (van Zyl, Neudecker, & Nel, 2000). With this assumption, the distribution of alpha is derived; as n$\rightarrow\infty$, then $\sqrt{n}(\hat{\alpha}-\alpha)$ has a Normal distribution with a mean of zero and a variance of:

$$Q = \left[\frac{2p^2}{(p-1)^2(j'Vj)^3}\right]\left[(j'Vj)(trV^2+tr^2V)-2(trV)(j'V^2j)\right], \qquad (2)$$

where n represents sample size, $\hat{\alpha}$ is the MLE of α, j is a px1 vector of ones and **V** is the population covariance matrix among the items (van Zyl, Neudecker & Nel, 2000).[3]

[2] K-R 20 is the Kuder-Richardson formula 20, applied to binary data, e.g., responses scored "right" or "wrong."

Armed with a variance, in this paper we derive alpha's standard error, ASE = $\sqrt{\frac{Q}{n}}$, use that in conjunction with the distribution (Normal) to calculate a z-score and p-value to assess the significance of alpha or to create confidence intervals to complement the knowledge about the size of alpha. Given that the information contained in a hypothesis test is comparable to that in a confidence interval (cf. Cortina & Dunlap, 1997), we choose the latter, computing the 95% confidence interval as:[4]

$$\hat{\alpha} \pm (1.96)\left(\sqrt{\frac{Q}{n}}\right). \qquad (3)$$

Computation of such intervals equips researchers with more insight into their measures. In addition to providing conclusive evidence about the relative superiority of certain tests, these intervals allow researchers to estimate bias due to measurement error.

For example, observe in Table 1 that for a two item scale with an item intercorrelation of .6 and a sample of 30, the alpha is .75 with a standard error of .091. The respective confidence interval for this alpha is .57 < $\hat{\alpha}$ <.93. Observed correlations involving this scale would be attenuated by 7% to 43%. Armed with this information, the researcher could interpret marginally significant findings with the knowledge that the observed correlation was up to 43% below its correct value (see Schmidt & Hunter, 1996 for many scenarios involving the impact of measurement error on validity). Note that this same point estimate of alpha (.75) may be obtained when p=7 and the average item intercorrelation is .3 (cf. Cortina 1993, p.101). In this latter case, for a sample of 200, the standard error is .027 and the confidence interval for alpha is .70 < $\hat{\alpha}$ <.80. In this case, with a smaller standard error, the potential

---

[3] In summation notation: $$Q = \left[\frac{2p^2}{(p-1)^2\left(\sum_{i=1}^{p}\sum_{j=1}^{p} v_{ij}\right)^3}\right]\left[\left(\sum_{i=1}^{p}\sum_{j=1}^{p} v_{ij}\right)\left(\sum_{i=1}^{p}\sum_{k=1}^{p} v_{ik}v_{ki} + \left(\sum_{i=1}^{p} v_{ii}\right)^2\right) - 2\left(\sum_{i=1}^{p} v_{ii}\right)\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} v_{ik}v_{kj}\right)\right].$$

[4] Even if α exceeds zero, $\hat{\alpha}$ and certainly the lower-bound of the confidence interval, may be estimated to be less than zero, particularly for small sample sizes. If a researcher computes the confidence interval and the lower-bound is negative, they should truncate the estimate and report the lower-bound to be zero.

downward correlational bias between two variables exhibiting these measurement properties ranges from 20% to only 30%.

Therefore, computing the standard error and confidence interval associated with a scale's alpha bears importance beyond the information it reveals about the stability of the measurement instrument itself. These examples demonstrate how these statistics may impact predictive validity for both theoretical and applied research. To cultivate an understanding of the standard error of alpha, we present an analytical illustration of these statistics for varied levels of the factors that effect the estimation: n (sample size), p (number of items comprising the scale), r_{ij} (the correlations among the items) and item variance.[5]

Study 1: Analytical Comparisons of the Behavior of Alpha's Standard Error (ASE)

In Table 1 we report α and its standard error, for each combination of n, p, and $\bar{r}_{ij}$. We let sample size, n, range from n=30, a relatively small sample for research reported in the literature, and a value at which point the Central Limit Theorem assists the behavior of test statistics. We then selected admittedly somewhat arbitrary values for higher n, but these selections were motivated by practical considerations. For instance, we reasoned that relatively few research articles report sample beyond n=200. Even for field research yielding larger samples, we note that alpha estimates are insensitive to sample size and the results on the confidence intervals that we point to momentarily indicate that similar precision is attained with samples of 200, 100 or 50, if there are sufficient numbers of items and strong item intercorrelations.

--- Table 1 goes about here ---

We selected values for p, beginning with the bare minimum of 2, adding the levels of 3, 5, 7, and 10. As we shall soon demonstrate, the empirical performance of the confidence intervals indicate

[5] Previous equations for standard errors had existed, e.g., Cortina (1993, p.101), but without distributional assumptions, there was "no real metric for judging the adequacy of the {alpha} statistic" (also see Hakstian & Barchard 2000).

that little additional information is typically gained beyond five or certainly seven items, thus obviating the need to consider p>10.

Finally, we explore the impact of $\bar{r}$, the average intercorrelation among the items, on the alpha and confidence intervals. We study the full range of $\bar{r}$, from 0.0, 0.1, 0.2, …, 0.9, 1.0. To use the terms in the literature, this factor explores internal consistency, the extent of inter-relatedness among the items (Cortina, 1993, p.100; Raykov, 1998b). We also investigate a variation on this factor. We illustrate the effect of the standard deviations of the items on the confidence intervals. This analytical manipulation was explored because we believed that most composite scale items in practice do not share equal variances.[6] To estimate the effect this inequality exerts on alpha, we display the unequal variance results in Table 2. By "differing slightly," we operationalize that the standard deviations will range from 1.0 to 2.0. In Table 3, we allow even greater heteroscedasticity; specifically the standard deviations range systematically from 1.0 to 5.0.[7]

--- Tables 2 and 3 go about here ---

At this point, our design is fully presented. We have incorporated all the factors that can impact Coefficient Alpha. As many researchers have pointed out, and as the reader can verify by examining equation (1), alpha is a function of p (the number of items), and r_{ij}'s, the correlations among the items (Cortina, 1993; Cronbach, 1988; *Journal of Consumer Psychology* Special Issue, 2001). Sample size does not enter into the computation of alpha, but its square root enters into the calculation of the confidence interval limits. Even intuitively, the reader can understand that the item intercorrelations themselves are more stably estimated as sample size increases. Finally, for good measure and to be thorough, we allowed item standard deviations to vary.

---

[6] Table 1 reflects the pure case of parallel tests (Allen & Yen, 1979), wherein item standard deviations are equal, and the item intercorrelations are constant. In these scenarios, the general alpha index also simplifies to the Spearman-Brown formulation, or that which is referred to as the case of compound symmetry by van Zyl, Neudecker and Nel (2000). Real data may vary more in scale across items (Tables 2 and 3).

[7] For standard deviations ranging from 1 to 2 and 1 to 5, the difference was evenly distributed in a step function across p; e.g., for p=3 and range 1 to 2, the standard deviation vector was [1, 1.50, 2.0]' and this vector pre- and post-multiplied the matrix of $\bar{r}$, to create a covariance matrix with the same structure of associations.

Results. The Coefficient Alphas and alpha's standard errors (ASE) are reported in each cell in Tables 1 through 3. Examining the standard errors in Table 1 yields four striking insights. First, the standard errors are smaller, that is, the estimation of alpha is more precise, as the item correlations increase. Standard errors begin large for $\bar{r}$ =0.0 regardless of p, and they decrease as $\bar{r}$ approaches 1.0. Second, standard errors are always larger for smaller sample sizes, as one might expect, though the differences between n=30 and n=200 are nominal for $\bar{r}$ =0.6 or higher even when there are only 2 items, and when $\bar{r}$ =0.4 or higher when p=5 or more. Third, the impact of p, the number of items, is also clear. The enhancement of $\bar{r}$ when p=2 is nearly linear, but when p=10, 7, or even 5, standard errors decrease rapidly from $\bar{r}$ =0.0 to $\bar{r}$ =0.4. Fourth, with a greater number of items, standard errors begin smaller, even for relatively small samples and relatively small item correlations.

Reflecting on these various trends, we note that the effect of sample size is the standard case of gaining power as one has obtained more information. There is an asymptotic effect in that a sample of size 200 is not much more effective in obtaining useful, precise estimates than a smaller sample of size of even n=30 if p and/or $\bar{r}$ are large. This information is important because the estimation methodology is based on asymptotic assumptions (recall that "as n$\rightarrow\infty$" the derivations hold). In many statistical applications, ∞ is not much beyond 30, and in our analysis also, we see that the statistic is generally well-behaved by the time n reaches 30, at least for p$\geq$5 and $\bar{r}\geq$0.5, or even when p$\geq$2 if $\bar{r}\geq$0.7. These results are encouraging, and should provide assurance to researchers, given that many studies reported in the literature are based on small samples.

Conducting one final analysis, we entered the terms n, p, and $\bar{r}$ into a multiple regression to predict ASE in Table 1 (the equal standard deviations case). Note that the effect of $\bar{r}$ ($\hat{\beta}$ =-.770, $\hat{\eta}^2$ =0.803) dominates that of p ($\hat{\beta}$ =-0.269, $\hat{\eta}^2$ =0.098) and n ($\hat{\beta}$ =-0.271, $\hat{\eta}^2$ =0.099). This finding is sensible given that alpha is an index intended to represent the internal consistency of the items in a scale. All coefficients are negative, as they should be, meaning that standard errors decrease, or

precision of estimation of alpha increases, with an increase in any term. Increasing item intercorrelation appears to be the most effective means of reducing alpha's standard error.

We have some understanding now of the effects of n, p, and $\bar{r}$ on the ASE, and therefore, the confidence intervals. Comparing Table 1 to Tables 2 and 3 allows us to explore the impact of item heteroscedasticity in the standard deviations across the p items. The condition of constant variances, that reported in Table 1, is the special case of parallel tests. The scenarios of varying variances, those reported in Tables 2 and 3, are probably more representative of real data.

The results in Tables 2 and 3 resemble those in Table 1, with the main difference being that the standard errors for the heterogeneity cases begin smaller for small p than their counterparts in Table 1 but are approximately the same magnitude for larger p. As the mean intercorrelation approaches 0.50 or 0.60, the differences diminish. Conversely, we see that heteroscedasticity has a large effect on alpha, even when $\bar{r}$ is high, e.g., for p=2, the homoscedastic alpha result is .949, whereas in the extreme variance heterogeneity condition, reliability drops to .584.

At this point, we have illustrated the confidence intervals across a large number of conditions for varying n, p, $\bar{r}$, and variances. This analytical exercise has hopefully imparted a sense of how each of these components affect measurement reliability.

Study 2: Exploring the Potential Biasing Effects of Covariance Heterogeneity on Standard Error

One additional rationale for reporting alpha's standard error is that it conveys unique information about scales. Even in cases when alpha reliabilities are deemed sufficiently high (greater than .70), significant differences in scales can exist. For instance, scale dimensionality should not affect the magnitude of alpha, but it can be a source of variance in the magnitude of alpha's standard error. This new estimate of standard error, ASE, is the first to not require the assumption of compound symmetry. In study 2, we investigate covariance heterogeneity as a potential source of bias, and to determine ASE's robustness.

Study Design. In study 2, we conducted a Monte Carlo simulation, generating 1000 replications. The study design incorporated the same levels of sample size and a reduced set of *p* and $\bar{r}$, in addition to the new covariance heterogeneity factor. We designed the new factor with three levels. The first level was designed to replicate the case of parallel tests, serving as a comparative benchmark. Specifically, all interitem correlations were constant, set to $\bar{r}$ (that is item homogeneity). The other two levels of this factor considered different means by which inter-item correlations might evince heterogeneity, rather than uniformity. The first scenario we modeled of nonhomogeneous interitem correlations would be for there to exist one (or more) 'poor' items. A classic step in scale development is to compute item-total correlations as initial diagnostics to detect items that do not load significantly or appear to be poor indicators of the construct at the focus of the research.

Operationally, when the covariance heterogeneity was due to one poor item, the item intercorrelations were created as follows. If all p items were consistent and homogeneous, the "sum" of the correlations would be p(p-1) x $\bar{r}$. In the presence of one poor item, two-thirds of "sum" was divided by (p-1)(p-2) and this value was evenly distributed over the (p-1)(p-2) elements in the matrix representing the intercorrelations among the p-1 good items. One third of "sum" was divided by 2(p-1) and placed in the 2(p-1) elements in the matrix representing the correlations between the one bad item and the (p-1) other good items. Stated another way, "sum" = [p(p-1)x $\bar{r}$] is divided by [2(p-1)(p-2) + 2(p-1)] and the result is = "weight." The intercorrelations among the (p-1)(p-2) good items were assigned "2weight," and the correlations between the good items and the bad item were assigned "weight." This assignment preserved the equality of the $\bar{r}$ across conditions. For example, for p=4 and $\bar{r}$ =.5, the matrix is: $\begin{bmatrix} 1.0 & .6667 & .6667 & .3333 \\ .6667 & 1.0 & .6667 & .3333 \\ .6667 & .6667 & 1.0 & .3333 \\ .3333 & .3333 & .3333 & 1.0 \end{bmatrix}$.

The final covariance heterogeneity condition considered the case of multidimensional scales. Because alpha is a measure of internal consistency reliability and therefore is not an assessment of

scale unidimensionality, scales with more than one underlying factor may still yield high levels of alpha [8]. Coefficient α is intended only for homogeneous scales, i.e., those that measure a single construct. However, the components of the formula clearly indicate that whether the items achieve the status of internal consistency, it is the extent of covariability, perhaps most easily seen in the equation using the mean item correlation, that drives the size of α, along with p, the scale length. That is, when the equation for α is written as: $\alpha = \frac{p\bar{r}_{ij}}{1+(p-1)\bar{r}_{ij}}$, one can see that constant correlations at $\bar{r}_{ij}$, or correlations centered about $\bar{r}_{ij}$ would yield comparable estimates. Thus, the introduction of a factor in a simulation that varies whether p items are homogeneous or heterogeneous should have no impact on the estimate of coefficient alpha itself.

However, scale multidimensionality should affect the size of alpha's standard error. Recall the equation for the standard error included terms that were quadratic functions of the covariance matrix. Heterogeneity in item correlation patterns would be exaggerated in these functions.

In terms of this simulation, when covariance heterogeneity was due to an underlying structural multidimensionality, we operationalized the item correlations as follows, to create two factors. For even values of p, the number of items loading on each factor was equal. For odd values of p, the extra variable loaded on the first of the two factors. Covariance heterogeneity was created by making the values within the clusters of items that loaded on common factors equal to twice that of the inter-factor item correlations. Thus, for p=4 items, the first two items loaded on the first factor, and the last two items loaded on the last factor. The correlations $r_{1,2}$ and $r_{3,4}$ were twice the size of the cross-factor correlation, $r_{1,3}$, $r_{1,4}$, $r_{2,3}$, $r_{2,4}$. For example, for p=4 and $\bar{r}$ =.5, the population correlation matrix was as follows: [9]

[8] Gerbing and Anderson (1988, p.190) state, "regardless of the dimensionality of the scale, its reliability tends to increase as the average off-diagonal item correlation increases and/or the number of items increases." Thus researchers are frequently counseled that a high alpha is not necessarily indicative of a unidimensional underlying scale.

[9] The multi-dimensionalities of these conditions were confirmed by exploratory factor analyses using promax rotation.

$$\begin{bmatrix} 1.0 & .750 & .375 & .375 \\ .750 & 1.0 & .375 & .375 \\ .375 & .375 & 1.0 & .750 \\ .375 & .375 & .750 & 1.0 \end{bmatrix}.$$

The multi-dimensionality or covariance heterogeneity required us to reduce the number of levels of *p* and $\bar{r}$ accordingly. The minimum number of items necessary in order to generate two stable underlying factors is four, therefore, the cases of p=2 and p=3 are not included in this design. Levels of $\bar{r}$ were also implicated. Scale items loading on the same factor must have higher correlations with each other than with items loading on other factors, so low levels of $\bar{r}$ are not possible. Therefore, $\bar{r}$ =0.0, $\bar{r}$ =0.1, $\bar{r}$ =0.2 are excluded. High levels of $\bar{r}$ such as $\bar{r}$ =0.8, $\bar{r}$ =0.9 and $\bar{r}$ =1.0 are similarly constrained, given that factor analyses would extract a single underlying factor. To summarize, the primary objective of study 2 was to examine the nature of the covariance heterogeneity effect and its relation to the other factors noted in study 1. The resulting design was a fully crossed 7 (n) x 7 (*p)* x 5 ($\bar{r}$) x 3 (covariance heterogeneity) factorial design.

--- Table 4 and Figure 1 go about here ---

Results. The focal criterion variables under examination in this study were alpha and its standard error. The modeling effects for these variables are displayed in Table 4. With regards to alpha, we are encouraged to observe the high level of congruence between the overall effects in study 2 and the analytical study 1. The significant main effect of $\bar{r}$ exudes the greatest influence on alpha, as determined by the $\hat{\eta}^2$ results, although, as expected, we observe that the alpha results are attenuated somewhat by sampling error. The effect of covariance heterogeneity is also significant, but the effect is small. This result supports conventional admonitions against using coefficient alpha as a diagnostic for assessing unidimensionality. Even when item intercorrelations vary by a great degree (due to multi-dimensionality or bad items), alpha was affected only nominally.

In contrast, the results pertaining to the standard error of alpha were more complex. First, we note that the standard error results with regards to the n, *p* and $\bar{r}$ effects were consonant with the study

1 results; standard errors decrease as n, p and $\bar{r}$ increase. In addition, we observe a significant and interpretable result of covariance heterogeneity (beta=0.073, $\hat{\eta}^2$=0.037). Covariance heterogeneity increases the standard error, although the magnitude of the bias is small (see Figure 1). In the case where a single item deviates from compound symmetry, the standard error estimates approximate the covariance homogeneity case, but are consistently significantly higher, for all values of p, n and $\bar{r}$ (the bias decreases as sample size increases). The standard error estimates depart further from the homogeneity condition in the case of two-dimensional scales, although again the magnitude of this bias is small. The degree of bias induced by either heterogeneity effect increases as $\bar{r}$ increases, e.g., when $\bar{r}$=0.7, the standard error estimates were more distinct than when $\bar{r}$=0.3, for all levels of n and p. To conclude, the results of study 2 demonstrate that covariance heterogeneity exerts minimal effect on the estimate of alpha, but has somewhat deleterious effects on the estimate of alpha's standard error.

As measurement experts have long averred, heterogeneity among the items is not desired in scale construction. The arguments have traditionally been theoretical, however, based on the concepts of domain sampling, internal consistency, item homogeneity and parsimony. Now, with the standard error, we see that as items are less uniformly correlated, estimator variances increase, hence confidence intervals widen, providing empirical support for the popular notion that there is a cost affiliated with generating multidimensional scales.

Study 3: Comparatively Assessing ASE against Competing Derivations

The primary objective in study 3 was to provide some context for the small biasing effect of covariance heterogeneity observed in study 2. As discussed, our derivation of alpha's standard error does not require tau equivalence among the scale items, a critical assumption of earlier derivations. In fact, previous empirical work has demonstrated the insufficiency of many of these test statistics under

violations of these assumptions (Barchard & Hakstian, 1997b). In the present study, we offer the first direct comparative test of the performance of these statistics.

Study Design. In study 3, the principal objective is to compare competing derivations of alpha's standard error over a comprehensive range of factors, including covariance heterogeneity. In Table 5, we offer an analytical framework comparing alternative forms of alpha's standard error. We see the two most prominent derivations preceding ASE, Feldt (1965) and Hakstian & Wahlen (1976), along with a standard error offered by Nunnally (Nunnally & Bernstein 1994) and a test-retest modified statistic attributed to Lord & Novick (1968) and revisited by Mendoza, Stauffer and Stauffer (2000). The final standard error statistic included in our simulation was a split half statistic (Charter 2000).

--- Table 5 and Figure 2 go about here ---

The design in study 3 was again a Monte Carlo simulation with 1000 replications. The factors of interest in this study were n, p, $\bar{r}$ and covariance heterogeneity. Unlike the focus of studies 1 and 2, where the goal was to document the behavior of alpha's standard error over a wide range of these components, the focus of study was to directly compare competing derivations of alpha's standard error. We simplified the design to include sample sizes of: n=30, 50, 100, 200; levels of p=5, 7; mean item intercorrelations ranging from 0.4 to 0.7. Finally, covariance heterogeneity was two levels: the unidimensional case representing compound symmetry and a multidimensional case representing two underlying factors. We computed the confidence interval estimates for ASE and the five other primary standard error statistics reported in the literature (equations in Table 5). To summarize, the design was a 4 (n) x 2 (p) x 4 ($\bar{r}$) x 2 (covariance heterogeneity) full factorial design.

Results. The dependent variables of interest in this comparative study differ slightly from those examined thus far. First, we assessed the degree of bias in each confidence interval (i.e., proportion of observations that contain true alpha). Second, we calculated the widths of the competing intervals. Ideally, one desires an interval that is both accurate, that is to say it contains an unbiased proportion of

true score alphas (i.e., 95% accuracy) *as well as* precise, that is to say an interval that is of a moderate width. These two complementary facets serve as the basis for the competing analysis in study 3.

Figure 2 displays the plots of the main effect of sample size, as an illustration, on the dependent variable of "proportion of confidence intervals containing the true underlying alpha." From these plots, we see that the ASE estimate most often contains the true underlying alpha at a 95% level of confidence. Examining the performance of the other standard error statistics displayed in the plots, we observe that the Feldt, Hakstian and Wahlen and Nunnally formulations all appear significantly less accurately. The intervals based on the split half and Lord and Novick standard error statistics exhibit a different bias. These interval estimates contain the true score significantly more often than their stated 95% level of accuracy, indicating a bias perhaps due to an imprecise interval estimate (we will examine this possibility shortly).

Examining the comparative performance across p, we again observe that the ASE statistic most closely approximates its stated 95% level of accuracy, and there is no bias attributable to scale length.[10] We observe the same pattern of bias in the other interval estimates, with Feldt, Hakstian and Wahlen and Nunnally all significantly less accurate than their stated 95% level of confidence and the split half and Lord and Novick estimates exhibit their same positive bias.

With respect to item intercorrelation, the ASE estimate is the least affected by changes in $\bar{r}$. The Nunnally estimation appears erratic, suggesting its serious vulnerabilities across varying ranges of item correlations. The Feldt and Hakstian and Wahlen intervals are also less accurate for higher levels of item intercorrelation.

The covariance factor indicates that in the multidimensional case, all interval estimates lose accuracy except ASE, which appears robust. The Feldt and Hakstian and Wahlen 95% estimates contain the true alpha less than 90% of the time in the multidimensional case when $\bar{r}$ =0.7.

[10] More figures are available from the authors.

Next we consider the precision of the confidence interval estimates by examining confidence interval width. With respect to sample size, Figure 2 shows distinctions across the interval widths. These differences are reduced slightly as n increases, but clearly the ASE, Feldt and Hakstian and Wahlen estimates appear significantly more precise compared to the other estimates. This result is particularly striking given the accuracy results reported previously. It thus appears that the bias in the split half and Lord and Novick estimates is due to their imprecision.

For all of the estimates, interval widths narrow as scale length increases, but ASE, Feldt and Hakstian and Wahlen remain significantly more precise. For item intercorrelation, we observe the same superior performance of ASE, Feldt, and Hakstian and Wahlen estimates, although the distinctions are not as pronounced.

To conclude, the results of study 3 offer strong evidence for the performance of the ASE estimate of alpha's standard error as determined by estimating both its accuracy and performance. In comparison to earlier derivations, this estimate was the most accurate across sample size, scale length, item intercorrelation and covariance heterogeneity. The performance of this estimate is particularly compelling given that our simulation modeled fairly extreme levels of covariance heterogeneity. Further, the analysis of the interval widths revealed that this estimate was also among the most precise. These results indicate alpha's standard error, ASE, is suitable for use in empirical research.

Two Independent Samples

We had mentioned previously the early research that attempted to provide inferential statistics for alpha. Different articles considered different scenarios, e.g., binary scales (Feldt, 1965; Feldt, 1969), pairs of independent (Feldt, 1965) and dependent samples (Alsawalmeh & Feldt, 1994; Feldt, 1980; Woodruff & Feldt, 1986), multiple independent samples (Hakstian & Whalen, 1976), and so on, but each carried restrictive assumptions, such as that of parallel test forms, and were not robust to their violation (e.g., Barchard & Hakstian, 1997a).

A natural question arises as to the extension of the approach in this paper to that of multiple (independent) samples. It is straightforward to generalize the variance estimate of equation (2) to that for the case of comparing coefficient alpha estimates obtained from samples representing two different, independent populations. A conservative (slightly larger) confidence interval would be obtained by estimating a standard error based on no assumption of homogeneity of variance (more precisely, equal covariance matrices) across samples:

$$(\hat{\alpha}_1 - \hat{\alpha}_2) \pm (1.96)\left(\sqrt{\frac{Q_1}{n_1} + \frac{Q_2}{n_2}}\right). \quad (4)$$

Armed with equation (4), the researcher comparing alpha estimates across two samples would examine whether zero was included in the multi-sample confidence interval. In such cases, the researcher may conclude with statistical confidence that the extent of measurement error affecting both samples was equal. This new test holds great promise as a means of statistically assessing group differences in reliability (i.e., differences in organizational or functional teams, etc.).

The SAS and SPSS Code

The SAS program to compute alpha, its standard error, and the confidence interval is presented in Appendix A. If a researcher simply relied on SAS's current Proc Corr Alpha option, the standard error and confidence intervals would not be available. SPSS produces analogously limited results. Now the SAS and SPSS code is now publicly available, it is easy to implement, and it yields precise estimates (i.e., for any combination of p, n, varying r's, etc.).

To use the program, the user specifies the number of items in the scale and the sample size. The program is versatile in allowing the inputs to be raw data, a correlation or covariance matrix produced by SAS's Proc Corr in an output statement, or a matrix input by hand (e.g., computed via another package). The standardized and unstandardized alphas are produced, along with the

confidence intervals, and a z-test of a hypothesis about alpha, and its corresponding p-value to denote significance. To offer versatility for different users, we also present the SPSS code in Appendix B.

Best Practices

We close with a few simple prescriptions regarding the reporting of coefficient alpha. First, researchers should follow the procedures outlined in Figure 3 in reporting all alpha statistics. We propound that conventional standard practice of merely reporting an alpha point estimate is no longer sufficient, when additional diagnostic information is clearly accessible. Alpha estimates should be supplemented with the standard error, ASE, and confidence interval. In Table 6, we see a summary of the nature of significant relationships between alpha and its component factors. As revealed in our studies and recapitulated in this summary, the factors influencing alpha do not always exert influence on alpha's standard error. Therefore, each component contributes uniquely to understanding the fundamental internal consistency properties of a scale and both should be reported. Also summarized in Table 6, we see the comparative performances of competing derivations of alpha's standard error. The findings of these studies indicate that ASE is the most well-behaved error statistic, as determined by both accuracy and precision. These findings indicate that it is suitable for use in empirical research, and thus should be reported. This more thorough reporting will enable the reader to independently assess the magnitude of the alpha, or conversely, the likely impact of measurement error, in the subsequent analytical use of the scale.

--Figure 3 goes about here--

Our analytical investigation suggests that researchers seeking to improve the stability of their measures via a reduction in the estimated standard errors should first seek to improve the inter-correlation among their scale items. Improvements in item inter-correlation lead to greater diminution of standard error estimates than is accomplished by adding sample or scale items. Also, many applied researchers will find the confidence interval procedure for comparing independent sample alphas

useful. Employing equation 4, researchers can now reach statistically-supported conclusions regarding the relative soundness of scale reliability across multiple samples, i.e., various populations of job applicants or employees' scores on selection or performance tests, different cultures' uses of psychological scales, etc.

Conclusion

The equations, analytical illustrations and empirical results presented in this paper should have great applicability to the behavioral researcher. The inferential testing requires assumptions of multivariate normality, but this assumption is fairly standard, required of many statistical modeling procedures, e.g., the widely pervasive LISREL-fitting of structural equations models (cf. Jöreskog & Sörbom 1996). Furthermore, early reports indicate that these alpha-related statistics might prove to be fairly hardy; Yuan and Bentler (2002) have shown in their extensive exploration of skewness and kurtosis that these indices are fairly robust to violations of the assumption of multivariate normality. In addition, previous attempts to derive inferential statistics for special reliability indices (e.g., Charter, 2000 and Drewes, 2000 for Spearman-Brown; Feldt, 1965 for K-R20; Feldt, Woodruff & Salih, 1987 with an analysis of variance approach; and Mendoza, Stafford and Stauffer, 2000 using selected samples and validity coefficients) may now be subsumed into this more general, elegant approach.

The results also provide largely good news about the behavior of the ASE. Unless one is working under extreme conditions, e.g., p=2 and r=0.0 (or very small), the alpha standard errors and confidence intervals function in a predictably robust manner, even for small samples (n=30).[11] Even when covariance heterogeneity is severe, we found that ASE performs quite well. Empirical researchers are aware that, in practice, developing scales with perfectly homogenous interitem correlations is a nearly impossible task. As studies 2 and 3 indicate, this heterogeneity does impact the

[11] Note these results suggest greater robustness than that found recently in Feldt & Ankenmann 1999, where sample sizes were required on the order of 100 to 200 and sometimes approached 1000 for small p.

precision of the estimate of alpha via its standard error. An interesting corollary to the covariance heterogeneity bias in standard errors was revealed through our results. Our findings demonstrate the increasing influence of such heterogeneity as $\bar{r}$ increases. Because researchers desire items with high intercorrelations, researchers should therefore be aware that attaining high levels of $\bar{r}$ could yield higher standard error estimates (i.e., more error-laden measures) if considerable covariance heterogeneity exists. Although study 2 found evidence for this bias, in sum, these findings suggest that our formulation of standard error is significantly less susceptible to bias stemming from covariance heterogeneity compared to earlier derivations (study 3).

Given these results and the ready availability of our program, our recommendation is that every alpha should be reported with its confidence interval to allow the reader to assess the size of the reliability index. Once inferential statistics are available, it becomes no longer sufficient to subjectively judge reliability solely on the basis of a point estimate. Computing estimates of ASE standard errors and forming confidence intervals around coefficient alpha provides more diagnostic information to the researcher, information that can be used in such tasks as the detection of differences between tests and estimation of test bias due to measurement error.

References

Aiken, L.R. (2002). *Psychological Testing and Assessment* (11th ed.), Boston, MA: Allyn and Bacon.

Allen, M.J. & Yen W.M. (1979). *Introduction to Measurement Theory*, Monterey, CA: Brooks/Cole.

Alsawalmeh, Y.M. & Feldt, L.S. (1994). "A Modification of Feldt's Test of the Equality of Two Dependent Alpha Coefficients," *Psychometrika*, 59, 49-57.

Anastasi, A. & Urbina, S. (1996). *Psychological Testing* (7th ed.), New York: Prentice Hall.

Barchard, K.A. & Hakstian, R. (1997a). "The Effects of Sampling Model on Inference with Coefficient Alpha," *Educational and Psychological Measurement*, 57, 893-905.

Barchard, K.A. & Hakstian, R. (1997b). "The Robustness of Confidence Intervals for Coefficient Alpha Under Violation of the Assumption of Essential Parallelism," *Multivariate Behavioral Research*, 32, 169-191.

Bentler, P.M. & Raykov, T. (2000). On Measures of Explained Variance in Nonrecursive Structrual Equations Models," *Journal of Applied Psychology*, 85, 125-131.

Chan, K.Y., Drasgow, F., & Sawin, L.L. (1999). "What is the Shelf Life of a Test: The effect of Time on the Psychometrics of a Cognitive Ability Test Battery," *Journal of Applied Psychology*, 84, 610-619.

Charter, R.A. (2000). "Confidence Interval Formulas for Split-Half Reliability Coefficients," *Psychological Reports*, 86, 1168-1170.

Colquitt, J.A. (2001). "On the Dimensionality of Organizational Justice: A Construct Validation of a Measure," *Journal of Applied Psychology*, 86, 386-400.

Cortina, J. M. (1993). "What is Coefficient Alpha? An Examination of Theory and Applications," *Journal of Applied Psychology*, 78, 98-104.

Cortina, J.M. & Dunlap, W.P. (1997). "On the Logic and Purpose of Significance Testing," *Psychological Methods*, 2, 161-172.

Cronbach, L. J. (1951). "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, 16,

297-334.

Cronbach, L.J. (1988). "Internal Consistency of Tests: Analyses Old and New," *Psychometrika*, 53, 63-70.

Drewes, Donald W. (2000). "Beyond the Spearman-Brown: A Structural Approach to Maximal Reliability," *Psychological Methods*, 5 (2), 214-227.

Feldt, L. S. (1965). "The Approximate Sampling Distribution of Kuder-Richardson Reliability Coefficient Twenty," *Psychometrika*, 30, 357-370.

Feldt, L.S. (1969). "A Test of the Hypothesis that Cronbach's Alpha or Kuder-Richardson Coefficient is the Same for Two Tests Administered to the Same Sample," *Psychometrika*, 34, 363-373.

Feldt, L.S. (1980). "A Test of the Hypothesis that Cronbach's Alpha Reliability Coefficient is the Same for Two Tests Administered to the Same Sample," *Psychometrika*, 45, 99-105.

Feldt, L.S. & Ankenmann, R.D. (1999). "Determining Sample Size for a Test of the Equality of Alpha Coefficients when the Number of Part-Tests is Small," *Psychological Methods*, 4, 366-377.

Feldt, L.S., Woodruff D.A., & Salih, F.A. (1987). "Statistical Inference for Coefficient Alpha," *Applied Psychological Measurement*, 11, 93-103.

Ghorpade, J., Hattrup, K., & Lackritz, J.R. (1999). "The Use of Personality Measures in Cross-cultural Research: A test of Three Personality Scales Across Two Countries," *Journal of Applied Psychology*, 84, 670-679.

Gregory, R.J. (2001). *Psychological Testing: History, Principles, and Applications* (3rd ed.) Boston: Allyn and Bacon.

Hakstian, A.R. & Barchard, K.A. (2000), "Toward More Robust Inferential Procedures for Coefficient Alpha Under Sampling of Both Subjects and Conditions," *Multivariate Behavioral Research*, 35, 427-456.

Hakstian, A.R. & Whalen T.E. (1976). "A K-Sample Significance Test for Independent Alpha Coefficients," *Psychometrika*, 41, 219-231.

Hogan, T. P., Benjamin, A. & Brezinski, K.L. (2000). "Reliability Methods: A Note on the Frequency of Use of Various Types," *Educational and Psychological Measurement*, 60, 523-531.

Jöreskog, K. G. and Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Chicago: Scientific Software International.

Journal of Consumer Psychology (2001). "Special Issue on Methodological and Statistical Concerns of the Experimental Behavioral Researcher," Dawn Iacobucci (ed.), 10, 55-62.

Kinicki, A., McKee-Ryan, F.M., Schriesheim, C. & Carson, K.P. (2002). "Assessing the Construct Validity of the Job Descriptive Index: A Review and Meta-analysis," *Journal of Applied Psychology*, 87, 14-32.

Kristof, W. (1963). "The Statistical Theory of Stepped-Up Reliability Coefficients when a Test has been Divided into Several Equivalent Parts," *Psychometrika*, 28, 221-238.

Li, Heng (1997). "A Unifying Expression for the Maximal Reliability of a Linear Composite," *Psychometrika*, 62, 245-249

Li, H., Rosenthal, R. & Rubin, D.B. (1996). "Reliability of Measurement in Psychology: From Spearman-Brown to Maximal Reliability," *Psychological Methods*, 1, 98-107.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theory of Mental Test Scores*, New York: McGraw-Hill.

Mendoza, J. L., Stafford, K.L., & Stauffer, J.M. (2000). "Large-Sample Confidence Intervals for Validity and Reliability Coefficients," *Psychological Methods*, 5, 356-369.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory* (3rd ed.), New York: WCB/McGraw-Hill.

Osburn, H. G. (2000). "Coefficient Alpha and Related Internal Consistency Reliability Coefficients," *Psychological Methods*, 5, 343-355.

Raykov, T. (1998a). "A Method for Obtaining Standard Errors and Confidence Intervals of Composite Reliability for Congeneric Items," *Applied Psychological Measurement*, 22, 369-374.

Raykov, T. (1998b). “Coefficient Alpha and Composite Reliability with Interrelated Nonhomogeneous Items,” *Applied Psychological Measurement*, 22 (4), 375-385.

Robinson, J.P. (ed.), Shaver, P.(ed.), Wrightsman, L.S. (ed.). (1991). *Measures of Personality and Social Psychological Attitudes*, New York: Academic Press.

Schmidt, F.L. & Hunter, John E. (1996). “Measurement Error in Psychological Research: Lessons from 26 Research Scenarios,” *Psychological Methods*, 1, 199-223.

van Zyl, J. M., Heinz. N., & Nel, D. G. (2000). “On the Distribution of the Maximum Likelihood Estimator of Cronbach’s Alpha,” *Psychometrika*, 65, 271-280.

Woodruff, D.J. & Feldt, L.S. (1986). “Tests for Equality of Several Alpha Coefficients when Their Sample Estimates are Dependent,” *Psychometrika*, 51, 393-413.

Yuan, K. & Bentler, P.M. (2002). “On Robustness of the Normal-Theory Based Asymptotic Distributions of Three Reliability Coefficient Estimates,” *Psychometrika*, 67, 251-259.

Appendix A

The SAS Code to Compute Alpha, Standard Error, Z, Confidence Intervals[12,13]

Option 1: User Provides a Correlation or Covariance Matrix

As the first four comments indicate, the user must provide 4 things: p, n, their matrix of item covariances or correlations, and a value to use as the hypothesized benchmark in the z-test.

```
proc iml;
*TO USER: you need to fill in the hypothesized
value you want your alpha tested against;          hypalpha = { 0.7 };
*TO USER: you need to fill in the number of items
you have in your scale;                            numbitem = { 3 };
*TO USER: you need to fill in your sample size;    numbsubj = { 100 };
*TO USER: you need to cut and
paste your item covariance matrix into the following
form (i.e., begin and end with braces, add commas to
delineate matrix rows);                            itemcov =
                                                   { 1 .5 .5 ,
                                                     .5 1 .5 ,
                                                     .5 .5 1 };

*next are analyses;
one=j(numbitem,1); jtphij = (one`)*itemcov*one;
myalpha = 1 – ((trace(itemcov))/jtphij); myalpha = (numbitem / (numbitem -1)) *myalpha;
trphisq = trace(itemcov*itemcov); trsqphi = (trace(itemcov))**2;
jtphisqj = (one`)*(itemcov*itemcov)*one; omega = jtphij*(trphisq+trsqphi);
omega = omega-(2*(trace(itemcov))*jtphisqj); omega = (2/(jtphij**3))*omega;
s2 = (numbitem**2) / ((numbitem-1)**2); s2 = s2*omega;
se = sqrt(s2/numbsubj); z = (myalpha-hypalpha)/se; pv = 1-probnorm(z);
cimin95 = myalpha – (1.96*se); cimax95 = myalpha + (1.96*se);
print 'Your Covariance Matrix was:'; print itemcov;
print 'Your number of items and sample size were:' numbitem numbsubj;
print 'Your coefficient alpha is:' myalpha;
print 'The z score for alpha and its p-value are:' z pv;
print 'The lower and upper 95% confidence limits follow:' cimin95 cimax95;
if cimin95 < .00 then print 'You should report your confidence interval as:  0.00 to ' cimax95;
*scale cov matrix to corr matrix; s=diag(itemcov); s=sqrt(s); s=s**(-1);  itemcov=s`*itemcov*s;
jtphij = (one`)*itemcov*one;
myalpha = 1 – ((trace(itemcov))/jtphij);  myalpha = (numbitem / (numbitem -1)) *myalpha;
trphisq = trace(itemcov*itemcov); trsqphi = (trace(itemcov))**2;
jtphisqj = (one`)*(itemcov*itemcov)*one;  omega = jtphij*(trphisq+trsqphi);
```

[12] An interested user need not type in this syntax. Our program is also available on our website (website address to go here; to Reviewers: the website is not yet available).

[13] The program uses SAS's IML module. IML stands for "interactive matrix language," but this program is not run interactively; just submit it the way you would any SAS job.

```
omega = omega-(2*(trace(itemcov))*jtphisqj); omega = (2/(jtphij**3))*omega;
s2 = (numbitem**2) / ((numbitem-1)**2); s2 = s2*omega;
se = sqrt(s2/numbsubj); z = (myalpha-hypalpha)/se; pv = 1-probnorm(z);
cimin95 = myalpha – (1.96*se); cimax95 = myalpha + (1.96*se);
print 'Your Correlation Matrix was:'; print itemcov;
print 'Standardized coefficient alpha equals:' myalpha;
print 'Your z score and its p-value are:' z pv;
print 'The lower and upper 95% confidence limits are:' cimin95 cimax95;
if cimin95 < .00 then print 'You should report your confidence interval as:  0.00 to' cimax95;
quit; run;
```

Option 2: User Reads in Raw Data

1. Prior to the "proc iml;" statement in the program above, insert:

```
data myabc; input x1 x2 x3; cards;
1 1 0
2 2 3
2 1 3
…
4 2 1
run;
```

2. After the "proc iml;" statement above, insert:

```
use myabc var {x1 x2 x3};
read all var {x1 x2 x3} into x;
```

3. Then delete the "itemcov = { … };" statement.

4. After the comment "*next are analyses;" insert:

```
bigone=j(numbsubj,1); means=((bigone`)*x)/numbsubj;
xd=x-(bigone*means); itemcov = (1/(numbsubj-1)) * ((xd`)*xd);
```

Option 3: User Reads Raw Data into Proc Corr, Produces a Correlation or Covariance Matrix to be used in Proc IML

1. Prior to the "proc iml;" statement in the program above, insert:

```
data myabc; input x1 x2 x3; cards;
1 1 0
2 2 3
2 1 3
…
4 2 1
proc corr cov outp=mycorrs; var x1 x2 x3; run;
```

2. After the "proc iml;" statement above, insert:

```
use mycorrs var {x1 x2 x3};
read point {1 2 3} var {x1 x2 x3} into itemcov;
```

Appendix B

The SPSS Code to Compute Alpha, Standard Error, Confidence Intervals

```
* USER: fill in #items in scale, sample size, covariance matrix.

matrix.
compute numbitem = 3.
compute numbsubj = 100.
compute itemcov = { 1.0, .5, .5;  .5, 1.0, .5;  .5, .5, 1.0}.

compute one=make(numbitem,1,1).
compute jtphij=transpos(one).
compute jtphij=jtphij*itemcov.
compute jtphij=jtphij*one.
compute trmy=trace(itemcov).
compute trmy=trmy/jtphij.
compute myalpha=1-trmy.
compute nn1=numbitem-1.
compute nn1=numbitem/nn1.
compute myalpha=nn1*myalpha.
compute trphisq=itemcov*itemcov.
compute trphisq=trace(trphisq).
compute trsqphi=trace(itemcov).
compute trsqphi=trsqphi**2.
compute ttp=itemcov*itemcov.
compute jtphisqj=transpos(one).
compute jtphisqj=jtphisqj*ttp.
compute jtphisqj=jtphisqj*one.
compute omega=trphisq+trsqphi.
compute omega=jtphij*omega.
compute omegab=trace(itemcov).
compute omegab=omegab*jtphisqj.
compute omega=omega-(2*omegab).
compute omega=(2/(jtphij**3))*omega.
compute s2=(numbitem**2) / ((numbitem-1)**2).
compute s2=s2*omega.
compute se=sqrt(s2/numbsubj).
compute cimin95=myalpha-(1.96*se).
compute cimax95=myalpha+(1.96*se).
print myalpha /format ="f8.3"/title= 'Your coefficient alpha is:'.
print cimin95  /format ="f8.3"/title= 'The lower 95% confidence limit follows:'.
print cimax95 /format ="f8.3"/title= 'The upper 95% confidence limit follows:'.
end matrix.
```

Tables 1-6 & Figures 1-3

Table 1

Coefficient Alpha and Standard Errors with Standard Deviations Equal One

Each cell contains the α followed by alpha's standard error, ASE.

	n	$\bar{r}$: 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
p=2	30	.000, .365	.182, .299	.333, .243	.462, .196	.571, .157	.667, .122	.750, .091	.824, .064	.889, .041	.947, .019	1.00, 0.00
	50	.000, .283	.182, .231	.333, .189	.462, .152	.571, .121	.667, .094	.750, .071	.824, .050	.889, .031	.947, .015	1.00, 0.00
	100	.000, .200	.182, .164	.333, .133	.462, .107	.571, .085	.667, .067	.750, .050	.824, .035	.889, .022	.947, .011	1.00, 0.00
	200	.000, .141	.182, .116	.333, .094	.462, .076	.571, .061	.667, .047	.750, .035	.824, .025	.889, .016	.947, .007	1.00, 0.00
p=3	30	.000, .316	.250, .237	.429, .181	.562, .138	.667, .105	.750, .079	.818, .057	.875, .040	.923, .024	.964, .011	1.00, 0.00
	50	.000, .245	.250, .184	.429, .140	.562, .107	.667, .082	.750, .061	.818, .045	.875, .031	.923, .019	.964, .009	1.00, 0.00
	100	.000, .173	.250, .130	.429, .099	.562, .076	.667, .058	.750, .043	.818, .031	.875, .022	.923, .013	.964, .006	1.00, 0.00
	200	.000, .122	.250, .092	.429, .070	.562, .054	.667, .041	.750, .031	.818, .022	.875, .015	.923, .009	.964, .004	1.00, 0.00
p=5	30	.000, .289	.357, .186	.556, .128	.682, .092	.769, .067	.833, .048	.882, .034	.921, .023	.952, .014	.978, .006	1.00, 0.00
	50	.000, .224	.357, .144	.556, .099	.682, .071	.769, .052	.833, .037	.882, .026	.921, .018	.952, .011	.978, .005	1.00, 0.00
	100	.000, .158	.357, .102	.556, .072	.682, .050	.769, .036	.833, .026	.882, .019	.921, .012	.952, .008	.978, .003	1.00, 0.00
	200	.000, .112	.357, .072	.556, .050	.682, .036	.769, .026	.833, .019	.882, .013	.921, .009	.952, .005	.978, .002	1.00, 0.00
p=7	30	.000, .279	.438, .157	.636, .101	.750, .069	.824, .049	.875, .035	.913, .024	.942, .016	.966, .010	.984, .004	1.00, 0.00
	50	.000, .216	.438, .122	.636, .079	.750, .054	.824, .038	.875, .027	.913, .019	.942, .012	.966, .007	.984, .003	1.00, 0.00
	100	.000, .152	.438, .086	.636, .056	.750, .038	.824, .027	.875, .019	.913, .013	.942, .009	.966, .005	.984, .002	1.00, 0.00
	200	.000, .108	.438, .061	.636, .039	.750, .027	.824, .019	.875, .014	.913, .009	.942, .006	.966, .004	.984, .002	1.00, 0.00
p=10	30	.000, .272	.526, .129	.714, .077	.811, .051	.870, .035	.909, .025	.938, .017	.959, .011	.976, .007	.989, .003	1.00, 0.00
	50	.000, .211	.526, .100	.714, .060	.811, .040	.870, .027	.909, .019	.938, .013	.959, .009	.976, .005	.989, .002	1.00, 0.00
	100	.000, .149	.526, .071	.714, .043	.811, .028	.870, .019	.909, .014	.938, .009	.959, .006	.976, .004	.989, .002	1.00, 0.00
	200	.000, .105	.526, .050	.714, .030	.811, .020	.870, .014	.909, .010	.938, .007	.959, .004	.976, .003	.989, .001	1.00, 0.00

Table 2

Unstandardized Coefficient Alpha and Standard Errors with Standard Deviations Ranging from 1.0 to 2.0

Each cell contains the α followed by alpha's standard error, ASE.

	n	$\bar{r}$: 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
p=2	30	.000, .292	.148, .248	.276, .210	.387, .176	.485, .146	.571, .118	.649, .094	.718, .071	.780, .050	.837, .030	.889, .000
	50	.000, .226	.148, .192	.276, .163	.387, .136	.485, .113	.571, .092	.649, .072	.718, .055	.780, .039	.837, .023	.889, .000
	100	.000, .160	.148, .136	.276, .115	.387, .096	.485, .080	.571, .065	.649, .051	.718, .039	.780, .027	.837, .016	.889, .000
	200	.000, .113	.148, .096	.276, .081	.387, .068	.485, .056	.571, .046	.649, .036	.718, .028	.780, .019	.837, .012	.889, .000
p=3	30	.000, .295	.228, .228	.396, .177	.525, .138	.626, .107	.709, .081	.777, .061	.835, .043	.884, .028	.926, .014	.963, .000
	50	.000, .229	.228, .176	.396, .137	.525, .107	.626, .083	.709, .063	.777, .047	.835, .033	.884, .021	.926, .011	.963, .000
	100	.000, .162	.228, .124	.396, .097	.525, .075	.626, .058	.709, .045	.777, .033	.835, .023	.884, .015	.926, .008	.963, .000
	200	.000, .114	.228, .088	.396, .068	.525, .053	.626, .041	.709, .032	.777, .023	.835, .017	.884, .011	.926, .006	.963, .000
p=5	30	.000, .278	.330, .185	.523, .131	.648, .095	.737, .070	.803, .051	.854, .036	.894, .025	.927, .015	.955, .008	.978, .000
	50	.000, .215	.330, .143	.523, .101	.648, .074	.737, .054	.803, .040	.854, .028	.894, .019	.927, .012	.955, .006	.978, .000
	100	.000, .152	.330, .101	.523, .072	.648, .052	.737, .038	.803, .028	.854, .020	.894, .014	.927, .008	.955, .004	.978, .000
	200	.000 .108	.330, .072	.523, .051	.648, .037	.737, .027	.803, .020	.854, .014	.894, .010	.927, .006	.955, .003	.978, .000
p=7	30	.000, .274	.418, .159	.615, .104	.730, .072	.806, .051	.859, .037	.898, .026	.929, .017	.953, .010	.973, .005	.989, .000
	50	.000, .212	.418, .123	.615, .081	.730, .056	.806, .040	.859, .028	.898, .020	.929, .013	.953, .008	.973, .004	.989, .000
	100	.000, .150	.418, .087	.615, .057	.730, .040	.806, .028	.859, .020	.898, .014	.929, .009	.953, .006	.973, .003	.989, .000
	200	.000, .106	.418, .061	.615, .040	.730, .028	.806, .020	.859, .014	.898, .010	.929, .007	.953, .004	.973, .002	.989, .000
p=10	30	.000, .269	.506, .132	.696, .081	.795, .054	.856, .037	.897, .026	.927, .018	.949, .012	.967, .007	.981, .003	.993, .000
	50	.000, .208	.506, .102	.696, .063	.795, .042	.856, .029	.897, .020	.927, .014	.949, .009	.967, .006	.981, .003	.993, .000
	100	.000, .147	.506, .072	.696, .044	.795, .030	.856, .020	.897, .014	.927, .010	.949, .007	.967, .004	.981, .002	.993, .000
	200	.000, .104	.506, .051	.696, .031	.795, .021	.856, .014	.897, .010	.927, .007	.949, .005	.967, .003	.981, .001	.993, .000

Table 3

Unstandardized Coefficient Alpha and Standard Errors with Standard Deviations Ranging from 1.0 to 5.0

Each cell contains the α followed by alpha's standard error, ASE.

	n	$\bar{r}$: 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
p=2	30	.000, .140	.074, .129	.143, .118	.207, .107	.267, .096	.323, .084	.375, .072	.424, .060	.471, .047	.514, .012	.556, .000
	50	.000, .109	.074, .100	.143, .092	.207, .083	.267, .074	.323, .065	.375, .056	.424, .046	.471, .036	.514, .032	556, .000
	100	.000, .077	.074, .071	.143, .065	.207, .059	.267, .052	.323, .046	.375, .040	.424, .033	.471, .026	.514, .025	556, .000
	200	.000, .054	.074, .050	.143, .046	.207, .041	.267, .037	.323, .033	.375, .028	.424, .023	.471, .018	.514, .017	556, .000
p=3	30	.000, .252	.174, .204	.312, .166	.424, .134	.517, .108	.594, .085	.661, .065	.719, .048	.769, .033	.813, .019	.852, .000
	50	.000, .195	.174, .158	.312, .129	.424, .104	.517, .083	.594, .066	.661, .051	.719, .037	.769, .026	.813, .015	.852, .000
	100	.000, .138	.174, .112	.312, .091	.424, .074	.517, .059	.594, .047	.661, .036	.719, .026	.769, .018	.813, .010	.852, .000
	200	.000, .098	.174, .079	.312, .064	.424, .052	.517, .042	.594, .033	.661, .025	.719, .019	.769, .013	.813, .007	.852, .000
p=5	30	.000, .259	.265, .185	.437, .137	.558, .103	.648, .077	.717, .058	.772, .042	.816, .029	.853, .019	.884, .010	.911, .000
	50	.000, .200	.265, .143	.437, .106	.558, .080	.648, .060	.717, .045	.772, .033	.816, .023	.853, .014	.884, .007	.911, .000
	100	.000, .142	.265, .101	.437, .075	.558, .056	.648, .042	.717, .032	.772, .023	.816, .016	.853, .010	.884, .005	.911, .000
	200	.000, .100	.265, .072	.437, .053	.558, .040	.648, .030	.717, .022	.772, .016	.816, .011	.853, .007	.884, .004	.911, .000
p=7	30	.000, .263	.367, .164	.558, .111	.675, .079	.755, .057	.812, .041	.856, .029	.889, .020	.917, .012	.939, .006	.958, .000
	50	.000, .203	.367, .l27	.558, .086	.675, .061	.755, .044	.812, .032	.856, .023	.889, .015	.917, .010	.939, .005	.958, .000
	100	.000, .144	.367, .090	.558, .061	.675, .043	.755, .031	.812, .023	.856, .016	.889, .011	.917, .007	.939, .003	.958, .000
	200	.000, .102	.367, .063	.558, .043	.675, .031	.755, .022	.812, .016	.856, .011	.889, .008	.917, .005	.939, .002	.958, .000
p=10	30	.000, .261	.453, .140	.644, .089	.749, .061	.815, .043	.861, .030	.895, .021	.920, .014	.941, .009	.957, .004	.970, .000
	50	.000, .203	.453, .109	.644, .069	.749, .047	.815, .033	.861, .023	.895, .016	.920, .011	.941, .007	.957, .003	.970, .000
	100	.000, .143	.453, .077	.644, .049	.749, .033	.815, .023	.861, .017	.895, .012	.920, .008	.941, .005	.957, .002	.970, .000
	200	.000, .101	.453, .054	.644, .034	.749, .023	.815, .016	.861, .012	.895, .008	.920, .005	.941, .003	.957, .002	.970, .000

Table 4

Study 2: Investigation into Covariance Heterogeneity

Dependent Variable: Alpha

Factors	df	MS	F	p-value	$\hat{\beta}$	se	η^2
n	6	1.722	1920.23	0.010	0.117	0.001	0.001
p	6	239.382	266926.00	<.0001	1.027	0.000	0.187
$\bar{r}$	4	1283.516	1431204.00	<.0001	1.413	0.000	0.669
cov hetero	2	0.003	3.80	<.0001	0.023	0.003	0.000
n x p	36	0.156	17.31	<.0001	-0.057	0.001	0.000
n x $\bar{r}$	24	0.139	154.62	<.0001	-0.080	0.000	0.000
p x $\bar{r}$	24	8.919	9945.12	<.0001	-0.864	0.000	0.028
n x cov hetero	12	0.001	0.74	0.768	-0.003	0.001	0.000
$\bar{r}$ x cov hetero	8	0.004	4.77	<.0001	-0.032	0.000	0.000
p x cov hetero	12	0.003	3.44	<.0001	-0.026	0.000	0.000
n x p x $\bar{r}$	96	0.002	1.80	<.0001	0.037	0.000	0.000
n x p x cov hetero	72	0.001	0.73	0.983	0.006	0.000	0.000
n x $\bar{r}$ x cov hetero	48	0.004	4.73	0.764	0.002	0.000	0.000
p x r x cov hetero	48	0.004	4.73	<.0001	0.033	0.000	0.000
n x p x r x cov hetero	288	0.001	1.12	0.042	-0.003	0.000	0.000
Error	734,313						
Total	734,999						

Dependent Variable: Alpha's Standard Error, ASE

Factors	df	MS	F	p-value	$\hat{\beta}$	se	η^2
n	6	31.405	584834.00	<.0001	-2.427	0.000	0.441
p	6	5.826	108502.00	<.0001	-0.926	0.000	0.082
$\bar{r}$	4	21.886	407579.00	<.0001	-1.208	0.001	0.000
cov hetero	2	0.087	1617.33	<.0001	0.073	0.000	0.037
n x p	36	0.443	8248.56	<.0001	1.217	0.000	0.095
n x $\bar{r}$	24	1.691	31490.8	<.0001	1.601	0.000	0.011
p x $\bar{r}$	24	0.194	3611.59	<.0001	0.793	0.000	0.000
n x cov hetero	12	0.006	120.65	<.0001	0.070	0.000	0.000
$\bar{r}$ x cov hetero	8	0.006	113.81	<.0001	0.152	0.000	0.000
p x cov hetero	12	0.001	21.14	<.0001	0.056	0.000	0.000
n x p x $\bar{r}$	96	0.15	283.67	<.0001	-0.772	0.000	0.005
n x p x cov hetero	72	0.000	1.90	<.0001	-0.046	0.000	0.000
n x $\bar{r}$ x cov hetero	48	0.000	8.94	<.0001	-0.119	0.000	0.000
p x r x cov hetero	48	0.000	3.89	<.0001	-0.102	0.000	0.000
n x p x r x cov hetero	288	0.000	1.34	<.0001	0.073	0.000	0.000
Error	734,313						
Total	734,999						

Table 5: A Framework of Confidence Interval Estimates for Reliability Coefficients

Source	Reliability Index	95% Confidence Interval	Comments
This paper (based on σ^2 of van Zyl, Heinz & Nel 2000 *Psychometrika*), ASE	α	$\hat{\alpha} \pm (1.96)\left(\sqrt{\frac{Q}{n}}\right)$, where Q = $\left[\frac{2p^2}{(p-1)^2(j'Vj)^3}\right] \times\left[(j'Vj)(trV^2+tr^2V)-2(trV)(j'V^2j)\right]$	f(x) of n, p, $\bar{r}$
Feldt 1965 *Psychometrika*; Feldt, Woodruff & Salih 1987 *Applied Psych Meas't*; Charter 1997 *Perceptual and Motor Skills*	α	(low,high), low = 1-(1-α)F_L high = 1-(1-α)F_U, where $F_L = F_{.975,df1,df2}$ $F_U = F_{.025,df1,df2}$ and df_1 = n-1 df_2 = (n-1)(p-1)	f(x) of n, p thru df indirect fn. $\bar{r}$ thru α
Hakstian & Whalen 1976 *Psychometrika*; Barchard & Hakstian 1997a *Educational and Psychological Measurement*	α	(low,high), low = $1-c^3\left[(1-\alpha)^{1/3}+1.96(\hat{\sigma})\right]^3$ high = $1-c^3\left[(1-\alpha)^{1/3}-1.96(\hat{\sigma})\right]^3$ where $\hat{\sigma}=\sqrt{\frac{18p(n-1)(1-\alpha)^{2/3}}{(p-1)(9n-11)^2}}$ and $c=\frac{(9n-11)(p-1)}{9(n-1)(p-1)-2}$	f(x) of n, p indirect fn. $\bar{r}$ thru α
Nunnally & Bernstein 1994 *Psychometric Theory*; Cortina 1993 *JAP*	α	$\hat{\alpha} \pm (1.96)(se)$ where se $=\frac{SD_r}{\sqrt{.5p(p-1)-1}}$, SD_r is the standard deviation of the item inter-correlations	indirect fn. $\bar{r}$ thru α no n
Lord & Novick 1968 *Statistical Theories of Mental Test Scores*; Mendoza, Stafford & Stauffer 2000 *Psychological Methods*	Lord & Novick Split-half reliability; test-retest modification (MS&S)	(lo,high), low = $\frac{1-k_{low}}{1+k_{low}}$, where $k_{low}=\frac{(n-1)(1-r_{xx})F_{.975,n-1,n}}{n(1+r_{xx})}$ high = $\frac{1-k_{high}}{1+k_{high}}$, where $k_{high}=\frac{(n-1)(1-r_{xx})F_{.025,n-1,n}}{n(1+r_{xx})}$	no p
Split Half Charter 2000 *Psychological Reports* Treats r_{XY} as correlation; also applicable to test-retest correlations as indices of reliability	Spearman-Brown split-half reliability coef $r_{SH}=\frac{2r_{xy}}{(1+r_{xy})}$	$z'(r_{xy}) \pm 1.96\left(\frac{1}{\sqrt{n-3}}\right) \rightarrow (z'_{low}, z'_{high})$ where z' = Fisher transform $(r_{xy\text{-}low}, r_{xy\text{-}high}) = \frac{\left(10^{z'low/1.1513}-1\right)}{\left(10^{z'low/1.1513}+1\right)}, \frac{\left(10^{z'high/1.1513}-1\right)}{\left(10^{z'high/1.1513}+1\right)}$ $(r_{SH\text{-}low}, r_{SH\text{-}high}) = \frac{2r_{xy-low}}{1+r_{xy-low}}, \frac{2r_{xy-high}}{1+r_{xy-high}}$	no p

Table 6

Comparing Sources of Bias for Alpha and its Standard Error

	Alpha	**Standard Error**
Sample size	No effect	Decrease
Scale length	Increase	Decrease
mean item correlation	Increase	Decrease
Covariance heterogeneity	No effect	Increase
Heteroscedasticity	Decrease	Slight Increase

Summarizing Accuracy and Precision of Competing Alpha Standard Errors

	Accuracy **(% C.I. containing α)**	**Precision** **(C.I. widths)**
ASE (this paper)	Approx. 95%	Among 3 narrowest
Feldt	<95%	Among 3 narrowest
Hakstian & Wahlen	<95%	Among 3 narrowest
Split-half	>95%	Among 2 widest
Lord and Novick	<95%	Among 2 widest
Nunnally	Ill-behaved	Ill-behaved

Figure 1

Study 2: Standard Error as a Function of Covariance Heterogeneity (e.g., p = 4, n = 100)

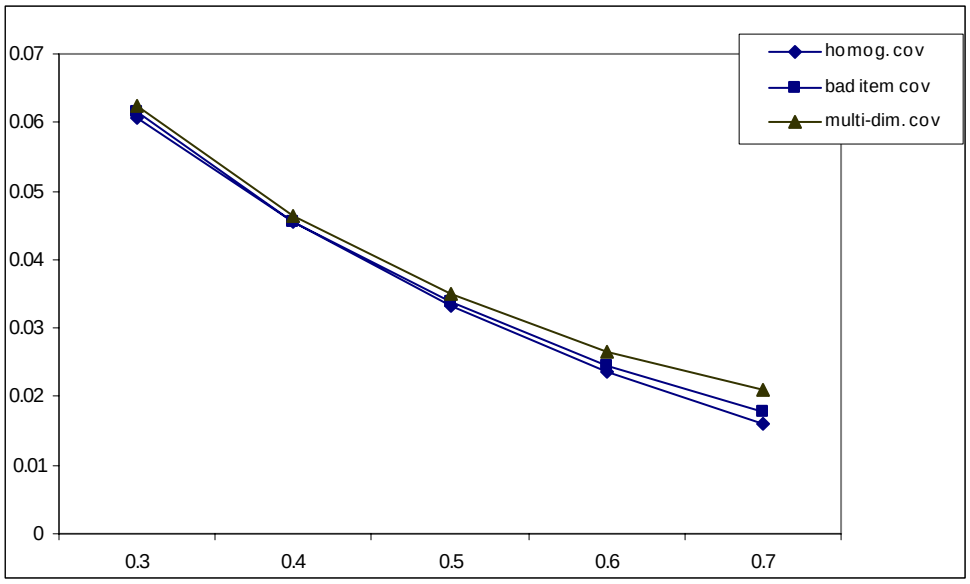
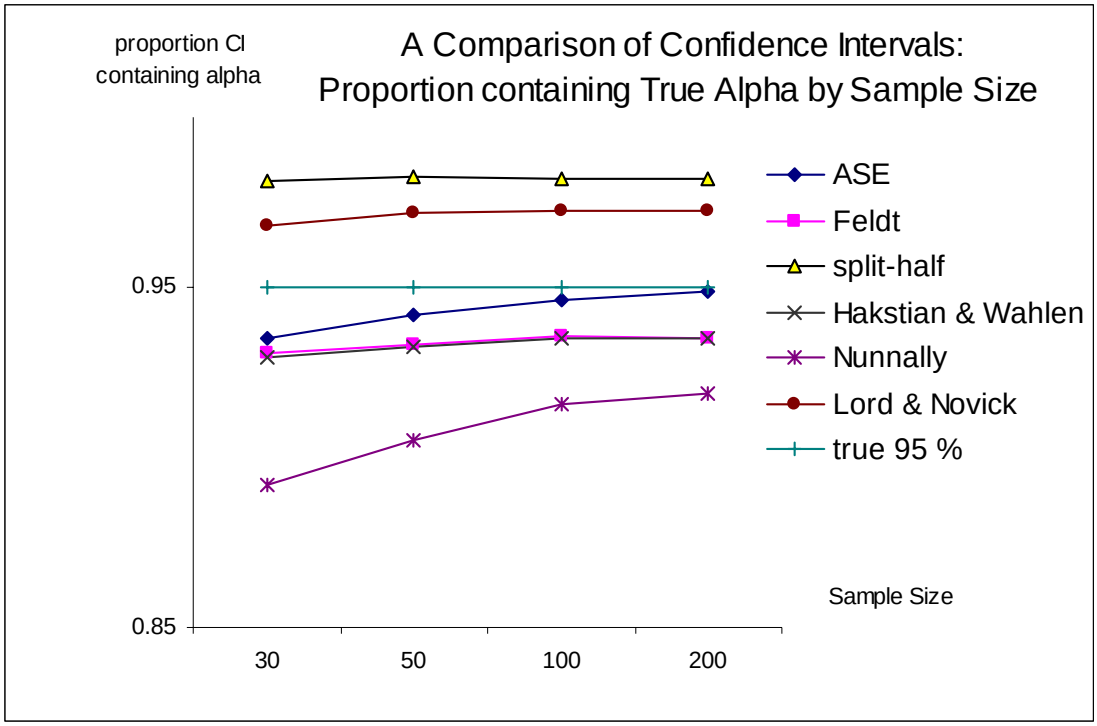


Figure 2

Study 3: Percent Confidence Intervals Containing Alpha and Interval Widths



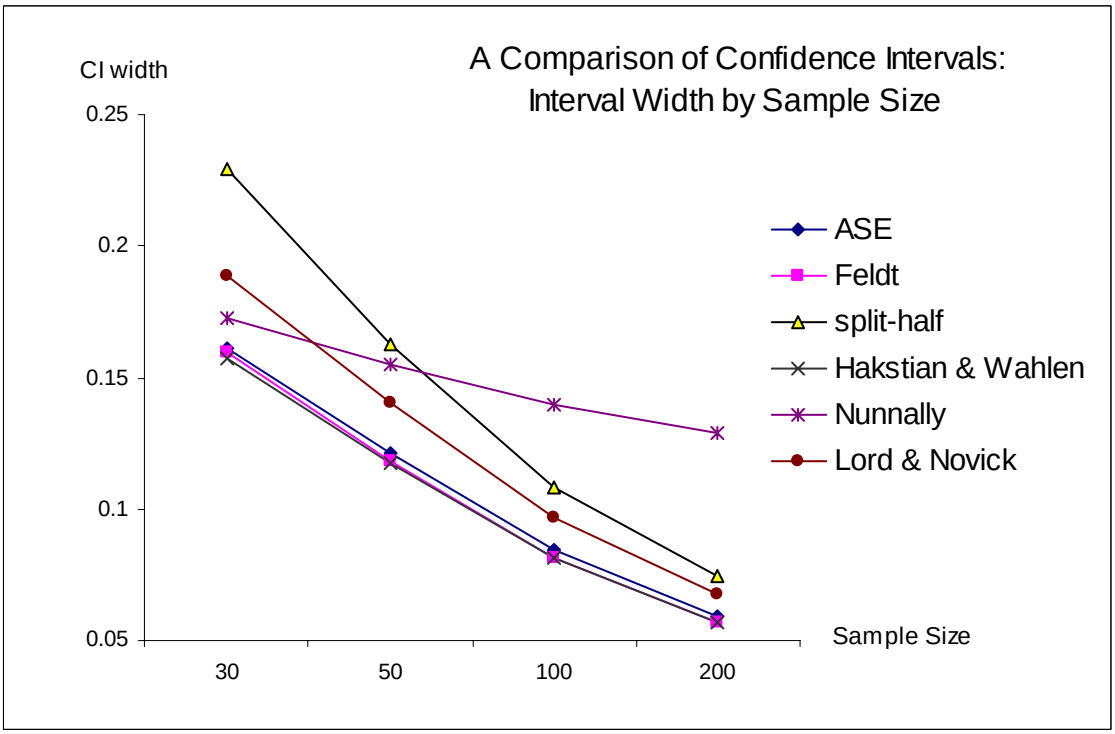


Figure 3

Best Practices for Reporting Internal Consistency Reliability

To Assess Coefficient Alpha Reliability within a Single Sample:

1. Compute point estimates for all scales per Equation 1.
2. Compute alpha's standard error ASE and confidence intervals for these point estimates using the programs in Appendix A or B.
3. To maximally reduce the size of the estimated standard error, first focus on increasing the inter-correlations among the scale items.
4. Report coefficient alpha point estimates, standard error estimates and 95% confidence intervals for all scales.

To Assess Differences in Coefficient Alpha Reliability across Independent Samples:

1. Compute point estimates for scales in each sample per Equation 1.
2. Computer standard errors for these point estimates using the programs in Appendix A or B.
3. Use Equation 4 or 5 to assess differences in reliability across sample.