Psychological Methods
2007, Vol. 12, No. 2, 157–176

1082-989X/07/$12.00 DOI: 10.1037/1082-989X.12.2.157

Asymptotically Distribution-Free (ADF) Interval Estimation of Coefficient Alpha

Alberto Maydeu-Olivares
University of Barcelona and Instituto de Empresa Business School

Donna L. Coffman
Pennsylvania State University

Wolfgang M. Hartmann
SAS Institute

The point estimate of sample coefficient alpha may provide a misleading impression of the reliability of the test score. Because sample coefficient alpha is consistently biased downward, it is more likely to yield a misleading impression of poor reliability. The magnitude of the bias is greatest precisely when the variability of sample alpha is greatest (small population reliability and small sample size). Taking into account the variability of sample alpha with an interval estimator may lead to retaining reliable tests that would be otherwise rejected. Here, the authors performed simulation studies to investigate the behavior of asymptotically distribution-free (ADF) versus normal-theory interval estimators of coefficient alpha under varied conditions. Normal-theory intervals were found to be less accurate when item skewness > 1 or excess kurtosis > 1. For sample sizes over 100 observations, ADF intervals are preferable, regardless of item skewness and kurtosis. A formula for computing ADF confidence intervals for coefficient alpha for tests of any size is provided, along with its implementation as an SAS macro.

Keywords: Likert-type items, nonnormality, categorical ordered items, model-based measurement, sampling variability

Supplementary materials: http://dx.doi.org/10.1037/1082-989X.12.2.157.supp

Arguably the most commonly used procedure to assess the reliability of a questionnaire or test score is by means of coefficient alpha (Hogan, Benjamin, & Brezinski, 2000). As McDonald (1999) pointed out, this coefficient was first proposed by Guttman (1945), with important contributions by Cronbach (1951). Coefficient alpha is a population parameter and, thus, an unknown quantity. In applications, it is typically estimated with the sample coefficient alpha, a point estimator of the population coefficient alpha. As with any point estimator, sample coefficient alpha is subject to variability around the true parameter, particularly in small samples. Thus, a better appraisal of the reliability of test scores is obtained by using an interval estimator for coefficient alpha. Duhachek and Iacobucci (2004; see also Duhachek, Coughlan, & Iacobucci, 2005, and Iacobucci & Duhachek, 2003) made a compelling argument for the use of an interval estimator, instead of a point estimator, for coefficient alpha.

Methods for obtaining interval estimators for coefficient alpha have a long history (see Duhachek & Iacobucci, 2004, for an overview). The initial proposals for obtaining confidence intervals for coefficient alpha were based on model as well as distributional assumptions. Thus, if a particular model held for the covariance matrix among the test items, *and* the test items followed a particular distribution, then a confidence interval for coefficient alpha could be obtained. The sampling distribution for coefficient alpha was first derived (independently) by Kristof (1963) and Feldt (1965),

Alberto Maydeu-Olivares, Department of Psychology, University of Barcelona, Barcelona, Spain, and Department of Marketing, Instituto de Empresa Business School, Madrid, Spain; Donna L. Coffman, Methodology and Prevention Research Centers, Pennsylvania State University; Wolfgang M. Hartmann, SAS Institute, Cary, North Carolina.

This research was supported by the Department of Universities, Research, and Information Society of the Catalan Government and by Spanish Ministry of Science and Technology Grants BSO2000-0661 and BSO2003-08507 to Alberto Maydeu-Olivares. Donna L. Coffman was supported by National Institute on Drug Abuse Training Grant T32 DA017629-01A1 and National Institute on Drug Abuse Center Grant P50 DA10075.

Correspondence concerning this article should be addressed to Alberto Maydeu-Olivares, Faculty of Psychology, University of Barcelona, Paseo Valle de Hebrón, 171, 08035, Barcelona, Spain. E-mail: amaydeu@ub.edu

who assumed that the test items are strictly parallel (Lord & Novick, 1968) and normally distributed. This model implies that all the item variances are equal and all item covariances are equal (i.e., a compound symmetric covariance structure). However, Barchard and Hakstian (1997) found that confidence intervals for coefficient alpha obtained from these results were not sufficiently accurate when model assumptions were violated (i.e., the items were not strictly parallel). As Duhachek and Iacobucci (2004) suggested, the lack of robustness of the interval estimators for coefficient alpha to violations of model assumptions have hindered the widespread use of these intervals in applications.

A major breakthrough in interval estimation occurred when van Zyl, Neudecker, and Nel (2000) derived the asymptotic (i.e., large sample) distribution of sample coefficient alpha without model assumptions.[1] The normal-theory (NT) interval estimator proposed by van Zyl et al. does not require the assumption of compound symmetry. In particular, these authors assumed only that the items comprising the test were normally distributed. Duhachek and Iacobucci (2004) recently investigated the performance of the confidence intervals for coefficient alpha using the results of van Zyl et al. versus procedures proposed by Feldt (1965) and those proposed by Hakstian and Whalen (1976) under violations of the parallel-measurement model. They found that the model-free, NT interval estimator proposed by van Zyl et al. uniformly outperformed competing procedures across all conditions.

However, the results of van Zyl et al. (2000) assume that the items composing the test can be well approximated by a normal distribution. In practice, tests are most often composed of binary or Likert-type items for which the normal distribution can be a poor approximation. Yuan and Bentler (2002) showed that the NT-based confidence intervals for coefficient alpha are asymptotically robust to violations of the normality assumptions under some conditions. Unfortunately, these conditions cannot be verified in applications. So whenever the observed data are markedly nonnormal, the researcher cannot verify if the necessary conditions put forth by Yuan and Bentler are satisfied or not.

Recently, using the scales of the Hopkins Symptom Checklist (HSCL; Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974), Yuan, Guarnaccia, and Hayslip (2003) compared the performance of the NT confidence intervals of van Zyl et al. (2000) with a newly proposed model-free, asymptotically distribution-free (ADF) confidence interval and with several confidence intervals based on bootstrapping. Yuan et al. concluded that the ADF intervals were more accurate for the Likert-type items of the HSCL than were the NT intervals but less accurate than were the bootstrapping procedures.

However, as Yuan et al. (2003, p. 7) pointed out, their conclusions may not be generalized to other Likert-type scales, because the item-distribution shapes, such as skewness and kurtosis, of the HSCL subscales may not be shared by other psychological inventories composed of Likert-type scales. The purpose of the current study is to investigate by means of a simulation study the behavior of the ADF interval estimator for coefficient alpha, introduced by Yuan et al., versus the NT interval estimator, proposed by van Zyl et al. (2000), with Likert-type data.[2] In so doing, we consider conditions in which the Likert-type items show skewness and kurtosis similar to those of normal variables but also conditions of high skewness, typically found in responses to questionnaires measuring rare events, such as employee drug usage, psychopathological behavior, and adolescent deviant behaviors such as shoplifting (see also Micceri, 1989). Computing the ADF confidence interval for coefficient alpha can be difficult when the number of variables is large. Our work provides some simplifications to the formulas that enable the computation of these confidence intervals for tests of any size. Yuan et al. did not provide these simplifications, and practical use of their equations would be limited in the number of variables. Further, we provide an SAS macro with the simplifications to compute the NT and ADF confidence intervals for coefficient alpha.

Coefficient Alpha and the Reliability of a Test Score

Consider a test composed of p items, $Y_1, \ldots, Y_p$, intended to measure a single attribute. One of the most common tasks in psychological research is to determine the reliability of the test score $X = Y_1 + \ldots + Y_p$, that is, the percentage of variance of the test score that is due to the attribute of which the items are indicators.

The most widely used procedure to assess the reliability of a questionnaire or test score is by means of coefficient alpha (Cronbach, 1951; Guttman, 1945). In the population of respondents, coefficient alpha is

$$\alpha = \frac{p}{p-1}\left(1 - \frac{\sum_i \sigma_{ii}}{\sum_{ij} \sigma_{ij}}\right), \quad (1)$$

where $\sum_i \sigma_{ii}$ denotes the sum of the p item variances in the

[1] Only a positive definite covariance matrix is assumed. All previous derivations, which assumed particular models (e.g., tau equivalence) for the covariance matrix, can be treated as special cases of their result.

[2] Bootstrap confidence intervals are not considered in this study. On the one hand, there are a variety of procedures that should be investigated (for an overview, see Hartmann, 2005). On the other hand, they are computationally more intensive. Most important, differences between ADF and bootstrap confidence intervals in Yuan et al.'s (2003) study are in all cases in the third decimal, a negligible difference for practical purposes.

population, and $\sum_{ij}^{\sigma_{ij}}\sigma_{ij}$ denotes the sum of the $\frac{p(p-1)}{2}$ unique item covariances. In applications, a sample of N respondents from the population is available, and a point estimator of the population alpha given in Equation 1 can be obtained using the sample coefficient alpha

$$\hat{\alpha} = \frac{p}{p-1}\left(1 - \frac{\sum_i s_{ii}}{\sum_{ij} s_{ij}}\right), \quad (2)$$

where s_{ij} denotes the sample covariance between items i and j, and s_{ii} denotes the sample variance of item i.

A necessary and sufficient condition for coefficient alpha to equal the reliability of the test score is that the items are *true-score equivalent* (i.e., essentially tau-equivalent items) in the population (Lord & Novick, 1968, p. 50; McDonald, 1999, Chapter 6). A true-score equivalent model is a one-factor model in which the factor loadings are equal for all items. The model implies that the population covariances are all equal, but the population variances need not be equal for all items.

A special case of the true-score equivalent model is the *parallel-items* model, in which, in addition to the assumptions of the true-score equivalent model, the unique variances of the error terms in the factor model are assumed to be equal for all items. The parallel-items model results in a population-covariance matrix with only two distinct parameters, a covariance common to all pairs of items and a variance common to all items. This covariance structure is commonly referred to as compound symmetry. In turn, a special case of the parallel-items model is the *strictly parallel-items* model. In this model, in addition to the assumptions of parallel items, the item means are assumed to be equal across items. When items are parallel or strictly parallel, coefficient alpha also equals the reliability of the test score.

When the items do not conform to a true score-equivalent model, coefficient alpha does not equal the reliability of the test score. For instance, if the items conform to a one-factor model with distinct factor loadings (i.e., *congeneric* items), then the reliability of the test score is given by coefficient omega.[3] Under a congeneric measurement model, coefficient alpha underestimates the true reliability. However, the difference between coefficient alpha and coefficient omega is small (McDonald, 1999), unless one of the factor loadings is very large (e.g., .9) and all the other factor loadings are very small (e.g., .2; Raykov, 1997). This condition is rarely encountered in practical applications.

NT and ADF Interval Estimators for Coefficient Alpha

In this section, we summarize the main results regarding the large sample distribution of sample coefficient alpha. Technical details can be found in the Appendix.

In large samples, $\hat{\alpha}$ is normally distributed with mean α and variance φ^2 (see the Appendix). As a result, in large samples, an x% confidence interval for the population coefficient alpha can be obtained as $(L_L; U_L)$. The lower limit of the interval, L_L, is $\hat{\alpha} - z_{x/2}\hat{\varphi}$, and the upper limit, U_L, is $\hat{\alpha} + z_{x/2}\hat{\varphi}$, where $\hat{\varphi}$ is the square root of the estimated large sample variance of sample alpha (i.e., its asymptotic standard error), and $z_{x/2}$ is the $(1 - x/2)$% quantile of a standard normal distribution. Thus, for instance, $z_{x/2} = 1.96$ for a 95% confidence interval for alpha.

No distributional assumptions have been made so far. The above results hold under NT assumptions (i.e., when the data are assumed to be normal) but also under the ADF assumptions set forth by Browne (1982, 1984).[4] Under normality assumptions, φ^2 depends only on population variances and covariances (bivariate moments), and as a result, it can be estimated from the sample variances and covariances (see the Appendix).

In contrast, under ADF assumptions, φ^2 depends on fourth-order moments (see Browne, 1982, 1984, for further details). As a result, the estimation of φ^2 requires computing an estimate of the asymptotic covariance matrix of the sample variances and covariances. This matrix is of dimensions $q \times q$, where $q = \frac{p(p+1)}{2}$, the number of unique variances and covariances. One consideration when choosing between the ADF and NT intervals is that the former are, in principle, computationally more intensive, because a $q \times q$ matrix must be stored, and the size of this matrix increases very rapidly as the number of items increases. However, we show in the Appendix that an estimate of the asymptotic variance of coefficient alpha under ADF assumptions can be obtained without storing this large matrix. This formula has been implemented in an SAS macro, which is available as supplementary material online. Detailed instructions on using the macro, along with a sample dataset for testing the macro, are provided. The macro is easy for applied researchers to use, and it also provides the NT confidence interval. It can be used to compute ADF confidence intervals for tests of any size, and in our implementation, the computation is only slightly more involved than that for the NT confidence intervals.

[3] The formula for coefficient omega can be found in the Appendix.

[4] ADF estimation replaces the normality assumption by the milder assumption that eighth-order moments of the distribution of the data are finite. This assumption is satisfied in the case of Likert-type items, where the distribution of each item is multinomial. The assumption ensures that the fourth-order sample moments are consistent estimators of their population counterparts (Browne, 1984).

Some Considerations in the Use of NT Versus ADF Interval Estimators

Both the NT and ADF interval estimators are based on large-sample theory. Hence, large samples will be needed for either of the confidence intervals to be accurate. Because larger samples are needed to accurately estimate the fourth-order sample moments involved in the ADF confidence intervals than the bivariate sample moments involved in the NT confidence intervals, in principle, larger samples will be needed to accurately estimate the ADF confidence intervals compared with the NT confidence intervals. On the other hand, because ADF confidence intervals are robust to nonnormality in large samples, we expect that when the test items present high skewness, high positive kurtosis, or both, the ADF confidence intervals will be more accurate than the NT confidence intervals. In other words, we expect that when the items are markedly nonnormal and large samples are available, the ADF confidence intervals will be more accurate than the NT confidence intervals. In contrast, we expect that when the data approach normality and sample size is small, the NT confidence intervals will be more accurate than the ADF confidence intervals. However, it is presently unknown under what conditions of sample size and nonnormality the ADF confidence intervals are more accurate than NT confidence intervals. This is investigated in the following sections by means of simulation.

Two simulation studies were performed. In the first simulation, data were simulated so that population alpha equaled the reliability of the test score. In the second simulation, data were simulated so that population alpha underestimated the reliability of the test score. This occurs, for instance, when the model underlying the items is a one-factor model with unequal factor loadings (i.e., a congeneric measurement model).

Previous researcher (e.g., Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992) has found that the ADF estimator performs poorly in confirmatory factor-analysis models with small sample sizes. In fact, they have recommended sample sizes over 1,000 for ADF estimation. However, our use of ADF theory differs from theirs in two key aspects. First, there is only one parameter to be estimated in this case, coefficient alpha. As in Yuan et al. (2003), we estimate this parameter using sample coefficient alpha. Thus, we use ADF theory only in the estimation of the standard error and not in the point estimation of coefficient alpha. Hu et al. (1992) and Curran et al. (1996) used ADF theory to estimate both the parameters and standard errors. Second, there is only one standard error to be computed here, the standard error of coefficient alpha. These key differences between the present usage of ADF theory and previous research on the behavior of ADF theory in confirmatory factor analysis led us to believe that much smaller sample sizes would be needed than in previous studies.

A Monte Carlo Investigation of NT Versus ADF Confidence Intervals When Population Alpha Equals the Reliability of the Test

Most often, tests and questionnaires are composed of Likert-type items, and coefficient alpha is estimated from ordered categorical data. To increase the validity and generalizability of the study, we used ordinal data in the simulation study. The procedure used to generate the data was similar to that of Muthén and Kaplan (1985, 1992). It enabled us to generate ordered categorical data with known population item skewness and kurtosis.

More specifically, the following sequence was used in the simulation studies:

1. Choose a correlation matrix $\mathbf{P}$ and a set of thresholds $\boldsymbol{\tau}$.
2. Generate multivariate normal data with mean zero and correlation matrix $\mathbf{P}$.
3. Categorize the data using the set of thresholds $\boldsymbol{\tau}$.
4. Compute the sample-covariance matrix among the items, $\mathbf{S}$, after categorization. Then, compute sample coefficient alpha using Equation 2 and its NT and ADF standard errors using Equations 5 and 7 (see the Appendix). Also, compute NT and ADF confidence intervals as described in the previous section.
5. Compute the true population-covariance matrix among the items, Σ, after categorization. Technical details on how to compute this matrix are given in the Appendix.
6. Compute the population coefficient alpha via Equation 1, using $\boldsymbol{\Sigma}$, the covariance matrix in the previous stage.
7. Determine if confidence intervals cover the true alpha, underestimate it, or overestimate it.

In the first simulation study, all elements of $\mathbf{P}$ were equal, as implied by a parallel-items model. Also, the same thresholds were used for all items.[5] These choices resulted in a compound symmetric-population covariance matrix $\boldsymbol{\Sigma}$ (i.e., equal covariances and equal variances) for the ordered categorical items (see the Appendix). In other words, $\boldsymbol{\Sigma}$ was consistent with a parallel-items model.

Overall, we investigated 144 conditions. These were obtained using a factorial design by crossing the following:

[5] The use of a common set of thresholds for all items simplifies the presentation of the findings, because all items have a common skewness and kurtosis. Additional simulations were performed with unequal thresholds, yielding results very similar to those reported in the article.

1. Four sample sizes (50, 100, 200, and 400 respondents);
2. Two test lengths (5 and 20 items);
3. Three different values for the common correlation in **P** (.16, .36, and .64; this is equivalent to assuming a one-factor model for these correlations with common factor loadings of .4, .6, and .8, respectively); and
4. Six item types (three types consisted of items with two categories, and three types consisted of items with five categories) that varied in skewness and/or kurtosis.

The sample sizes were chosen to be small to very large in typical questionnaire-development applications. Also, in our experience, 5 and 20 items are the typical shortest and longest lengths for questionnaires measuring a single attribute. Finally, we included items with typical small (.4) to large (.8) factor loadings.

The item types used in the study, along with their population skewness and kurtosis, are depicted in Figure 1. Details on how to compute the population item skewness and kurtosis are given in the Appendix. These item types were chosen to be typical of a variety of applications. We report results only for positive skewness, because the effect was symmetric for positive and negative skewness. Items of Types 1–3 consisted of only two categories. Type 1 items had the highest skewness and kurtosis. The threshold was chosen such that only 10% of the respondents endorsed the items. Type 2 items were endorsed by 15% of the respondents, resulting in smaller values of skewness and kurtosis. Items of Types 1 and 2 are typical of applications in which items are seldom endorsed. On the other hand, Type 3 items were endorsed by 40% of the respondents. These items had low skewness, and their kurtosis was smaller than that of a standard normal distribution.[6] Items of Types 4–6 consisted of five categories. The skewness and kurtosis of Type 5 items closely matched those of a standard normal distribution. Type 4 items were also symmetric (skewness = 0); however, the kurtosis was higher than that of a standard normal distribution. These items can be found in applications in which the middle category reflects an undecided position, and a large number of respondents choose this middle category. Finally, Type 6 items showed a substantial amount of skewness and kurtosis. For these items, the thresholds were chosen so that the probability of endorsing each category decreased as the category label increased.

For each of the 144 conditions, we obtained 1,000 replications. For each replication, we computed the sample coefficient alpha, the NT and ADF standard errors, and the NT and ADF 95% confidence intervals. Then, for each condition, we computed (a) the relative bias of the point estimate of coefficient alpha as $\text{bias}(\hat{\alpha}) = \frac{\text{mean}_{\hat{\alpha}} - \alpha}{\alpha}$, (b) the relative bias of the NT and ADF standard errors as $\text{bias}(\hat{\varphi}) = \frac{\text{mean}_{\hat{\varphi}} - \text{std}_{\hat{\alpha}}}{\text{std}_{\hat{\alpha}}}$, and (c) the coverage of the NT and ADF 95% confidence intervals (i.e., the proportion of estimated confidence intervals that contain the true population alpha).

The accuracy of ADF versus NT confidence intervals was assessed by their coverage. Coverage should be as close to the nominal level (.95 in our study) as possible. Larger coverage than the nominal level indicates that the estimated confidence intervals are too wide; they overestimate the variability of sample coefficient alpha. Smaller coverage than the nominal level indicates that the estimated confidence intervals are too narrow; they underestimate the variability of sample coefficient alpha.

Note that there are two different population correlations within our framework: (a) the population correlations before categorizing the data (i.e., the elements of **P**) and (b) the population correlations after categorizing the data (i.e., the correlations that can be obtained by dividing each covariance in $\boldsymbol{\Sigma}$ by the square root of the product of the corresponding diagonal elements of $\boldsymbol{\Sigma}$). We refer to the former as *underlying correlations* and to the latter as *interitem population correlations*. Also note that in our factorial design, probabilities of responding to item alternatives are manipulated to yield different item types. Thus, skewness and excess kurtosis are not independently manipulated. Rather, the effects of skewness and kurtosis are confounded. Either one can be used to illustrate the effect of departure from normality on coverage rates for ADF and NT intervals. In this simulation, coverage results are shown as a function of skewness. In the second simulation, they are shown as a function of excess kurtosis.

Table 1 summarizes the relationship between the average interitem correlations in the population after categorization of the data and the underlying correlation before categorization. The average interitem correlation is the extent of interrelatedness (i.e., internal consistency) among the items (Cortina, 1993). There are three levels for the average population interitem correlation, corresponding to the three underlying correlations. Table 1 also summarizes the population alpha corresponding to the three levels of the average population interitem correlations. As may be seen in this table, the population coefficient alpha used in our study

[6] The skewness and kurtosis of a standard normal distribution are 0 and 3, respectively. We subtracted 3 from the kurtosis values so that 0 indicated no excess kurtosis, a positive value indicated excess kurtosis greater than that of a normal distribution, and a negative kurtosis value indicated excess kurtosis less than that of a normal distribution.

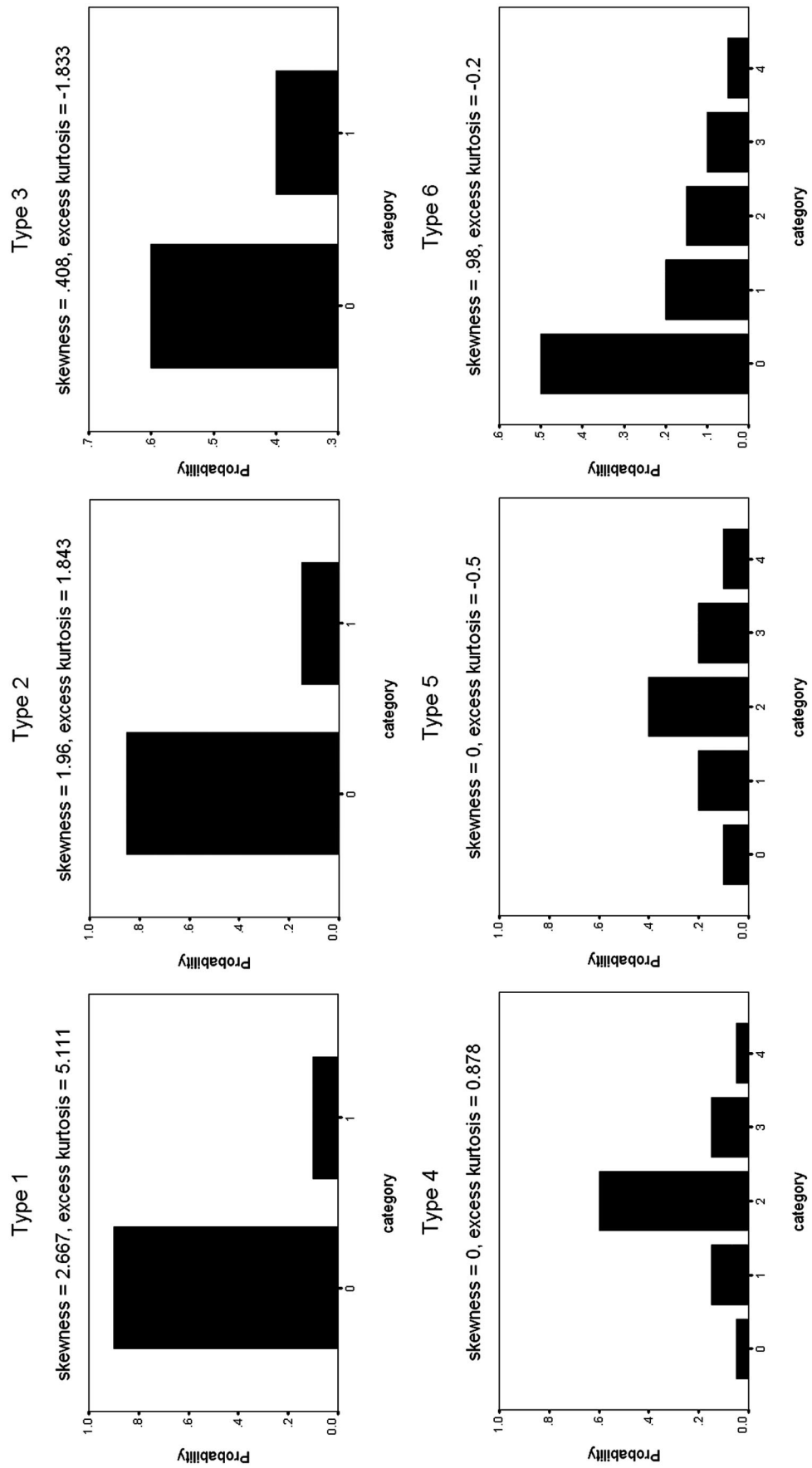

Figure 1. Histograms of the different types of items employed in the simulation study.

Table 1
Relationship Between Underlying Polychoric Correlation (ρ), the Average Population Interitem Correlation ($\bar{\rho}$), and the Population Coefficient Alpha (α)

		$\bar{\rho}$			Population α		
ρ	Level	*M*	Minimum	Maximum	*M*	Minimum	Maximum
.16	Low	.11	.06	.15	.53	.25	.77
.36	Medium	.25	.16	.33	.74	.49	.91
.64	High	.48	.36	.59	.88	.73	.97

ranges from .25 to .97, and the population interitem correlations range from .06 to .59. Thus, in the present study, we consider a wide range of values for both the population coefficient alpha and the population interitem correlations.

Empirical Behavior of Sample Coefficient Alpha: Bias and Sampling Variability

To our knowledge, the behavior of the point estimate of coefficient alpha when computed from ordered categorical data under conditions of high skewness and kurtosis has never been investigated. Consequently, we report results on the bias and variability of sample alpha under these conditions.

The results for the bias of the point estimates of coefficient alpha are best depicted graphically as a function of the true population alpha. The results for the 144 conditions investigated are shown in Figure 2. Three trends are readily apparent from Figure 2. First, bias increases with decreasing true population alpha. Second, bias is consistently negative. In other words, the point estimate of coefficient alpha consistently underestimates the true population alpha. Third, the variability of the bias increases with decreasing sample size. For fixed sample size and true reliability, bias increases with increased kurtosis and increased skewness. This is not shown in the figure, for ease of presentation. Nevertheless, the coefficient-alpha point estimates are remarkably robust to skewness and kurtosis: Provided sample size is larger than 100, relative bias is less than 5% whenever population alpha is larger than .3.

Results on the variability of sample alpha are reported graphically in Figure 3. This figure depicts the standard

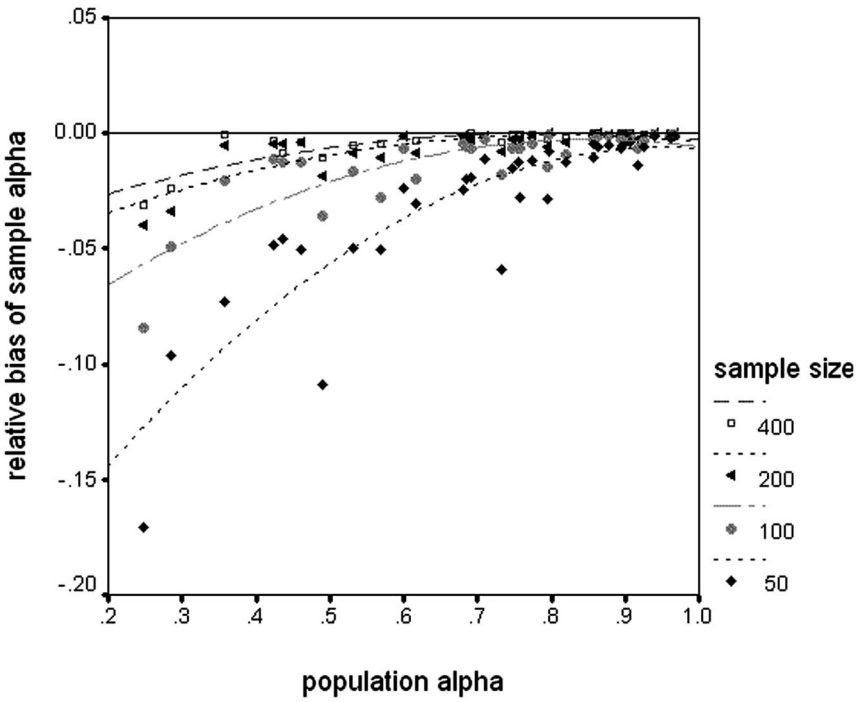

Figure 2. Relative bias of the coefficient-alpha point estimates as a function of the true population alpha. A quadratic model has been fit to the points to model the relationship between relative bias and true alpha by sample size. Bias increases with decreasing sample size and decreasing population alpha.

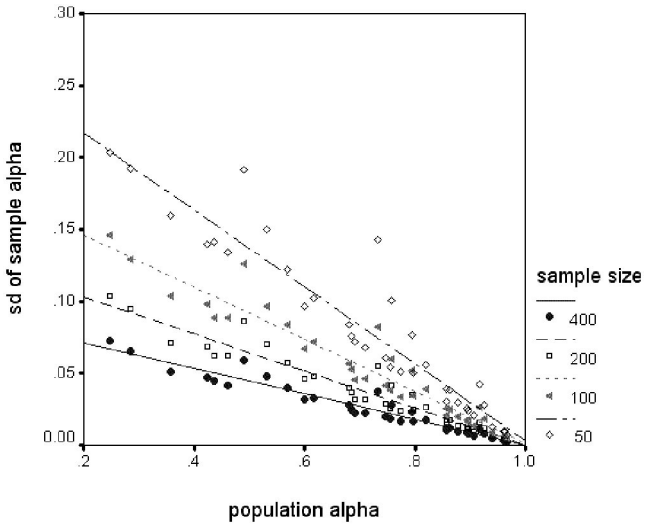

Figure 3. Variability of the coefficient-alpha point estimates as a function of the true population coefficient alpha by sample size. Linear functions have been fit to the points to model the relationship between the standard deviation of sample coefficient alpha and the true population coefficient alpha.

deviation of the point estimate of coefficient alpha as a function of the true population alpha. As can be seen in this figure, the variability of the point estimate of coefficient alpha is the result of the true population coefficient alpha and sample size. As the population coefficient alpha approaches 1.0, the variability of the point estimate of coefficient alpha approaches zero. As the population coefficient alpha becomes smaller, the variability of the point estimates of coefficient alpha increases. The increase in variability is larger when the sample size is small. An interval estimator for coefficient alpha is most needed when the variability of the point estimate of coefficient alpha is largest. In those cases, a point estimator can be quite misleading. Figure 3 clearly suggests that an interval estimator is most useful when sample size is small and the population coefficient alpha is not large.

Do NT and ADF Standard Errors Accurately Estimate the Variability of Coefficient Alpha?

The relative bias of the estimated standard errors for all conditions investigated is reported in Tables 2 and 3. Results for NT standard errors are displayed in Table 2, and results for ADF standard errors are displayed in Table 3.

As can be seen in Table 3, the ADF standard errors seldom overestimate the variability of sample coefficient alpha. When it does occur, the overestimation is small (at most 3%). More generally, the ADF standard errors *underestimate* the variability of sample coefficient alpha. The bias can be substantial (−30%), but on average it is small (−5%). The largest amount of bias appears for the smallest sample size considered. For sample sizes of 200 observations, relative bias is at most −9%.

NT standard errors (see Table 2) also occasionally overestimate the variability of sample coefficient alpha. As in the case of ADF standard errors, the overestimation of NT standard errors is small (at most 4%). More generally, the NT standard errors underestimate the variability of sample coefficient alpha. The underestimation can be very severe (up to −55%). Overall, the average bias is unacceptably large (−14%). Bias increases with increasing skewness, as well as with an increasing average interitem correlation. For the two most extreme skewness conditions, and the highest level of average interitem correlation considered (.36–.59), bias is at least −30%.

As can be seen by comparing Tables 2 and 3, of the 144 different conditions investigated, the NT standard errors were more accurate than the ADF standard errors in 45 conditions (31.3% of the time). NT standard errors were more accurate than ADF standard errors when skewness was less than .5 (nearly symmetrical items) and the average interitem correlation was low (.06–.15) or medium (.16–.33). Even in these cases, the differences were very small. The largest difference in favor of NT standard errors is 5%. In contrast, in all remaining conditions (68.7% of the time), the ADF standard errors were considerably more accurate than NT standard errors. The average difference in favor of ADF standard errors is 12%, with a maximum of 44%.

Accuracy of NT and ADF Interval Estimators

Figure 4 shows the coverage rates of NT and ADF confidence intervals as a function of skewness. The coverage rates of NT confidence intervals decrease dramatically as a function of the combination of increasing skewness and increasing average interitem correlations. The coverage rates can be as low as .68 when items are severely skewed (Type 1 items) and the average interitem correlation is high (.36–.59).

Figure 4 also shows the coverage rates of ADF confidence intervals as a function of item skewness by sample size. The ADF confidence intervals behave much better than NT confidence intervals. The effect of skewness on their coverage is mild. The effect of sample size is more important. For sample sizes of at least 200 observations, ADF coverage rates are at least .91, regardless of item skewness. For a sample size of 50, the smallest coverage rate is .82. The maximum coverage rate is .96, as was also the case for NT intervals.

Table 4 provides the average coverage for NT and ADF 95% confidence intervals at each level of sample size and skewness. This table reveals that the average coverage of ADF intervals is as good as or better than the average coverage of NT intervals whenever item skewness is larger than .5, regardless of sample size (i.e., sample size $\geq$ 50). Also, ADF intervals are uniformly more accurate than NT

Table 2
Relative Bias of Normal-Theory Standard Errors

		Skewness					
		2.667	1.960	0.980	0.408	0	0
		Excess kurtosis					
N	No. variables	5.111	1.843	−0.200	−1.833	−0.500	0.878
		Low $\bar{\rho}$					
50	5	−.15	−.15	−.05	−.07	−.07	−.08
	20	−.24	−.20	−.11	−.08	−.06	−.05
100	5	−.17	−.12	−.07	−.01	−.03	.01
	20	−.24	−.20	−.09	.00	−.02	−.07
200	5	−.18	−.15	−.06	.01	−.03	.01
	20	−.23	−.16	−.08	−.01	−.01	−.03
400	5	−.17	−.13	−.04	−.01	.04	.01
	20	−.21	−.14	−.06	−.01	−.01	−.02
		Medium $\bar{\rho}$					
50	5	−.36	−.27	−.11	−.04	−.03	−.05
	20	−.40	−.31	−.12	−.02	−.07	−.09
100	5	−.35	−.22	−.10	−.04	−.01	−.04
	20	−.40	−.31	−.12	.01	−.01	−.05
200	5	−.33	−.25	−.11	−.03	.01	−.03
	20	−.39	−.28	−.11	.00	−.01	−.04
400	5	−.31	−.22	−.08	−.01	.03	.02
	20	−.36	−.26	−.10	.00	−.01	−.04
		High $\bar{\rho}$					
50	5	−.53	−.42	−.18	−.07	−.05	−.13
	20	−.55	−.41	−.16	−.04	.02	−.13
100	5	−.46	−.35	−.13	−.06	.00	−.09
	20	−.51	−.38	−.15	−.02	.01	−.10
200	5	−.45	−.34	−.13	−.08	−.01	−.07
	20	−.46	−.34	−.14	−.05	.00	−.09
400	5	−.43	−.31	−.10	−.05	.02	−.04
	20	−.45	−.34	−.14	−.05	.00	−.09

Note. $\bar{\rho}$ = average population interitem correlation.

intervals with large samples (≥ 400; i.e., regardless of item skewness). When sample size is smaller than 400 and item skewness is smaller than .5, the behavior of both methods is indistinguishable for all practical purposes. NT confidence intervals are more accurate than ADF confidence intervals only when the items are perfectly symmetric (skewness = 0) and sample size is 50. In summary, the empirical behavior of ADF confidence intervals is better than that of the NT confidence intervals.

A Monte Carlo Investigation of NT Versus ADF Confidence Intervals When Population Coefficient Alpha Underestimates the Reliability of the Test

When the population covariances are not equal, then population coefficient alpha generally underestimates the true reliability of a test score.[7] As a result, on average, sample coefficient alpha will also underestimate the true reliability, and so should the NT and ADF confidence intervals for coefficient alpha. Here, we investigate the empirical behavior of these intervals under different conditions. Using a factorial design, we crossed

1. Four sample sizes (50, 100, 400, and 1,000),
2. Three test lengths (7, 14, and 21 items), and

[7] Coefficient alpha is a lower bound to the reliability of a test score when (a) the items can be decomposed as $X_i = T_i + E_i$, with T_i and E_i being uncorrelated, and (b) the covariance matrix of the E_is is diagonal (Bentler, in press).

Table 3
Relative Bias of Asymptotically Distribution-Free Standard Errors

		Skewness					
		2.667	1.960	0.980	0.408	0	0
		Excess kurtosis					
N	No. variables	5.111	1.843	−0.200	−1.833	−0.500	0.878
		Low $\bar{\rho}$					
50	5	−.16	−.14	−.07	−.08	−.10	−.13
	20	−.19	−.17	−.13	−.12	−.10	−.09
100	5	−.12	−.08	−.06	−.02	−.04	−.01
	20	−.13	−.12	−.08	−.03	−.04	−.09
200	5	−.07	−.08	−.04	.01	−.03	.00
	20	−.07	−.05	−.05	−.03	−.03	−.04
400	5	−.04	−.03	.00	.00	.03	.00
	20	−.02	−.01	−.02	−.02	−.02	−.03
		Medium $\bar{\rho}$					
50	5	−.26	−.17	−.08	−.04	−.06	−.09
	20	−.25	−.18	−.10	−.06	−.11	−.13
100	5	−.16	−.05	−.05	−.03	−.03	−.05
	20	−.17	−.13	−.07	−.01	−.05	−.06
200	5	−.08	−.05	−.04	−.02	−.01	−.03
	20	−.09	−.05	−.04	−.02	−.03	−.04
400	5	−.02	.00	.00	.00	.02	.03
	20	−.01	−.01	−.02	−.02	−.03	−.03
		High $\bar{\rho}$					
50	5	−.30	−.21	−.10	−.02	−.09	−.11
	20	−.30	−.17	−.09	−.02	−.02	−.11
100	5	−.12	−.06	−.02	.00	−.03	−.05
	20	−.16	−.09	−.05	−.01	−.03	−.06
200	5	−.06	−.03	−.01	−.02	−.03	−.02
	20	−.04	−.01	−.03	−.03	−.03	−.03
400	5	−.01	.02	.02	.02	.00	.02
	20	−.01	−.02	−.02	−.03	−.04	−.03

Note. $\bar{\rho}$ = average population interitem correlation.

3. The six item types used in the previous simulation (three types consisted of items with two categories, and three types consisted of items with five categories),

resulting in 72 conditions. We categorized the data using the same thresholds as in our previous simulation. Thus, items with the same probabilities, and therefore with the same values for skewness and kurtosis, were used (see Figure 1).

We used the same procedure described in the previous section except for two differences. First, in Step 1 we used a correlation matrix **P** with a one-factor structure and factor loadings of .3, .4, .5, .6, .7, .8, and .9. Thus, the data were generated assuming a congeneric measurement model. For the test length with 14 items, these loadings were assigned to Items 1–7 and then repeated for Items 8–14. For the test length with 21 items, these loadings were repeated once again for Items 15–21. Second, Steps 6 and 7 consisted of two parts, as we computed both the population coefficient alpha and population reliability (in this case, population alpha underestimates reliability). We then examined the behavior of the ADF and NT confidence intervals with respect to both population parameters.

Under the conditions of this simulation study, true reliability is obtained using coefficient omega (see McDonald, 1999). Details on how the true reliabilities for each of the experimental conditions can be computed are given in the Appendix. Coefficient omega, ω, (i.e., true reliability) ranges from .60 to .92. To obtain smaller true reliabilities, we could have used fewer items and smaller factor loadings.

Also, for each condition, we computed (a) the absolute

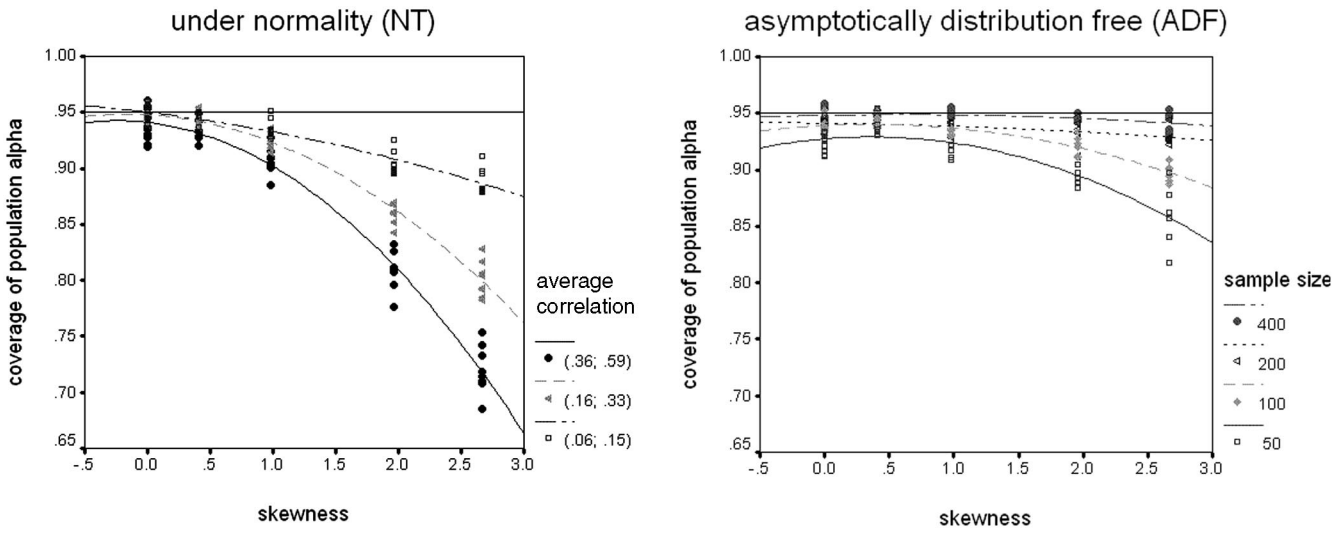

Figure 4. Proportion of times (coverage) that 95% confidence intervals (CIs) for alpha include population reliability as a function of skewness. Data have been generated so that population alpha equals reliability. Coverage rates should be close to nominal rates (95%). A quadratic model has been fit to the points to model the relationship between coverage and skewness. The accuracy of normal-theory (NT) CIs worsens as average interitem correlation gets smaller and skewness increases. The accuracy of asymptotically distribution-free (ADF) CIs worsens as sample size decreases and skewness increases. The accuracy of both CIs is similar for items with low skewness ($< |1|$); for higher skewness, ADF CIs are more accurate than NT CIs, provided sample size > 100 observations.

bias of sample coefficient alpha in estimating the true reliability as $\text{mean}_{\hat{\alpha}} - \omega$, (b) the relative bias of sample coefficient alpha in estimating the true reliability $\frac{\text{mean}_{\hat{\alpha}} - \omega}{\omega}$, (c) the proportion of estimated NT and ADF 95% confidence intervals that contain the true population alpha (i.e., coverage of alpha), and (d) the proportion of estimated NT and ADF 95% confidence intervals that contain the true population reliability (i.e., coverage of omega).

Empirical Behavior of Sample Coefficient Alpha: Bias

Under the conditions of this simulation study, the absolute bias of population alpha ranged from −.01 to −.02, with a median of −.01. Thus, the bias of popula-

Table 4
Average Coverage Rates for Normal-Theory (NT) and Asymptotically Distribution-Free (ADF) 95% Confidence Intervals at Each Level of Sample Size and Skewness, When Population Coefficient Alpha Equals True Reliability

		Skewness				
N	Method	0	0.41	0.98	1.96	2.67
50	ADF	.92	**.94**	**.92**	**.89**	**.86**
	NT	**.94**	**.94**	**.92**	.85	.80
100	ADF	**.94**	**.94**	**.93**	**.92**	**.90**
	NT	**.94**	**.94**	.92	.86	.80
200	ADF	**.94**	**.94**	**.94**	**.93**	**.93**
	NT	**.94**	**.94**	.92	.86	.80
400	ADF	**.95**	**.95**	**.95**	**.95**	**.94**
	NT	.94	.94	.93	.87	.81

Note. Coverage rates should be close to nominal rates (.95). Boldface type indicates the more accurate method for each combination of sample size and skewness.

tion alpha is small, as would be expected in typical applications in which a congeneric model holds (McDonald, 1999).

The same trends regarding the bias and variability of sample alpha observed in the previous simulation were found in this simulation. First, the bias of sample coefficient alpha in estimating population reliability increases with decreasing population reliability. Second, bias is consistently negative. In other words, the point estimate of coefficient alpha consistently underestimates the true population reliability. Third, the variability of the bias increases with decreasing sample size. For fixed sample size and true reliability, bias increases with increased kurtosis and increased skewness.

However, in this simulation, the magnitude of the bias is larger. In the first simulation, when population coefficient alpha equals reliability, the bias of sample alpha was negligible (relative bias less than 5%), provided that (a) sample size was equal to or larger than 100 and (b) population reliability was larger than .3. In contrast, when population coefficient alpha underestimates the reliability of test scores, relative bias is negligible, provided sample size is larger than 100 only whenever population reliability is larger than .6. This is because in this simulation sample, alpha combines the effects of two sources of downward bias. One source of downward bias is the bias of the true population alpha. The second source of downward bias is induced by using a small sample size.

The results of the two combined sources of downward bias are displayed in Figure 5. In this figure, we have plotted the absolute bias of sample alpha as a function of the true population reliability by sample size. Because the absolute bias of population alpha equals (to two significant digits) the estimated bias of sample alpha when sample size is 1,000, the points in this figure for sample size 1,000 are also the absolute bias of population alpha. We see in this figure that absolute bias of population alpha ranges from −.01 to −.02. We also see in this figure that the underestimation does not increase much when sample size is 400 or larger. However, the underestimation increases substantially for sample size 100 if the population reliability is .6 or smaller.

Do NT and ADF Standard Errors Accurately Estimate the Variability of Coefficient Alpha?

It is interesting to investigate how accurately NT and ADF standard errors estimate the variability of sample alpha

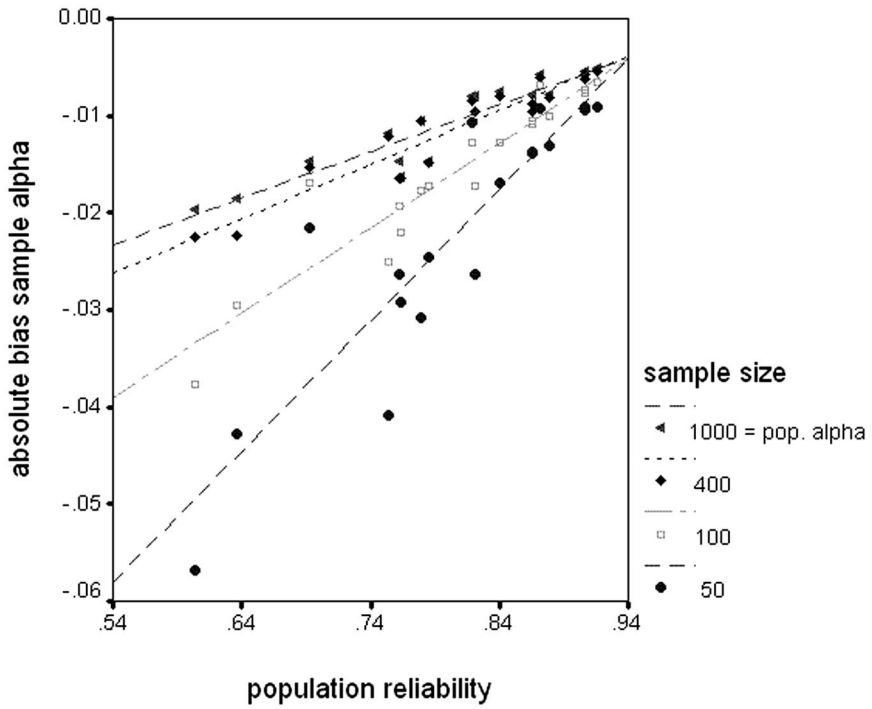

Figure 5. Absolute bias of the coefficient-alpha point estimates as a function of the true population reliability when population alpha underestimates true reliability. A linear model has been fit to the points to model the relationship between bias and true reliability by sample size. Bias increases with decreasing sample size and decreasing population reliability. The absolute bias of population alpha equals the estimated bias of sample alpha (to two significant digits), when sample size is 1,000. Therefore, the points for sample size 1,000 are also the absolute bias of population alpha.

when population alpha is a biased estimator of reliability. The mean standard errors versus the standard deviations of sample alpha for each of the conditions investigated are shown separately for NT and ADF in Figure 6.

Ideally, for every condition, the mean of the standard errors should be equal to the standard deviation of sample alpha. This ideal situation has been plotted along the diagonal of the scatterplot. Points on the diagonal or very close to the diagonal indicate that the standard error (either NT or ADF) accurately estimates the variability of sample alpha. Points below the line indicate underestimation of the variability of sample alpha (leading to confidence intervals that are too narrow). Points above the line indicate overestimation of the variability of sample alpha (leading to confidence intervals that are too wide). As can be seen in Figure 6, neither NT nor ADF standard errors are too large. Also, the accuracy of NT standard errors depends on the excess kurtosis of the items, whereas the accuracy of ADF standard errors depends on sample size. NT standard errors negligibly underestimate the variability of alpha when excess kurtosis is less than 1. However, when excess kurtosis is larger than 1, the underestimation of NT standard errors can no longer be neglected, particularly as the variability of sample alpha increases. On the other hand, Figure 6 shows that for sample sizes greater than or equal to 400, ADF standard errors are exactly on target. ADF standard errors underestimate the variability of sample alpha for smaller sample sizes, but for sample sizes over 100 ADF standard errors are more accurate than NT standard errors.

We next investigated how the bias of sample coefficient alpha and the accuracy of its standard errors affect the accuracy of the NT and ADF interval estimators.

Do NT and ADF Interval Estimators Accurately Estimate Population Coefficient Alpha?

Figure 7 shows the proportion of times that 95% confidence intervals for alpha include population alpha as a function of kurtosis and sample size. Coverage rates should be close to nominal rates (95%). For items with excess kurtosis less than 1, the behavior of both estimators is somewhat similar: Both estimators accurately estimate population coefficient alpha, with NT confidence intervals being slightly more accurate than ADF confidence intervals when sample size is 50. However, for items with excess kurtosis higher than 1, coverage rates of NT confidence intervals decrease dramatically for increasing kurtosis, regardless of sample size. On the other hand, ADF confidence intervals remain accurate regardless of kurtosis, provided that sample size is at least 400. As sample size decreases, ADF intervals become increasingly more inaccurate. However, they maintain a coverage rate of at least 90% when sample size is 100.

Table 5 provides the average coverage for NT and ADF 95% confidence intervals at each level of sample size and item kurtosis. This table reveals that the average coverage of ADF intervals is as good as or better than the average coverage of NT intervals whenever sample size is 400. Even

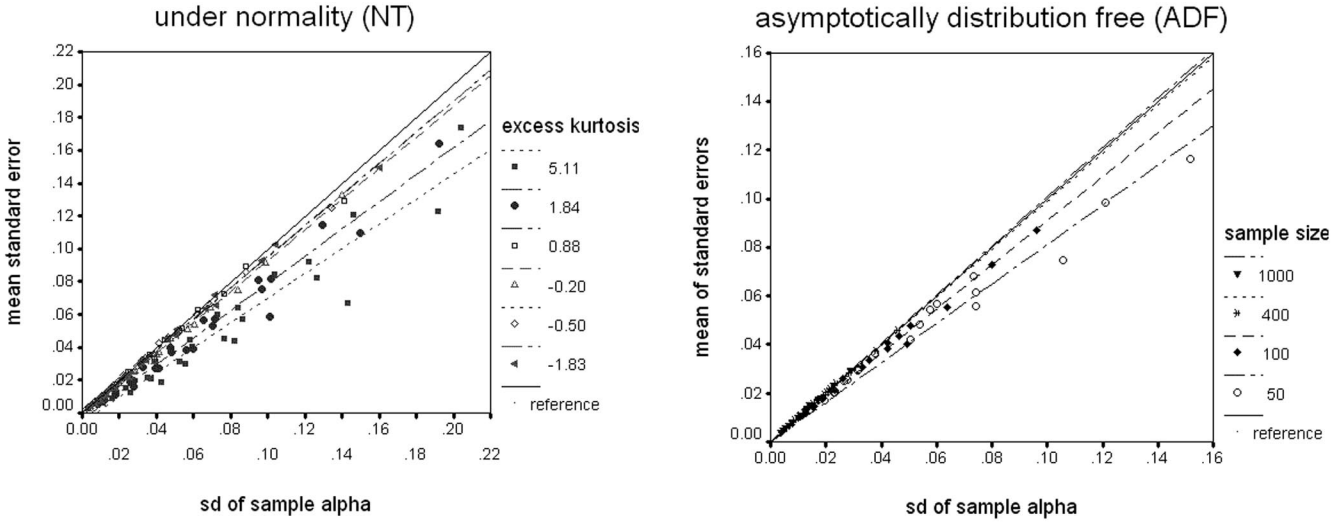

Figure 6. Scatterplot of mean standard errors (SEs) versus standard deviation of sample coefficient alpha. The mean SEs should be equal to the standard deviation of sample coefficient alpha. This is indicated by the reference line in the diagonal of the graph. Points below the line indicate underestimation of the variability of sample coefficient alpha. Normal-theory (NT) SEs underestimate the variability of coefficient alpha when excess kurtosis > 1. Asymptotically distribution-free (ADF) SEs underestimate the variability of coefficient alpha, when sample size ≤ 100. Across levels of kurtosis, ADF SEs are more accurate than NT SEs, provided sample size ≥ 100.

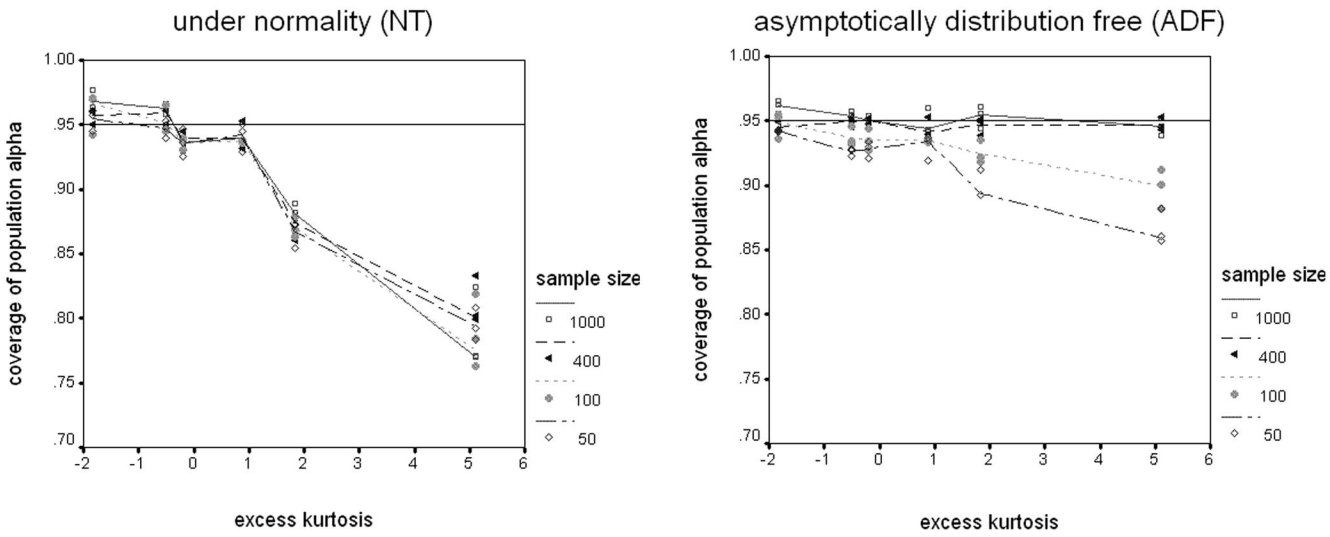

Figure 7. Proportion of times (coverage) that 95% confidence intervals (CIs) for coefficient alpha include population coefficient alpha as a function of kurtosis and sample size. Data have been generated according to a congeneric model. Coverage rates should be close to nominal rates (95%). A nonparametric procedure has been used to model the relationship between coverage and excess kurtosis by sample size. The accuracy of both CIs is similar (and adequate) for items with low excess kurtosis (< 1). For items with higher excess kurtosis, asymptotically distribution-free (ADF) intervals are more accurate, particularly when sample size is greater than 100 observations.

with sample sizes of 100, ADF confidence intervals are preferable to NT intervals, because the NT intervals underestimate coefficient alpha when excess kurtosis is larger than 1. Only with sample sizes of 50 do NT confidence intervals consistently outperform ADF intervals when excess kurtosis is less than 1, and even in this situation, the advantage of NT over ADF intervals is small.

In summary, ADF intervals are preferable to NT intervals. They portray accurately the population alpha, even when this underestimates true reliability, provided sample size is at least 100. However, because population alpha underestimates the true reliability, it is of interest to investigate the extent to which ADF and NT confidence intervals are able to capture true reliability.

Do NT and ADF Interval Estimators Accurately Estimate Population Reliability?

Figure 8 shows the proportion of times (coverage) that 95% confidence intervals for coefficient alpha include the true reliability of the test scores as a function of kurtosis and sample size. For items with excess kurtosis less than 1, the

Table 5
Average Coverage of Population Coefficient Alpha for Normal-Theory (NT) and Asymptotically Distribution-Free (ADF) 95% Confidence Intervals at Each Level of Sample Size and Kurtosis, When Population Coefficient Alpha Underestimates True Reliability

		Excess kurtosis					
N	Method	−1.83	−0.50	−0.20	0.88	1.84	5.11
50	ADF	.94	.93	.93	.93	**.90**	**.87**
	NT	**.95**	**.95**	**.94**	**.94**	.87	.79
100	ADF	**.95**	.94	**.94**	**.94**	**.93**	**.90**
	NT	.96	**.95**	**.94**	**.94**	.87	.79
400	ADF	**.95**	**.95**	**.95**	**.94**	**.95**	**.95**
	NT	.96	.96	.94	**.94**	.87	.81
1000	ADF	**.96**	**.95**	**.95**	**.95**	**.95**	**.95**
	NT	.97	.96	.94	.94	.88	.79

Note. Coverage rates should be close to nominal rates (.95). Boldface type indicates the more accurate method for each combination of sample size and kurtosis.

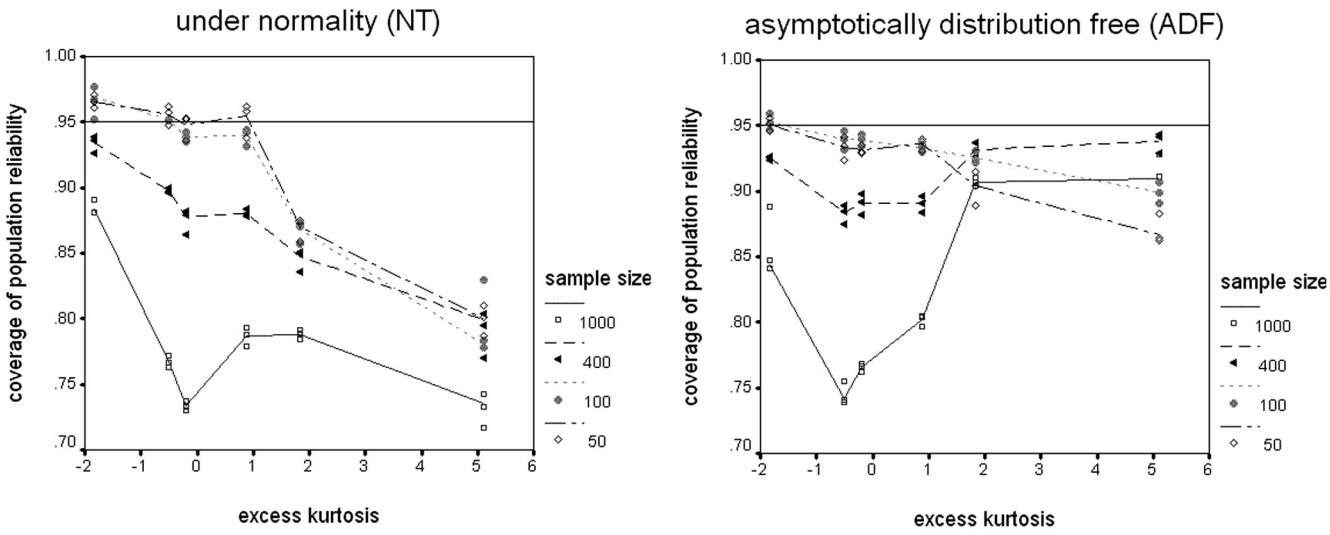

Figure 8. Proportion of times (coverage) that 95% confidence intervals (CIs) for coefficient alpha include population reliability as a function of kurtosis. Data have been generated according to a congeneric model, and population coefficient alpha is smaller than population reliability. As a result, coverage rates should be smaller than nominal rates (95%). A nonparametric procedure has been used to model the relationship between coverage and excess kurtosis by sample size. The accuracy of both CIs is similar for items with low excess kurtosis (< 1). For items with higher kurtosis, asymptotically distribution-free (ADF) CIs are more accurate.

behavior of both estimators is somewhat similar. Confidence intervals contain the true reliability only when sample size is less than 400. For larger sample sizes, confidence intervals for alpha increasingly miss true reliability.

For excess kurtosis larger than 1, the behavior of both confidence intervals is different. NT confidence intervals miss population reliability, and they do so with increasing sample size. On the other hand, ADF intervals for population alpha are reasonably accurate at including the true population reliability (coverage over 90%), provided sample size is larger than 100. They are considerably more accurate than NT intervals, even with a sample size of 50.

To understand these findings, one must notice that the confidence intervals for coefficient alpha can be used to test the null hypothesis that the population alpha equals a fixed value; for instance, $\alpha = .60$. In Figure 7, we examine whether the confidence intervals for alpha include the population alpha. This is equivalent to examining the empirical rejection rates at a $(1 - .95) = 5\%$ level of a statistic that tests for each condition whether $\alpha = \alpha_0$, where α_0 is the population alpha in that condition. In contrast, in Figure 8, we examine whether the confidence intervals for alpha include the population reliability, which is given by coefficient omega, say ω_0. This is equivalent to examining the empirical rejection rates at a 5% level of a statistic that tests for each condition whether $\alpha = \omega_0$, where ω_0 is the population reliability in that condition. However, in this simulation, study population alpha is smaller than population reliability. Thus, the null hypothesis is false, and the coverage rates shown in Figure 8 are equivalent to empirical power rates.

Figure 8 shows that when items are close to being normally distributed, both confidence intervals have power to distinguish population alpha from the true reliability, when sample size is large. In other words, when sample size is large and the items are close to being normally distributed, both interval estimators will reject the null hypothesis that population alpha equals the true population reliability. On the other hand, when excess kurtosis is greater than 1, the ADF confidence intervals, but not the NT confidence intervals, will contain the true reliability. The ADF confidence interval contains the true reliability in this case, because it does not have enough power to distinguish population alpha from true reliability, even with a sample size of 1,000. However, the NT confidence intervals do not contain the true reliability, because, as we show in Figure 7, they do not contain alpha.

These findings are interesting. A confidence interval is most useful when sample coefficient alpha underestimates true reliability the most, which is when sample size is small. It is needed the least when sample size is large (i.e., 1,000), because in this case, sample alpha underestimates true reliability the least. When sample size is small, the ADF interval estimator may compensate for the bias of sample alpha, because the rate with which it contains true reliability is acceptable (over 90% for 95% confidence intervals).

However, when sample size is large and items are close to being normally distributed, both the NT and ADF intervals miss true reliability by, on average, the difference between true reliability and population coefficient alpha. Under the conditions of our simulation study, this difference is at most .02.

Discussion

Coefficient alpha equals the reliability of the test score when the items are tau-equivalent, that is, when they fit a one-factor model with equal factor loadings. In applications, this model seldom fits well. In this case, applied researchers have two options: (a) Find a better-fitting model and use a reliability estimate based on such model, or (b) use coefficient alpha.

If a good-fitting model can be found, the use of a model-based reliability estimate is clearly the best option. For instance, if a one-factor model is found to fit the data well, then the reliability of the test score is given by coefficient omega, and the applied researcher should employ this coefficient. Although this approach is preferable in principle, there may be practical difficulties in implementing it. For instance, if the best-fitting model is a hierarchical factor-analysis model, it may not be straightforward to many applied researchers to figure out how to compute a reliability estimate on the basis of the estimated parameters of such a model. Also, model-based reliability estimates depend on the method used to estimate the model parameters. Thus, for instance, different coefficient-omega estimates will be obtained for the same dataset, depending on the method used to estimate the model parameters: ADF, maximum likelihood, unweighted least squares, and so on. There has not been much research on which of these parameter-estimation methods lead to the most accurate reliability estimate.

Perhaps the most common situation in applications is that no good-fitting model can be found (i.e., the model is rejected by the chi-square-test statistic). That is, the best-fitting model has a nonnegligible amount of model misfit. In this case, an applied researcher can still compute a model-based reliability estimate on the basis of his or her best-fitting model. Such a model-based reliability estimator will be biased. The direction and magnitude of this bias will be unknown, because it depends on the direction and magnitude of the discrepancy between the best-fitting model and the unknown true model. When no good-fitting model can be found, the use of coefficient alpha as an estimator of the true reliability of the test score becomes very attractive for two reasons. First, coefficient alpha is easy to compute. Second, if the mild conditions discussed in Bentler (in press) are satisfied, the direction of the bias of coefficient alpha is known: It provides a conservative estimate of the true reliability. These reasons explain the popularity of alpha among applied researchers.

As with any other statistic, sample coefficient alpha is subject to variability around its true parameter, in this case, the population coefficient alpha. The variability of sample coefficient alpha is a function of sample size and the true population coefficient alpha. When the sample size is small and the true population coefficient alpha is not large, the sample-coefficient-alpha point estimate may provide a misleading impression of the true population alpha and, hence, of the reliability of the test score.

Furthermore, sample coefficient alpha is consistently biased downward. It is therefore more likely to yield a misleading impression of poor reliability. The magnitude of the bias is greatest precisely when the variability of sample alpha is greatest (small population reliability and small sample size). The magnitude is negligible when the model assumptions underlying alpha are met (i.e., when coefficient alpha equals the true reliability). However, as coefficient alpha increasingly underestimates reliability, the magnitude of the bias need not be negligible.

To take into account the variability of sample alpha, one should use an interval estimate instead of a point estimate. In this paper, we investigated the empirical performance of two confidence-interval estimators for population alpha under different conditions of skewness and kurtosis, as well as sample size: (a) the confidence intervals proposed by van Zyl et al. (2000), who assumed that items are normally distributed (NT intervals), and (b) the confidence intervals proposed by Yuan et al. (2003), on the basis of asymptotic distribution-free assumptions (ADF intervals). Our results suggest that when the model assumptions underlying alpha are met, ADF intervals are preferred to NT intervals, provided sample size is larger than 100 observations. In this case, the empirical coverage rate of the ADF confidence intervals is acceptable (over .90 for 95% confidence intervals), regardless of the skewness and kurtosis of the items. Even with samples of size 50, the NT confidence intervals outperform the ADF confidence intervals only when skewness is zero.

We found similar results for the coverage of alpha when we generated data in which coefficient alpha underestimates true reliability. Also, our simulations revealed that the confidence intervals for alpha may contain the true reliability. In particular, we found that if the bias of population alpha is small, as in typical applications in which a congeneric measurement model holds, the ADF intervals contain true reliability for items with excess kurtosis larger than 1. If item excess kurtosis is smaller than 1 (i.e., close to being normally distributed), ADF intervals also contain population reliability, for samples smaller than 400. For larger samples, the ADF intervals underestimate population reliability slightly, because there is power to distinguish between true reliability and population alpha. For near normally distributed items, the behavior of NT intervals is similar. However, for items with excess kurtosis larger than

1, NT confidence intervals miss the true reliability of the test, because they do not even contain coefficient alpha.

As with any other simulation study, our study is limited by the specification of the conditions employed. For instance, when generating congeneric items, population alpha underestimated population reliability only slightly, by a difference of between −.02 and −.01. This amount of misspecification was chosen to be typical in applications (McDonald, 1999). Further simulation studies are needed to explore whether the robustness of the interval estimators for coefficient alpha hold (i.e., whether they contain population coefficient alpha) under alternative-model misspecification, such as bifactor models. Also, as the bias of population alpha increases, confidence intervals for alpha should not include the population reliability. Finally, further research should compare the symmetric confidence intervals employed here against asymmetric confidence intervals, because the upper limit of symmetric confidence intervals for alpha may exceed the upper bound of 1 when sample alpha is near 1.

Conclusions

Following Duhachek and Iacobucci (2004), we strongly encourage researchers to report confidence intervals as well as point estimates of coefficient alpha when evaluating the reliability of a test score. Failing to do so may result in an underestimation of the true population coefficient alpha of the test score, leading to rejection of reliable tests. NT confidence intervals can be safely used when items are approximately normally distributed. Also, NT intervals can be used with very small sample sizes, provided items are approximately normally distributed. Duhachek and Iacobucci reported that accurate NT confidence intervals can be obtained with sample sizes as small as 30.

Because test and questionnaire items are usually ordered categorical variables, they may show considerable skewness and kurtosis, thereby violating the normality assumption. Accurately estimating the standard errors without normality assumptions requires larger samples, but our results indicate that for sample sizes over 100, the ADF confidence intervals provide an accurate perspective on population alpha. In fact, for sample sizes over 100, they are definitely preferred to NT confidence intervals if the items show skewness over 1 or excess kurtosis over 1. Also, when item responses greatly depart from normality (as in questionnaires measuring rare events), the difference between the NT and ADF intervals can be substantial.

References

Barchard, K. A., & Hakstian, R. (1997). 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. (in press). Covariance structure models for maximal reliability of unit-weighted composites. In S.-Y. Lee (Ed.), *Handbook of structural equation models.* Amsterdam: Elsevier.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, MA: Cambridge University Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

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

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

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. (1974). The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory. *Behavioral Science, 19,* 1–15.

Duhachek, A., Coughlan, A. T., & Iacobucci, D. (2005). Results on the standard error of the coefficient alpha index of reliability. *Marketing Science, 24,* 294–301.

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

Feldt, L. S. (1965). The approximate sampling distribution of Kuder–Richardson reliability coefficient twenty. *Psychometrika, 30,* 357–370.

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

Hakstian, A. R., & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219–231.

Hartmann, W. M. (2005). Resampling methods in structural equation modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift to R. P. McDonald* (pp. 341–376). Mahwah, NJ: Erlbaum.

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.

Hu, L.-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351–362.

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

Kristof, W. (1963). The statistical theory of stepped-up reliability when a test has been divided into several equivalent parts. *Psychometrika, 28,* 221–238.

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

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of nonnormal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19–30.

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

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

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

Yuan, K.-H., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficients. *Psychometrika, 67,* 251–259.

Yuan, K.-H., Guarnaccia, C. A., & Hayslip, B. (2003). A study of the distribution of sample coefficient alpha with the Hopkins symptom checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement, 63,* 5–23.

Appendix

Technical Details

Computation of the Normal-Theory (NT) and Asymptotically Distribution-Free (ADF) Standard Errors of Sample Alpha

In matrix notation, population alpha is $\alpha = \frac{p}{p-1}\left(1 - \frac{\text{tr}(\boldsymbol{\Sigma})}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right)$, where $\boldsymbol{\Sigma}$ is the covariance matrix of the items in the population, tr() denotes the trace operator, and $\mathbf{1}$ is a $p \times 1$ vector of ones. Sample alpha is $\hat{\alpha} = \frac{p}{p-1}\left(1 - \frac{\text{tr}(\mathbf{S})}{\mathbf{1}'\mathbf{S}\mathbf{1}}\right)$, where $\mathbf{S}$ denotes the sample covariance matrix.

Let $\mathbf{s} = \text{vecs}(\mathbf{S})$, and let $\boldsymbol{\sigma} = \text{vecs}(\boldsymbol{\Sigma})$, where vecs() is an operator that takes the elements of a symmetric matrix on or below the diagonal and stacks them onto a column vector. Asymptotically (i.e., in large samples), the vector $\sqrt{N}\,\mathbf{s}$ is normally distributed with mean $\boldsymbol{\sigma}$ and covariance matrix $\boldsymbol{\Gamma}$ of dimensions $q \times q$. Because $\hat{\alpha}$ is a function of $\mathbf{s}$, asymptotically, $\hat{\alpha}$ is normally distributed with mean α and variance

$$\varphi^2 = \frac{1}{N}\boldsymbol{\delta}'\boldsymbol{\Gamma}\boldsymbol{\delta}, \tag{3}$$

where $\boldsymbol{\delta}' = \frac{\partial \alpha}{\partial \boldsymbol{\sigma}'}$ is a $1 \times q$ vector of derivatives of α with respect to $\boldsymbol{\sigma}$. The elements of $\boldsymbol{\delta}$ are:

$$\frac{\partial \alpha}{\partial \sigma_{ij}} = \begin{cases} \frac{-p}{p-1}\frac{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} - \text{tr}(\boldsymbol{\Sigma})}{(\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1})^2} & \text{if} \quad i = j \\ \frac{2p}{p-1}\frac{\text{tr}(\boldsymbol{\Sigma})}{(\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1})^2} & \text{if} \quad i \neq j. \end{cases} \tag{4}$$

The above results hold under NT assumptions but also under ADF assumptions. However, the $\boldsymbol{\Gamma}$ matrix differs under NT and ADF assumptions. Henceforth, we use $\boldsymbol{\Gamma}_{\text{NT}}$ and $\boldsymbol{\Gamma}_{\text{ADF}}$ to distinguish them.

If we are willing to assume that the test items are normally distributed, then Equation 3 can be estimated as (van Zyl et al., 2000)

$$\hat{\varphi}^2_{\text{NT}} = \frac{1}{N}\frac{p^2}{(p-1)^2}\frac{2[(\mathbf{1}'\mathbf{S}\mathbf{1})(\text{tr}(\mathbf{S}^2) + \text{tr}(\mathbf{S})^2) - 2\text{tr}(\mathbf{S})(\mathbf{1}'\mathbf{S}^2\mathbf{1})]}{(\mathbf{1}'\mathbf{S}\mathbf{1})^3}. \tag{5}$$

On the other hand, estimation of the asymptotic variance of sample coefficient alpha under ADF assumptions requires estimating $\boldsymbol{\Gamma}_{\text{ADF}}$. Let $\mathbf{y}_i$ be the $p \times 1$ vector of data for observation i, and $\bar{y}$ be the $p \times 1$ vector of sample means. Also, let $\mathbf{s}_i = \text{vecs}[(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})']$ be a $q \times 1$ vector of squared deviations from the mean. Then, $\boldsymbol{\Gamma}_{\text{ADF}}$ can be estimated (Satorra & Bentler, 1994) as

$$\hat{\boldsymbol{\Gamma}}_{ADF} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{s}_i - \mathbf{s})(\mathbf{s}_i - \mathbf{s})'. \tag{6}$$

However, an estimate of the asymptotic variance of coefficient alpha under ADF assumptions can be obtained directly without storing $\hat{\boldsymbol{\Gamma}}_{ADF}$ using

$$\hat{\varphi}^2_{ADF} = \frac{1}{N}\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Gamma}}_{ADF}\hat{\boldsymbol{\delta}} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\delta}}'(\mathbf{s}_i - \mathbf{s})\right)^2. \tag{7}$$

To see this, insert Equation 6 in Equation 3,

$$\hat{\varphi}^2_{ADF} = \frac{1}{N}\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Gamma}}_{ADF}\hat{\boldsymbol{\delta}} = \frac{1}{N}\hat{\boldsymbol{\delta}}'\left\{\frac{1}{N-1}\left[\sum_{i=1}^{N}(\mathbf{s}_i - \mathbf{s})(\mathbf{s}_i - \mathbf{s})'\right]\right\}\hat{\boldsymbol{\delta}}$$
$$= \frac{1}{N(N-1)}\left[\sum_{i=1}^{N}\hat{\boldsymbol{\delta}}'(\mathbf{s}_i - \mathbf{s})(\mathbf{s}_i - \mathbf{s})'\hat{\boldsymbol{\delta}}\right],$$

but because $\hat{\boldsymbol{\delta}}'$ is a $1 \times q$ vector and $(\mathbf{s}_i - \mathbf{s})$ is a $q \times 1$ vector, $\hat{\boldsymbol{\delta}}'(\mathbf{s}_i - \mathbf{s})$ is a scalar. As a result, $\hat{\boldsymbol{\delta}}'(\mathbf{s}_i - \mathbf{s})(\mathbf{s}_i - \mathbf{s})'\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\delta}}'(\mathbf{s}_i - \mathbf{s}))^2$, and we obtain Equation 7. Our SAS macro computes the NT standard error of $\hat{\alpha}$ via Equation 5, and the ADF standard error of $\hat{\alpha}$ via Equation 7.

Computation of Population Reliability for Categorized Normal Variables

To compute the population coefficient alpha, one needs the population variances and covariances. In our simulation study, each observed variable Y_i is multinomial, with $m = 2$ or 5 categories. The categories are scored as $k = 0, \ldots, m - 1$. For categorical variables,

$$\sigma_{ii} = Var[Y_i] = \left(\sum_{k=0}^{m-1} k^2 \Pr(Y_i = k)\right) - \mu_i^2, \quad (8)$$

$$\sigma_{ij} = Cov[Y_i Y_j] = \left(\sum_{k=0}^{m-1}\sum_{l=0}^{m-1} kl \Pr[(Y_i = k) \cap (Y_j = l)]\right) - \mu_i\mu_j, \quad (9)$$

where $\Pr[(Y_i = k) \cap (Y_j = l)]$ stands for the probability that item i takes the value k and item j takes the value l, and

$$\mu_i = E[Y_i] = \sum_{k=0}^{m-1} k \Pr(Y_i = k). \quad (10)$$

Data are generated as follows: First we generate multivariate normal data. In the first simulation, we used $\mathbf{z}^* \sim N(\mathbf{0},\mathbf{P})$, where $\mathbf{P} = \rho\mathbf{1}\mathbf{1}' + (1 - \rho)\mathbf{I}$. That is, the covariance matrix used to generate data is a correlation matrix with a common correlation. The normal variables are categorized via the threshold relationship $Y_i = k_i$, if $\tau_{i_k} < z_i^* < \tau_{i_{k+1}}$, $k_i = 0, \ldots, K - 1$, where $\tau_{i_0} = -\infty$ and $\tau_{i_k} = \infty$. The thresholds were selected so that the items had the marginal probabilities shown in Figure 1. In the second simulation, we used the same procedure, except that to generate multivariate normal data, we used $\mathbf{P} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{I} - diag(\boldsymbol{\lambda}\boldsymbol{\lambda}')$, where $\boldsymbol{\lambda}' = (.3, .4, .5, .6, .7, .8, .9)$, when $p = 7$. That is, in the second simulation, we generated data using a correlation matrix with a one-factor model structure.

Under this model of ordered categorized normal variables,

$$\Pr(Y_i = k_i) = \int_{\tau_{i_k}}^{\tau_{i_{k+1}}} \phi_1(z_i^*{:}0,1)dz_i^*, \quad (11)$$

$$\Pr[(Y_i = k) \cap (Y_j = k')] = \int_{\tau_{i_k}}^{\tau_{i_{k+1}}}\int_{\tau_{jk'}}^{\tau_{ik'+1}} \phi_2(z_i^*,z_j^*{:}0,0,1,1,\rho_{ij})dz_i^*dz_j^*, \quad (12)$$

where ρ_{ij} is an element of $\mathbf{P}$.

The population skewness and kurtosis reported in Figure 1 were computed using skewness $= \frac{K_3}{K_2^{3/2}}$ and excess kurtosis $= \frac{K_4}{K_2^2} - 3$, where

$$K_m = \sum_{k=0}^{K-1}[(k - \mu_i)^m \Pr(Y_i = k)], \quad m = 2, \ldots, 4, \quad (13)$$

and μ_i is the population mean given in Equation 10.

Also, the population correlation between two items can be obtained using $\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ and Equations 8 and 9. Finally, the average population interitem correlation is

$$\bar{\rho} = \frac{1}{q}\sum_{i<j}\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \quad (14)$$

where $q = \frac{p(p+1)}{2}$.

To illustrate, consider the condition with $p = 5$ items of Type 3 in Figure 1 and $\mathbf{P} = \rho\mathbf{1}\mathbf{1}' + (1 - \rho)\mathbf{I}$, where $\rho = .8$. We generated multivariate normal data with mean zero and correlation structure $\mathbf{P}$. We dichotomized the data using the threshold $\tau = .253$, as this is the threshold that yields Type 3 items. To obtain the population alpha, we computed the population covariance matrix using Equations 8–12. For this condition, all the variances in $\boldsymbol{\Sigma}$ are equal to .24, and all covariances are equal to .11. As a result, the population $\alpha = .796$. Also, on the basis of Equation 14, the average population interitem correlation is .438. When $\mathbf{P} = \rho\mathbf{1}\mathbf{1}' + (1 - \rho)\mathbf{I}$, the covariances in $\boldsymbol{\Sigma}$ are all equal, and population alpha equals the reliability of the test score.

Consider now the case where $\mathbf{P} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{I} - diag(\boldsymbol{\lambda}\boldsymbol{\lambda}')$. In this case, the covariances in $\boldsymbol{\Sigma}$ are not equal, and as a result, population alpha underestimates reliability. When $\mathbf{P} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{I} - diag(\boldsymbol{\lambda}\boldsymbol{\lambda}')$ and the same thresholds are used for all items, the population covariance matrix $\boldsymbol{\Sigma}$

(*Appendix continues*)

obtained using Equations 8–12 can be fitted exactly by a one-factor model, say $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}' + \boldsymbol{\Psi}$. In this decomposition, $\tilde{\boldsymbol{\lambda}} \neq \boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ are the factor loadings used to generate the data. Because $\boldsymbol{\Sigma}$ follows a one-factor model, population reliability is given by coefficient omega:

$$\omega = \frac{\left(\sum_{i=1}^{p} \tilde{\lambda}_i\right)^2}{\left(\sum_{i=1}^{p} \tilde{\lambda}_i\right)^2 + \sum_{i=1}^{p} \psi_i^2}, \tag{15}$$

where ψ_i^2 is the element of the diagonal matrix $\boldsymbol{\Psi}$ corresponding to the ith item. Because the model fits exactly in the population, any method can be used to estimate $\tilde{\boldsymbol{\lambda}}$ and $\boldsymbol{\Psi}$ from $\boldsymbol{\Sigma}$. They all yield the same result.

To illustrate, consider the condition with $p = 7$ items of Type 3 in Figure 1. Before dichotomization, the simulated data has population correlation matrix $\mathbf{P} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{I} - diag(\boldsymbol{\lambda}\boldsymbol{\lambda}')$, with $\boldsymbol{\lambda}' = (.3, .4, .5, .6, .7, .8, .9)$. We dichotomized the data, using the threshold $\tau = .253$ to obtain Type 3 items. Now, using Equations 8–12, we obtain the following population covariance matrix:

$$\sum = \begin{bmatrix} .24 & .02 & .02 & .03 & .03 & .04 & .04 \\ .02 & .24 & .03 & .04 & .04 & .05 & .06 \\ .02 & .03 & .24 & .05 & .05 & .06 & .06 \\ .03 & .04 & .05 & .24 & .07 & .08 & .09 \\ .03 & .04 & .05 & .07 & .24 & .09 & .10 \\ .04 & .05 & .06 & .08 & .09 & .24 & .24 \\ .04 & .06 & .07 & .09 & .10 & .12 & .24 \end{bmatrix}.$$

This $\boldsymbol{\Sigma}$ follows a one-factor model where $\boldsymbol{\lambda}' = (.11, .15, .19, .23, .28, .33, .37)$, and the elements of the diagonal matrix $\boldsymbol{\Psi}$ are (.22, .22, .20, .18, .16, .13, .10). Thus, for this condition, the population $\alpha = .677$, and the population reliability is $\omega = .692$.

Received October 24, 2005
Revision received December 20, 2006
Accepted December 27, 2006 ■