ASYMPTOTICALLY DISTRIBUTION FREE (ADF) INTERVAL ESTIMATION OF COEFFICIENT ALPHA

IE Working Paper WP06-24 05-12-2006

Alberto Maydeu Olivares

Instituto de Empresa
Marketing Dept.
C/Maria de Molina 11-15
28006 Madrid – Spain
Alberto.Maydeu@ie.edu

Donna L. Coffman

The Methodology Center
The Pensylvania State University
204E, Calder Way, Ste. 400
State College, PA 16801-USA.
dlc30@psu.edu

Abstract

Asymptotic distribution free (ADF) interval estimators for coefficient alpha were introduced in the context of an application by Yuan, Guarnaccia, and Hayslip (2003). Here, simulation studies were performed to investigate the behavior of ADF vs. normal theory (NT) interval estimators of coefficient alpha for tests composed of ordered categorical items under varied conditions of sample size, item skewness and kurtosis, number of items, and average inter-item correlation. NT intervals were found to be inaccurate when item *skewness* > 1 or *kurtosis* > 4. But for sample sizes over 100 observations, ADF intervals provide an accurate perspective of the population coefficient alpha of the test 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 a SAS macro.

Keywords

coefficient omega, reliability, Likert-type ítems.

1. Introduction

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) points 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 using the sample coefficient alpha, a point estimator of the population coefficient alpha. As with any other 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 Iacobucci & Duhachek, 2003, and Duhachek, Coughlan, & Iacobucci, 2005) have made a compelling argument to use an interval estimator for coefficient alpha instead of a point estimator.

Methods for obtaining interval estimators for coefficient alpha have a long history (see Duhachek and 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 population covariance matrix, *and* the observed data followed a particular distribution, then a confidence interval for coefficient alpha could be obtained. The sampling distribution for coefficient alpha was independently derived by Kristof (1963) and Feldt (1965) assuming 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. However, Barchard and Haskstian (1997) found that confidence intervals for coefficient alpha obtained using these results were not sufficiently accurate when model assumptions were violated (i.e. the items were not strictly parallel). As Duhachek and Iacobucci (2004) have suggested, the lack of robustness of the interval estimators for coefficient alpha to violations of model assumptions have hindered the widespread use of interval estimators for coefficient alpha 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. (2000) does not require the assumption of compound symmetry. In particular, these authors assumed only that the items composing 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. (2000) 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. (2000) 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) have shown 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 non-normal, the researcher can not verify if the necessary conditions put forth by Yuan and Bentler (2002) 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) have compared the performance of the NT confidence intervals of van Zyl et al. (2000) to a newly proposed model-free asymptotically distribution free (ADF) confidence interval, and several confidence intervals based on bootstrapping. Yuan et al. (2003) concluded that the ADF intervals were more accurate for the Likert-type items of the HSCL than the NT intervals, but less accurate than the bootstrapping procedures.

Also, as Yuan et al. (2003: p. 7) point 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. (2003) versus the NT interval estimator proposed by van Zyl et al. (2000) with Likert-type data [2]. In so doing, we consider conditions where 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 formulae that enable the computation of these confidence intervals for tests of any size. Yuan et al. (2003) did not provide these simplifications and practical use of their equations would be limited in the number of variables. Further, we provide a 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 $X = Y_1 + \mathrm{L} + Y_p$, 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 (Guttman, 1945; Cronbach, 1951). In the population of respondents, coefficient alpha is

$$a = \frac{p}{p-1}\left(1 - \frac{\sum_i s_{ii}}{\sum_{ij} s_{ij}}\right), \qquad (1)$$

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

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

where s_{ij} denote the sample covariance between items i and j, and s_{ii} denote 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* (a.k.a. *essentially tau-equivalent* items) in the population (Lord & Novick, 1968: p. 50; McDonald, 1999: Chapter 6). A true-score equivalent model is simply a one factor model for the item scores where the factor loadings are equal for all items. The model implies that the population covariances are all equal, but that the population variances are not equal for all items.

A special case of the true-score equivalent model is the *parallel items* model, where 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 items 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.

However, when the items do not conform to a true score 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 (a.k.a., *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 (say .9) and all the other factor loadings are very small (say .2) (Raykov, 1997). This condition is rarely encountered in practical applications.

NT and ADF interval estimators for coefficient alpha

This section summarizes the main results regarding the large sample distribution of sample coefficient alpha. Technical details can be found in the Appendix.

In large samples, $\hat{a}$ is normally distributed with mean α and variance j^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 $\hat{a} - z_{x/2}\hat{j}$, whereas the upper limit, U_L, is $\hat{a} + z_{x/2}\hat{j}$. $\hat{j}$ 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 α.

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, j^2 depends only on population variances and covariances (bivariate moments), whereas under ADF assumptions j^2 depends on fourth order moments (see Browne 1982, 1984 for further details).

Under normality assumptions, j^2 can be estimated from the sample variances and covariances (see the Appendix). In contrast, the estimation of j^2 under ADF assumptions requires computing an estimate of the asymptotic covariance matrix of the sample variances and covariances. This is a $q \times q$ matrix, where $q = \frac{p(p+1)}{2}$. 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 a SAS macro which is available from the authors upon request. The macro is easy to use for applied researchers. 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 for the NT confidence intervals. The macro also provides the NT confidence interval.

Some considerations in the use of NT vs. 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 to the NT confidence intervals. On the other hand, because ADF confidence intervals are robust to non-normality in large samples, we expect that when the test items present high skewness and/or kurtosis, the ADF confidence intervals will be more accurate than the NT confidence intervals. In other words, we expect that when the items are markedly non-normal and large samples are available the ADF confidence intervals will be more accurate than the NT confidence intervals. Yet, we expect that when the data approaches 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 non-normality the ADF confidence intervals are more accurate than NT confidence intervals. This will be investigated in the next sections by means of simulation.

Two simulation studies were performed. In the first simulation, data were simulated so that population alpha equals the reliability of the test score. In the second simulation, data were simulated so that population alpha underestimates the reliability of the test score. This occurs for instance when the model underlying the item scores is a one factor model with unequal factor loadings (e.g., McDonald, 1999).

Previous research (e.g., Hu, Bentler & Kano, 1992; Curran, West & Finch, 1996) 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 1000 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 simply 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, Bentler, and Kano (1992) and Curran, West, and Finch (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. Even though the ADF asymptotic covariance matrix of the sample variances and covariances can be quite unstable in small samples, we concentrate its information to estimate a single standard error, that 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. This was investigated by means of two simulation studies to which we now turn.

2. A Monte Carlo investigation of NT vs. 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, ordinal data were used in the simulation study. The procedure used to generate the data was similar to that of Muthén and Kaplan (1985, 1992). It enables 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) in 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, $\boldsymbol{\Sigma}$, 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 over-estimate it.

In the first simulation study, $\mathbf{P}$ had all its elements equal. Also, the same thresholds were used for all items. These choices result 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, $\boldsymbol{\Sigma}$ is consistent with a parallel items model. This simplifies the presentation of the findings as all items have a common skewness and kurtosis.

Overall, we investigated 144 conditions. These were obtained by crossing

a) 4 sample sizes (50, 100, 200, and 400 respondents)
b) 2 test lengths (5 and 20 items)

c) 3 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.
d) 6 item types (3 types consist of items with 2 categories, and 3 types consist of items with 5 categories), that varied in skewness and/or kurtosis.

The sample sizes were chosen to be very small to large in typical questionnaire development applications. Also, 5 and 20 items are the typical shortest and longest lengths for questionnaires measuring a single attribute. Finally, we include items with typical low (.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 items 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 to 3 consist of only two categories. Type 1 items have the highest skewness and kurtosis. The threshold was chosen such that only 10% of the respondents endorse the items. Type 2 items are endorsed by 15% of the respondents, resulting in smaller values of skewness and kurtosis. Items of Types 1 and 2 are typical of applications where items are seldom endorsed. On the other hand, Type 3 items are endorsed by 40% of the respondents. These items have low skewness and their kurtosis is smaller than that of a standard normal distribution [5]. Items of Types 4 through 6 consist of 5 categories. The skewness and kurtosis of Type 5 items closely match those of a standard normal distribution. Type 4 items are also symmetric (skewness = 0), however, the kurtosis is higher than that of a standard normal distribution. These items can be found in applications where the middle category reflects an undecided position and a large number of respondents choose this middle category. Finally, Type 6 items show 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.

Insert Figure 1 about here

For each of the 144 conditions, 1000 replications were obtained. 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 $\text{bias}(\hat{a}) = \frac{\text{mean}_{\hat{a}} - a}{a}$, (b) the relative bias of the NT and ADF standard errors as $\text{bias}(\hat{f}) = \frac{\text{mean}_{\hat{f}} - \text{std}_{\hat{a}}}{\text{std}_{\hat{a}}}$, 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 vs. 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 **Σ** by the square root of the product of the corresponding diagonal elements of **Σ**). We refer to the former as underlying correlations, and to the latter as inter-item population correlations.

Table 1 summarizes the relationship between the average inter-item correlations in the population after categorizing the data and the underlying correlation before categorization. The average inter-item correlation is the extent of interrelatedness (i.e. internal consistency) among the items (Cortina, 1993). There are three levels for the average population inter-item correlation corresponding to the three underlying correlations. Table 1 also summarizes the population alpha corresponding to the three levels of the average population inter-item correlations. As may be seen in this table, the population coefficient alpha used in our study ranges from .25 to .97, and the population inter-item correlations range from .06 to .59. Thus, in the present study we are considering a wide range of values for both the population coefficient alpha and the population inter-item correlations.

Insert Table 1 about here

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. 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, it is reassuring to see in this figure that the coefficient alpha point estimates are remarkably robust to skewness and kurtosis for the sample sizes considered here provided sample size is larger than 100. In this case relative bias is less than 5% whenever population alpha is larger than .3.

Insert Figures 2 and 3 about here

Figure 3 depicts graphically the variability 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.

Insert Tables 2 and 3 about here

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) can also 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 inter-item correlation. For the two most extreme skewness conditions, and the highest level of average inter-item correlation considered (.36 to .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 times). NT standard errors were more accurate than ADF standard errors when skewness was less than .5 (nearly symmetrical items) and the average inter-item correlation was low (.06 to .15) or medium (.16 to .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 times), 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

We show in Figure 4 the coverage rates of NT and ADF confidence intervals as a function of skewness. We see in Figure 4 how the coverage rates of NT confidence intervals decrease dramatically as a function of the combination of increasing skewness and increasing average inter-item correlations. The coverage rates can be as low as .68 when items are severely skewed (Type 1 items) and the average inter-item correlation is high (.36 to .59).

Insert Figure 4 and Table 4 about here

We also show in this figure the coverage rates of ADF confidence intervals as a function of item skewness by sample size. We clearly see in this figure that 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.

Further insight is obtained by inspecting Table 4. In this table we provide 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 ≥ 50). Also, ADF intervals are uniformly more accurate than NT 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 almost indistinguishable. NT confidence intervals are more accurate than ADF confidence intervals only when the items are perfectly symmetric (skewness = 0) and sample size is 50. All in all, the empirical behavior of ADF confidence intervals is better than that of the NT confidence intervals.

3. A Monte Carlo investigation of NT vs. 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 score test [6]. 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. In particular, we crossed

a) 4 sample sizes (50, 100, 400, and 1000),
b) 3 test lengths (7, 14, and 21 items), and
c) the 6 item types used in the previous simulation (3 types consist of items with 2 categories, and 3 types consist of items with 5 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 model structure with 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 repeated once and for the test length with 21 items, they were repeated twice. Second, Steps 6) and 7) now consist of two parts, as we compute both the population coefficient alpha and population reliability (in this case population alpha underestimates reliability). We then examine 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 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

With these factor loadings, the absolute bias of population alpha ranges from -.01 to -.02, with a median of -.01. Thus, the bias of population alpha is small as one would expect in typical applications where a congeneric model holds (McDonald, 1999).

As for the bias of sample alpha in this setup, the same trends observed in the previous simulation study were found in this case. 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, now 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 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 both 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 1000, the points in this figure for sample size 1000 are also the absolute bias of population alpha. We see in this figure that absolute bias of population alpha ranges from -.01 to -.02, with a median of -.01. Thus, population alpha underestimates only slightly population reliability under the conditions of our simulation. 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 when population alpha is a biased estimator of reliability. To investigate this, we simply plotted the mean standard errors vs. the standard deviations of sample alpha for each of the conditions investigated. These are shown separately for NT and ADF in Figure 6.

Insert Figures 5 and 6 about here

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 estimate the variability of sample alpha. Points below the line indicate underestimation of the variability of sample alpha (leading to too narrow confidence intervals). Points above the line indicate overestimation of the variability of sample alpha (leading to too wide confidence intervals). As can be seen in Figure 5, neither NT or ADF standard errors are too large. Also, the accuracy of NT standard errors depends on the 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 kurtosis was less than 4. However, when kurtosis was larger than 4, the underestimation of NT standard errors can not longer be neglected, particularly as the variability of sample alpha increases. On the other hand, we see in Figure 6 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 investigate how the bias of sample coefficient alpha and the accuracy of standard errors affect the accuracy of the NT and ADF interval estimators.

Do NT and ADF interval estimators accurately estimate population coefficient alpha?

To answer this question, we show graphically in Figure 7 the percentage of times that 95% confidence intervals for alpha include population alpha as a function of kurtosis and sample size. In this figure coverage rates should be close to nominal rates (95%). We see in this Figure that for items with kurtosis less than 4, 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 kurtosis higher than 4, 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.
Further insight is obtained by inspecting Table 5. In this table we provide 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 with samples of size 100, ADF confidence intervals are preferable to NT intervals as the NT intervals underestimate coefficient alpha when kurtosis is larger than 4. Only at samples of size 50 does NT confidence intervals consistently outperform ADF intervals when kurtosis is less than 4, and even in this situation the advantage of NT over ADF intervals is small.

Insert Figure 7 and Table 5 about here

All in all, 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, in the conditions investigated population alpha underestimates the true reliability, and hence 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 percentage 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. We see in this Figure that for items with kurtosis less than 4, the 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.

Insert Figure 8 about here

For kurtosis larger than 4 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 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 an (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 kurtosis is higher than 4, 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 of size 1000. However, the NT confidence intervals do not contain the true reliability because, as we have seen 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. 1000) as 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 as 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 will miss true reliability. By how much? On average by the difference between true reliability and population coefficient alpha. Under the conditions of our simulation study this difference is at most .02.

4. 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 face 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 based on the estimated parameters of such 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 (ML), unweighted least squares (ULS), etc. 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 presents some amount of model misfit that can not be attributed to chance. In this case, an applied researcher can still compute a model-based reliability estimate based on her best fitting model. Such a model-based reliability estimator will be biased. The direction and magnitude of this bias will be unknown as 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 for instance 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.

Yet, 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 downwards. Hence it will 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 no be negligible.

In order to take into account the variability of sample alpha, an interval estimator should be used instead of a point estimate. In this paper, we have investigated the empirical performance of two confidence interval estimators for population alpha under different conditions of skewness and kurtosis, as well as sample size: 1) the confidence intervals proposed by van Zyl et al. (2000) which assumes that items are normally distributed (NT intervals), and 2) the confidence intervals proposed by Yuan et al. (2003) based on asymptotic distribution free assumptions (ADF intervals). Our results suggest that when the model assumptions underlying alpha are met, ADF intervals are to be 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.

Similar results for the coverage of alpha were found when we generated data where 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 where a congeneric measurement model holds, the ADF intervals contain true reliability when item kurtosis is larger than 4. If item kurtosis is smaller than 4 (i.e., close to being normally distributed), ADF intervals will also contain population reliability for samples smaller than 400. For larger samples, the ADF intervals will underestimate very slightly population reliability because the intervals have 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 kurtosis larger than 4, NT confidence interval misses the true reliability of the test because it does 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 only underestimated population reliability slightly, by a difference of between -.02 and -.01. This amount of misspecification was chosen to be typical in applications (McDonald, 1999). We feel that further simulation studies are needed to explore if the robustness of the interval estimators for coefficient alpha hold (i.e., if they contain population coefficient alpha) under alternative setups of model misspecification (such as bifactor models). Also, as the bias of population alpha increases, one should not expect confidence intervals for alpha to include the population reliability. Finally, further research should compare the symmetric confidence intervals employed here against asymmetric confidence intervals. This is because, as a reviewer pointed out, the upper limit of the symmetric confidence intervals for alpha may exceed the upper bound of one when sample alpha is near one.

5. 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. Because test and questionnaire items are usually ordered categorical variables, they may show considerable skewness and kurtosis, thereby violating the normality assumption (see Micceri, 1989). 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 of population alpha. Also, for sample sizes over 100 they are definitely preferred to NT confidence intervals if the items show skewness over 1 or kurtosis over 4. NT confidence intervals can be safely used within those bounds (i.e., 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 (2004) report that accurate confidence intervals can be obtained with sample sizes as small as 30.

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: 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., and Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(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., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89*, 792-808.

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.

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 and J.J. McArdle (Eds.). *Contemporary Psychometrics. A Festschrift to R.P. McDonald*. (pp. 341-376). Mahwah, NJ: Lawrence 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: Lawrence 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 non-normal 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 and 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 NT and ADF standard errors of sample alpha

In matrix notation, population alpha is $a = \frac{p}{p-1}\left(1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right)$, 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 $\hat{a} = \frac{p}{p-1}\left(1 - \frac{\mathrm{tr}(\mathbf{S})}{\mathbf{1}'\mathbf{S}\mathbf{1}}\right)$, where $\mathbf{S}$ denotes the sample covariance matrix.

Let $\mathbf{s} = \mathrm{vecs}(\mathbf{S})$ and let $\boldsymbol{\sigma} = \mathrm{vecs}(\boldsymbol{\Sigma})$, 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 $\hat{a}$ is a function of $\mathbf{s}$, asymptotically, $\hat{a}$ is normally distributed with mean α and variance

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

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

$$\frac{\partial a}{\partial s_{ij}} = \begin{cases} \frac{-p}{p-1}\frac{\mathbf{1}'\mathbf{S}\mathbf{1} - \mathrm{tr}(\mathbf{S})}{(\mathbf{1}'\mathbf{S}\mathbf{1})^2} & \text{if } i = j \\ \frac{2p}{p-1}\frac{\mathrm{tr}(\mathbf{S})}{(\mathbf{1}'\mathbf{S}\mathbf{1})^2} & \text{if } i \neq j \end{cases}. \tag{4}$$

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 shall use $\boldsymbol{\Gamma}_{NT}$ and $\boldsymbol{\Gamma}_{ADF}$ to distinguish them.

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

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

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

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

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

$$\hat{f}^2_{ADF} = \frac{1}{N}\hat{\mathbf{d}}'\hat{\mathbf{G}}_{ADF}\hat{\mathbf{d}} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\left(\hat{\mathbf{d}}'(\mathbf{s}_i - \mathbf{s})\right)^2. \quad (7)$$

To see this insert Equation (6) in Equation (3),

$$\hat{f}^2_{ADF} = \frac{1}{N}\hat{\mathbf{d}}'\hat{\mathbf{G}}_{ADF}\hat{\mathbf{d}} = \frac{1}{N}\hat{\mathbf{d}}'\left[\frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{s}_i - \mathbf{s})(\mathbf{s}_i - \mathbf{s})'\right]\hat{\mathbf{d}} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\hat{\mathbf{d}}'(\mathbf{s}_i - \mathbf{s})(\mathbf{s}_i - \mathbf{s})'\hat{\mathbf{d}},$$

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

Computation of population reliability for categorized normal variables

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

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

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

where

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

Data is generated as follows: First we generate multivariate normal data. In the first simulation we used $\mathbf{z}^*$: $N(\mathbf{0}, \mathbf{R})$, where $\mathbf{R} = r\mathbf{1}\mathbf{1}' + (1-r)\mathbf{I}$. 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 $t_{i_k} < z_i^* < t_{i_{k+1}}$, k_i = 0, ..., K - 1, where $t_{i_0} = -\infty$, $t_{i_K} = \infty$. 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 $\mathbf{R} = \mathbf{l}\mathbf{l}' + \mathbf{I} - diag(\mathbf{l}\mathbf{l}')$ where $\mathbf{l}' = (.3, .4, .5, .6, .7, .8, .9)$ 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,

$$\Pr(Y_i = k_i) = \int_{\tau_{ik}}^{\tau_{ik+1}} f_1(z_i^* : 0, 1)\, dz_i^*, \tag{11}$$

$$\Pr[(Y_i = k) \cap (Y_j = k')] = \int_{\tau_{ik}}^{\tau_{ik+1}} \int_{\tau_{jk'}}^{\tau_{jk'+1}} f_2(z_i^*, z_j^* : 0, 0, 1, 1, r_{ij})\, dz_i^* dz_j^*, \tag{12}$$

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

The population skewness and kurtosis reported in Figure 1 were computed using $\text{skewness} = \frac{m_3}{m_2^{3/2}}$, and $\text{kurtosis} = \frac{m_4}{m_2^2}$, where

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

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 inter-item correlation is simply

$$\bar{r} = \frac{1}{q} \sum_{i<j} \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \tag{14}$$

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

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

Consider now the case where $\mathbf{R} = \mathbf{l}\mathbf{l}' + \mathbf{I} - diag(\mathbf{l}\mathbf{l}')$. In this case, the covariances in $\mathbf{\Sigma}$ are not equal and as a result population alpha underestimates reliability. When

$\mathbf{R} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{I} - diag(\boldsymbol{\lambda}\boldsymbol{\lambda}')$ and the same thresholds are used for all items, the population covariance matrix $\boldsymbol{\Sigma}$ obtained using Equations (8) to (12) can be fitted exactly by a one-factor model, say $\mathbf{S} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. In this decomposition, $\boldsymbol{\Lambda}$, where $\boldsymbol{\lambda}$ are the factor loadings used to generate the data. Because $\boldsymbol{\Sigma}$ admits a one-factor model decomposition, population reliability is given by coefficient omega

$$w = \frac{\left(\sum_{i=1}^{p} \lambda_i\right)^2}{\left(\sum_{i=1}^{p} \lambda_i\right)^2 + \sum_{i=1}^{p} \psi_i^2} \tag{15}$$

where ψ_i^2 is the element of the diagonal matrix $\boldsymbol{\Psi}$ corresponding to the i^{th} item. As the model fits exactly in the population, any method can be used to estimate $\boldsymbol{\Lambda}$ 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 $\mathbf{R} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{I} - diag(\boldsymbol{\lambda}\boldsymbol{\lambda}')$ 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) to (12) we obtain the following population covariance matrix

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

This $\boldsymbol{\Sigma}$ admits a one-factor model decomposition 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 α is .677 and the population reliability is $\omega = .692$.

Footnotes

[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 importantly, 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

[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).

[5] The skewness and kurtosis of a standard normal distribution are 0 and 3, respectively.

[6] 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_i's is diagonal (Bentler, in press).

Table 1
Relationship between underlying polychoric correlation (ρ), the average population inter-item correlation ($\bar{r}$), and the population coefficient alpha (α)

Polychoric corr. (ρ)	Average population inter-item correlation ($\bar{r}$)				Population α		
	Levels	Mean	Min.	Max.	Mean	Min.	Max.
.16	Low	.11	.06	.15	.53	.25	.77
.36	Medium	.25	.16	.33	.74	.49	.91
.64	High	.48	.36	.59	.88	.73	.97

Table 2
Relative bias of NT standard errors

			skewness					
			2.667	1.960	.980	.408	0	
			kurtosis					
			8.111	4.843	2.800	1.167	2.500	3.878
$\bar{r}$	N	p						
low	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	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	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

Notes: p = number of variables, N = sample size, $\bar{r}$ = average population inter-item correlation

Table 3
Relative bias of ADF standard errors

skewness			2.667	1.960	.980	.408	0	
kurtosis			8.111	4.843	2.800	1.167	2.500	3.878
	$\bar{r}$	N	p					
low		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		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		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

Notes: p = number of variables, N = sample size, $\bar{r}$ = average population inter-item correlation

Table 4
Average coverage rates for NT and ADF 95% confidence intervals at each level of sample size and skewness when population coefficient alpha equals true reliability

Sample size	method	skewness				
		0	.41	.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). We have shaded the most accurate method for each combination of sample size and skewness.

Table 5
Average coverage of population coefficient alpha for NT and ADF 95% confidence intervals at each level of sample size and kurtosis when population coefficient alpha underestimates true reliability

Sample size	method	kurtosis					
		1.17	2.50	2.80	3.88	4.84	8.11
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). We have shaded the most accurate method for each combination of sample size and kurtosis.

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

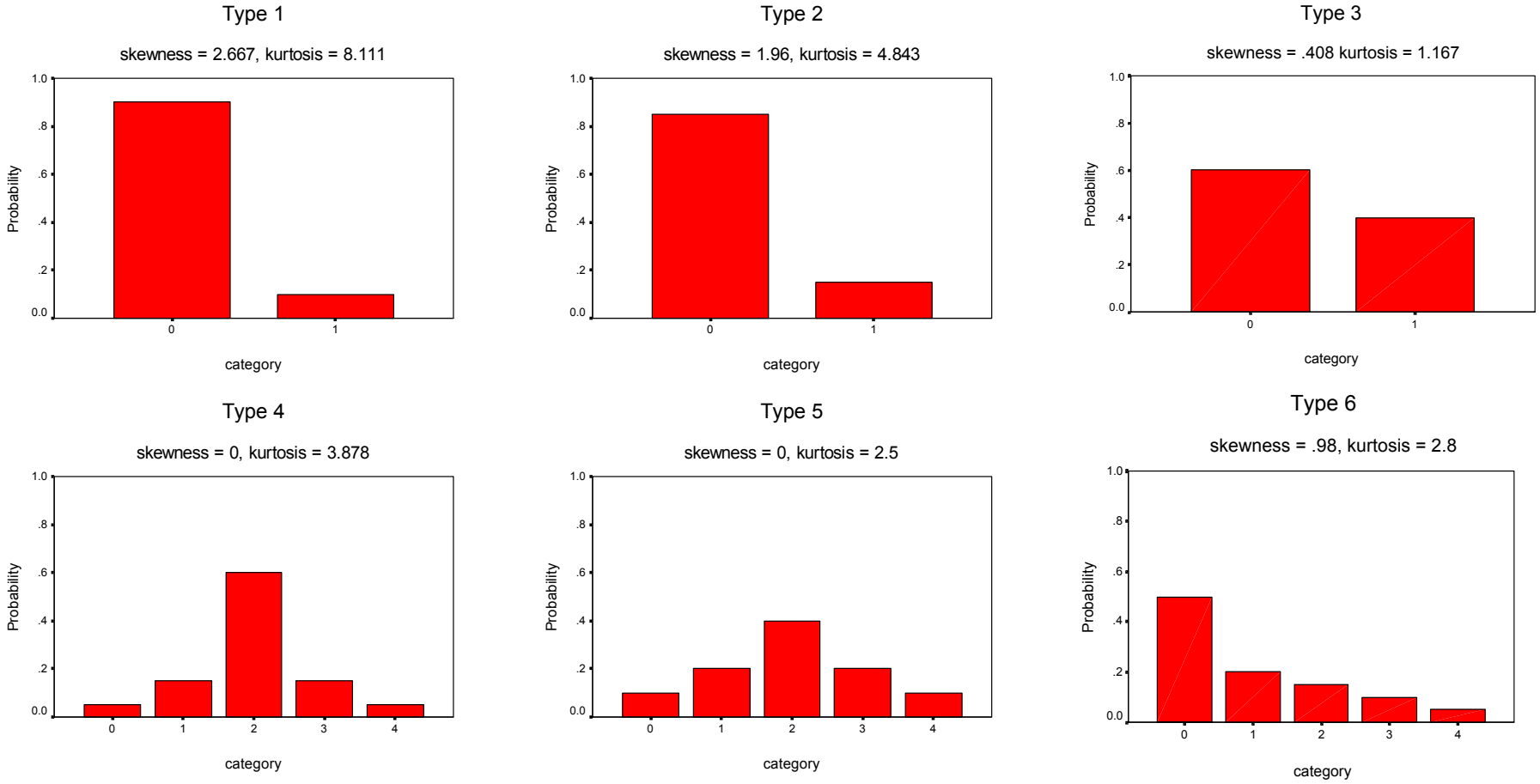


Figure 2
Relative bias of the coefficient alpha point estimates as a function of the true population alpha. A quadratic model has been fit to the points to model the relationship between relative bias and true alpha by sample size. Bias increases with decreasing sample size and decreasing population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

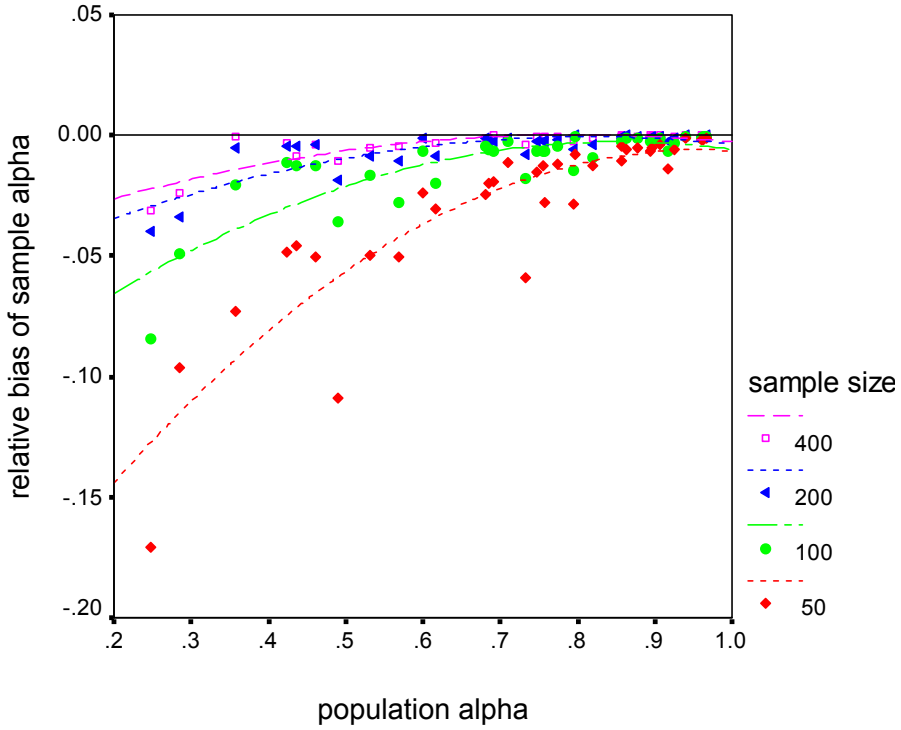
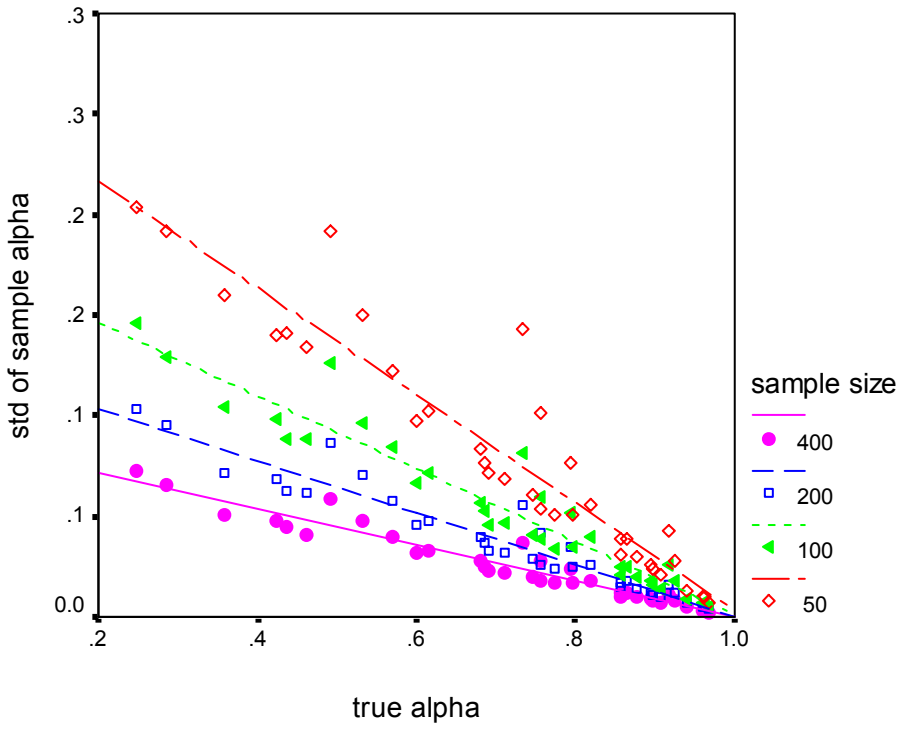


Figure 3
Variability of the coefficient alpha point estimates as a function of the true population coefficient alpha by sample size. Linear functions have been fit to the points to model the relationship between the standard deviation of sample coefficient alpha and the true population coefficient alpha.



:

Figure 4
Percentage of times (coverage) that 95% confidence intervals for alpha include population reliability as a function of skewness. Data has been generated so that population alpha equals reliability. Coverage rates should be close to nominal rates (95%). The accuracy of NT CIs worsens as average inter-item correlation gets smaller and skewness increases. The accuracy of ADF CIs worsens as sample size decreases and skewness increases. The accuracy of both CIs is somewhat similar for items with low skewness (< |1|); for higher skewness, ADF CIs are more accurate than NT CIs provided sample size > 100 observations.

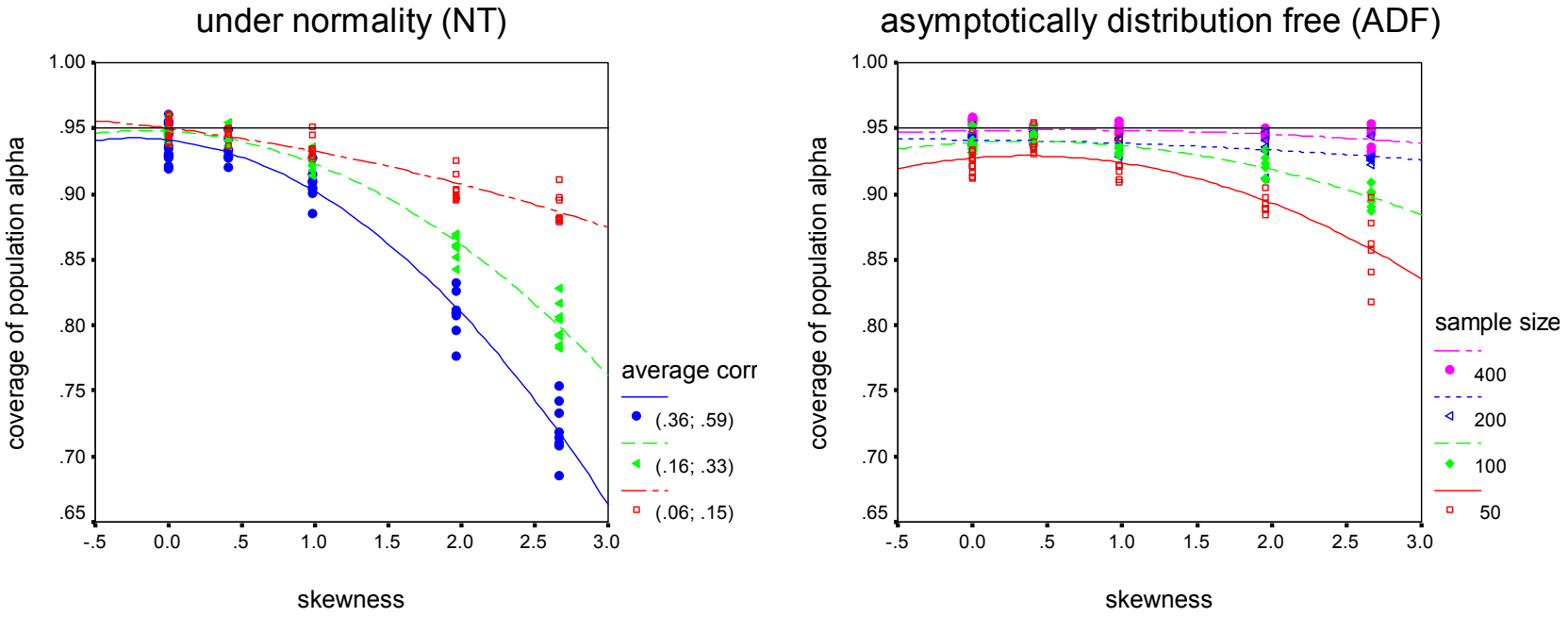


Figure 5
Absolute bias of the coefficient alpha point estimates as a function of the true population reliability when population alpha underestimates true reliability. A linear model has been fit to the points to model the relationship between bias and true reliability by sample size. Bias increases 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 1000. Therefore, the points for sample size 1000 are also the absolute bias of population alpha.

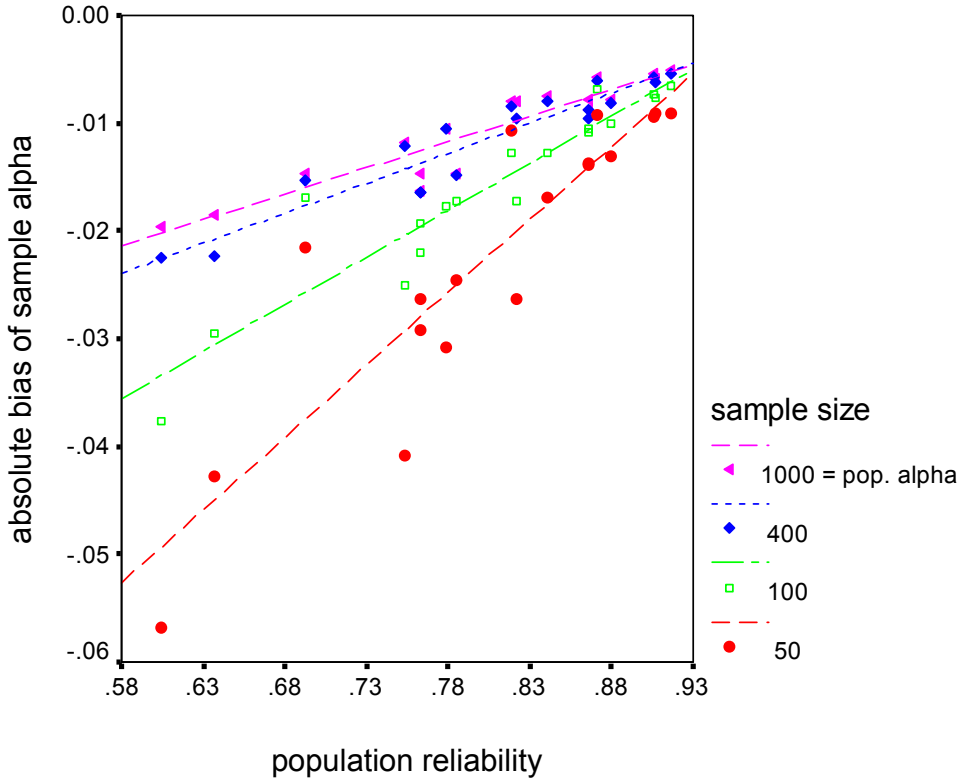


Figure 6
Scatterplot of mean standard errors (SEs) vs. standard deviation of sample coefficient alpha. The mean SEs should be equal to the standard deviation of sample coefficient alpha. This is indicated by the reference line in the diagonal of the graph. Points below the line indicate underestimation of the variability of sample coefficient alpha. NT SEs underestimate the variability of coefficient alpha when kurtosis > 4. 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.

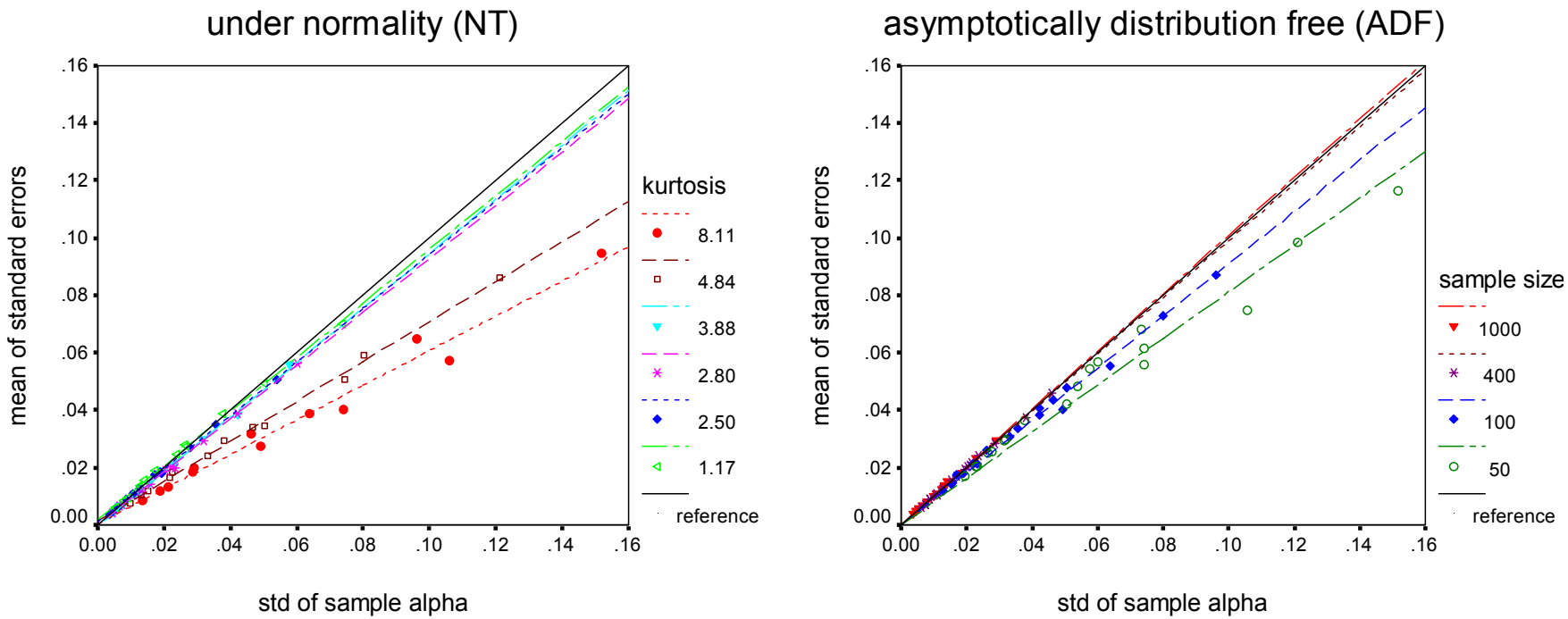


Figure 7
Percentage of times (coverage) that 95% confidence intervals (CIs) for coefficient alpha include population coefficient alpha as a function of kurtosis and sample size. Data has been generated according to a congeneric model. Coverage rates should be close to nominal rates (95%). The accuracy of both CIs is somewhat similar (and adequate) for items with low kurtosis (< 4). For items with higher kurtosis, ADF intervals are more accurate, particularly when sample size > 100 observations.

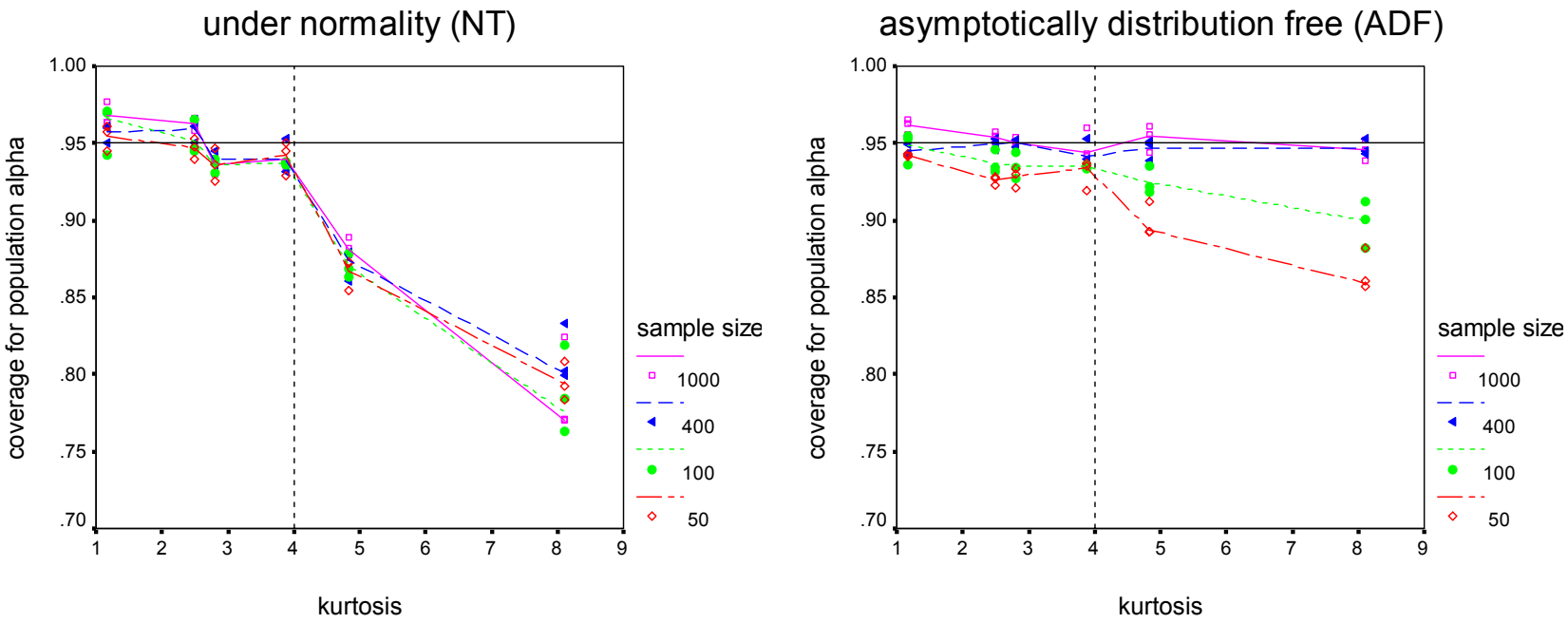
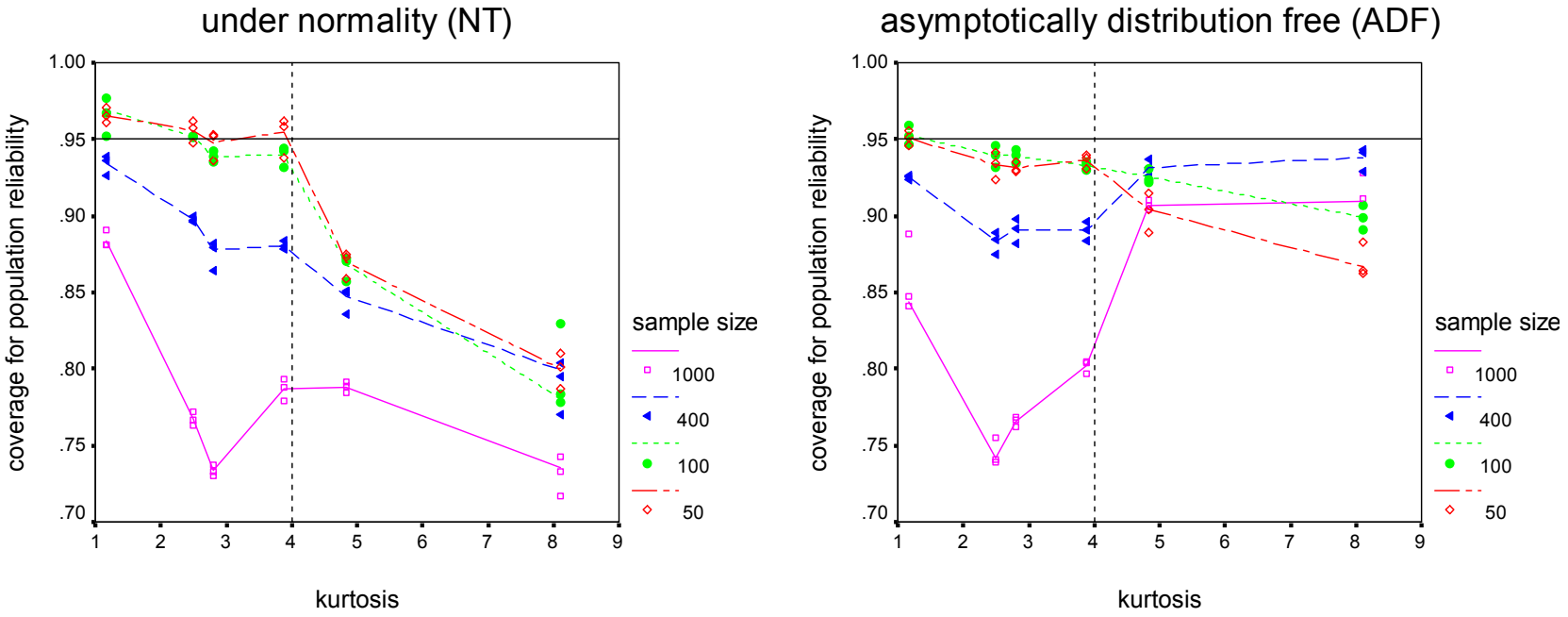


Figure 8
Percentage of times (coverage) that 95% confidence intervals (CIs) for coefficient alpha include population reliability as a function of kurtosis. Data has been generated according to a congeneric model, and population coefficient alpha is smaller than population reliability. As a result, coverage rates should be smaller than nominal rates (95%). The accuracy of both CIs is somewhat similar for items with low kurtosis (< 4). For items with higher kurtosis, ADF confidence intervals are definitively more accurate.



NOTAS

Depósito Legal: M-20073 I.S.S.N.: 1579-4873