MULTIVARIATE BEHAVIORAL RESEARCH, *40*(1), 115–148

Fit Indices Versus Test Statistics

Ke-Hai Yuan
University of Notre Dame

Model evaluation is one of the most important aspects of structural equation modeling (SEM). Many model fit indices have been developed. It is not an exaggeration to say that nearly every publication using the SEM methodology has reported at least one fit index. Most fit indices are defined through test statistics. Studies and interpretation of fit indices commonly assume that the test statistics follow either a central chi-square distribution or a noncentral chi-square distribution. Because few statistics in practice follow a chi-square distribution, we study properties of the commonly used fit indices when dropping the chi-square distribution assumptions. The study identifies two sensible statistics for evaluating fit indices involving degrees of freedom. We also propose linearly approximating the distribution of a fit index/statistic by a known distribution or the distribution of the same fit index/statistic under a set of different conditions. The conditions include the sample size, the distribution of the data as well as the base-statistic. Results indicate that, for commonly used fit indices evaluated at sensible statistics, both the slope and the intercept in the linear relationship change substantially when conditions change. A fit index that changes the least might be due to an artificial factor. Thus, the value of a fit index is not just a measure of model fit but also of other uncontrollable factors. A discussion with conclusions is given on how to properly use fit indices.

In social and behavioral sciences, interesting attributes such as *stress*, *social support*, *socio-economic status* cannot be observed directly. They are measured by multiple indicators that are subject to measurement errors. By segregating measurement errors from the true scores of attributes, structural equation modeling (SEM), especially its special case of covariance structure analysis, provides a methodology for modeling the latent variables directly. Although there are many

The research was supported by Grant DA01070 from the National Institute on Drug Abuse (Peter M. Bentler, Principal Investigator) and NSF Grant DMS-0437167. I am thankful to Peter M. Bentler and Robert C. MacCallum for their comments that have led the article to a significant improvement over the previous version.

Correspondence concerning this article should be addressed to Ke-Hai Yuan, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556. E-mail: kyuan@nd.edu

aspects to modeling, such as parameter estimation, model testing, and evaluating the size and significance of specific parameters, overall model evaluation is the most critical part in SEM. There is a huge body of literature on model evaluation that can be roughly classified into two categories: (a) overall-model-test statistics that judge whether a model fits the data exactly; (b) fit indices that evaluate the achievement of a model relative to a base model.

Fit indices and test statistics are often closely related. Actually, most interesting fit indices *F*s are defined through the so called chi-square statistics *T*s. The rationales behind these fit indices are often based on the properties of *T*. For example, under idealized conditions, *T* may approximately follow a central chi-square distribution under the null hypothesis and a noncentral chi-square distribution under an alternative hypothesis. In practice, data and model may not satisfy the idealized conditions and *T* may not follow (noncentral) chi-square distributions. Then, the rationales motivating these fit indices do not hold. There are a variety of studies on the performance of statistics; there also exist many studies on the performance of fit-indices. However, these two classes of studies are not well connected. For example, most of the studies on fit indices use just simulation with the normal theory based likelihood ratio statistic. There are also a few exceptions (e.g., Anderson, 1996; Hu & Bentler, 1998; Marsh, Hau, & Wen, 2004; Wang, Fan, & Willson, 1996; Zhang, 2004) but no study focused on the relationship between fit-indices and test statistics. This article will formally explore the relationship of the two. We are especially interested in conditions that affect the distributions of the commonly used fit indices. The purpose is to identify statistics that are most appropriate for calculating fit indices, to use fit indices more wisely and to evaluate models more scientifically.

We will use both analytical and empirical approaches to study various properties of fit indices. Our study will try to answer the following questions.

1. As point estimators, what are the population counterparts of the commonly used fit indices?
2. How the population counterparts related to model misspecifications?
3. Do we ever know the distribution of a fit index with real or even simulated data?
4. Are cutoff values such as 0.05 or 0.95 related to the distributions of the fit indices?
5. Are measures of model fit/misfit defined properly when the base-statistic does not follow a chi-square distribution? If not, can we have more sensible measures?
6. Whether confidence intervals for fit indices as printed in standard software cover the model fit/misfit with the desired probability?
7. How to reliably evaluate the power or sensitivity of fit indices?
8. Can we ever get an unbiased estimator of the population model fit/misfit as commonly defined?

Some of the questions have positive answers, some have negatives, and some may need further study. We will provide insightful discussions when definite answers are not available.

Although mean structure is an important part of SEM, in this article we will focus on covariance structure models due to their wide applications. In the next section, in order to facilitate the understanding of the development in later sections, we will give a brief review of the existing statistics and their properties, as well as fit indices and their rationales. We will discuss properties of fit indices under idealized conditions in the section entitled "Mean Values of Fit Indices Under Idealized Conditions." Of course, idealized conditions do not hold in practice. In the section entitled "Approximating the Distribution of *T* Using a Linear Transformation," we will introduce a linear transformation on the distribution of *T* to understand the difference between idealization and realization. With the help of the linear transformation, we will discuss the properties of fit indices in the section entitled "Properties of Fit Indices When *T* Does Not Follow a Chi-Square Distribution." In the section entitled "Matching Fit Indices with Statistics," we will match fit indices and statistics based on existing literature. An ad hoc correction to some existing statistics will also be given. The corrected statistics are definitionally more appropriate to define most fit indices. In the section entitled "Stability of Fit Indices When Conditions Change," we discuss the sensitivity of fit indices to changes in other conditions besides model misspecification. Power issues related to fit indices will be discussed in the section entitled "The Power of a Fit Index." In the Discussion section, we will discuss several critical issues related to measures of model fit and test statistics. We conclude the article by providing recommendations and pointing out remaining issues for further research.

SOME PROPERTIES OF STATISTICS AND RATIONALES FOR COMMONLY USED FIT INDICES

Let $\mathbf{x}$ represent the underlying p-variate population from which a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ with $N = n + 1$ is drawn. We will first review properties of three classes of statistics. Then we discuss the rationales behind several commonly used fit-indices. This section will provide basic background information for later sections, where we discuss connections between fit indices and the existing statistics.

Statistics

The first class of statistics includes the normal theory likelihood ratio statistic and its rescaled version; the second one involves asymptotically distribution free statistics. These two classes are based on modeling the sample covariance matrix $\mathbf{S}$ by a proposed model structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The third class is based on robust procedures which treat each observation $\mathbf{x}_i$ individually instead of using the summary statistic $\mathbf{S}$.

The most widely utilized test statistic in SEM is the classical likelihood ratio statistic T_{ML}, based on the normal distribution assumption of the data. When data are truly normally distributed and the model structure is correctly specified, T_{ML} approaches a chi-square distribution χ^2_{df} as the sample size N increases. Under certain conditions, this statistic asymptotically follows χ^2_{df} even when data are not normally distributed (Amemiya & Anderson, 1990; Browne & Shapiro, 1988; Kano, 1992; Mooijaart & Bentler, 1991; Satorra, 1992; Satorra & Bentler, 1990; Yuan & Bentler, 1999b). Such a property is commonly called asymptotic robustness. However, procedures do not exist for verifying the conditions for asymptotic robustness. It seems foolish to blindly trust that T_{ML} will asymptotically follow χ^2_{df} when data exhibit nonnormality. When data possess heavier tails than that of a multivariate normal distribution, the statistic T_{ML} is typically stochastically greater than the chi-square variate $\xi \sim \chi^2_{df}$. When the fourth-order moments of $\mathbf{x}$ are all finite, the statistic T_{ML} can be decomposed into a linear combination of independent $\xi_j \sim \chi^2_1$. That is,

$$T_{ML} = \sum_{j=1}^{df} \kappa_j \xi_j + o_p(1),$$

where the κ_js depend on the fourth-order moments of $\mathbf{x}$ as well as the model structure, and $o_p(1)$ is a term that approaches zero in probability as sample size N increases. When $\mathbf{x}$ follows elliptical or pseudo elliptical distributions with a common kurtosis κ, $\kappa_1 = \kappa_2 = \ldots = \kappa_{df} = \kappa$. Then (see Browne, 1984; Shapiro & Browne, 1987; Yuan & Bentler, 1999b)

$$T_{ML} = \kappa\xi + o_p(1). \tag{1}$$

When a consistent estimator $\hat{\kappa}$ of κ is available, one can divide T_{ML} by $\hat{\kappa}$ so that the resulting statistic still asymptotically approaches χ^2_{df}. Satorra and Bentler (1988) proposed $\hat{\kappa} = (\hat{\kappa}_1 + \ldots + \hat{\kappa}_{df})/df$ and the resulting statistic

$$T_R = \hat{\kappa}^{-1} T_{ML},$$

is often referred to as the Satorra-Bentler rescaled statistic. Like T_{ML}, T_R can also follow a chi-square distribution when certain asymptotic robustness conditions are satisfied (Kano, 1992; Yuan & Bentler, 1999b). Simulation studies indicate that T_R performs quite robustly under a variety of conditions (Chou, Bentler, & Satorra, 1991; Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992). However, data

In EQS, when specifying Method=ML, Robust;
T_R is the "SATORRA-BENTLER SCALED CHI-SQUARE"

generation in some of the studies is not clearly stated and may satisfy the asymptotic robustness condition for T_R (see Yuan & Bentler, 1999b). In general, T_R does not approach a chi-square distribution. Instead, it approaches a variate η with $E(\eta) = df$. It is likely that the distribution shape of η is far from that of a chi-square. In such cases, T_R will not behave like a chi-square. It can also lead to inappropriate conclusions when referring T_R to a chi-square distribution.

With typical nonnormal data in the social and behavioral sciences (Micceri, 1989), the ideal is to have a statistic that approximately follows a chi-square distribution regardless of the underlying distribution of the data. One of the original proposals in this direction was made by Browne (1984). His statistic is commonly called the asymptotically distribution free (ADF) statistic T_{ADF}, due to its asymptotically following χ^2_{df} as long as **x** has finite fourth-order moments. The ADF property is desirable. However, the distribution of T_{ADF} can be far fr
for typical sample sizes encountered in practice (Hu et al., 1992). S
mean and variance of T_{ADF} are much greater than those of χ^2_{df}.
specified models are rejected if using $T_{ADF} \sim \chi^2_{df}$. In an effort to fin
perform better in rejection rate with smaller *N*s, Yuan and Bentle
posed a corrected statistic

In EQS, when specifying

METHOD=AGLS;

T_CADF is the

"YUAN-BENTLER

CORRECTED AGLS TES

STATISTIC"

$$T_{CADF} = T_{ADF}/(1 + T_{ADF}/n).$$

In EQS, when specifying

METHOD=AGLS;

T_ADF is the

"CHI-SQUARE"

statistic

ADF asymptotically follows χ^2_{df} as long as **x** has finite fourth-order
s, it is asymptotically distribution free. The mean of T_{CADF} approx-
ls *df* for all sample sizes across various distributions (Yuan &
b). However, at small sample sizes T_{CADF} over-corrects the behav-
due to its rejection rate with correct models being smaller than the nominal level. Furthermore, T_{CADF} also carries the drawback of the ADF estimation method with nonconvergences at smaller sample sizes. In addition to T_{ADF}, Browne (1984) also proposed a residual-based ADF statistic T_{RADF} in which the estimator just needs to be consistent. However, T_{RADF} behaves almost the same as T_{ADF}, rejecting most correct models at smaller sample sizes. Parallel to T_{CADF}, Yuan and Bentler (1998b) proposed T_{CRADF} whose performance is almost the same as T_{CADF}, with its empirical mean approximately equal to *df* and under-rejecting the correct model for small sample sizes (Bentler & Yuan, 1999; Yuan & Bentler, 1998b).

The third class of statistics is obtained from robust procedures. It is well-known that the sample covariance matrix **S** is very sensitive to influential observations and is biased for $\boldsymbol{\Sigma}_0 = \mathrm{Cov}(\mathbf{x})$ when data contain outliers (see Yuan & Bentler, 1998c). In such a situation, removing these outliers followed by modeling **S** will lead to proper analysis of the covariance structure model. However, in a given data set, determining which cases are outliers may be difficult. The heavy tails, often indicated by larger marginal kurtoses or Mardia's (1970) multivariate kurtosis, might

In EQS, when specifying

METHOD=ML, Robust;

T_CRADF is the

"YUAN-BENTLER RESIDUAL-BASED

TEST STATISTIC"

In EQS, when specifying

METHOD=ML, Robust;

T_RADF is the

"RESIDUAL-BASED TEST

be due to the heavy tails in the distribution of $\mathbf{x}$. When a data set possesses moderate heavy tails, $\mathbf{S}$ will be an inefficient estimator of its population counterpart $\mathbf{\Sigma}_0$ (Tyler, 1983). When the heavy tails are severe, the population 4th-order moments do not exist. Then modeling $\mathbf{S}$ and referring T_{ML}, T_R, T_{ADF}, T_{RADF}, T_{CADF} or T_{CRADF} to χ^2_{df} is meaningless. In such cases, it is better to model a robust covariance matrix $\mathbf{S}_r$. Since the empirical experimentations by Huba and Harlow (1987), various technical procedures have been developed for modeling robust covariance matrices (Yuan & Bentler, 1998a, 1998c, 2000; Yuan, Chan, & Bentler, 2000). They differ in how to control the weight attached to each case (see Yuan & Bentler, 1998a, 1998c; Yuan & Hayashi, 2003). Three types of statistics are proposed related to modeling $\mathbf{S}_r$. One is to model $\mathbf{S}_r$ using the rescaled statistic T_R (Yuan & Bentler, 1998c); another is to just use T_{ML} treating $\mathbf{S}_r$ as $\mathbf{S}$ (Yuan et al., 2000; Yuan & Hayashi, 2003); the third is an ADF type statistic using the inverse of a robust version of the fourth-order moment matrix as the weight matrix (Yuan & Bentler, 1998a, 2000). The most preferred is T_{ML} coupling with a proper weight-control scheme (Yuan, Bentler, & Chan, 2004; Yuan & Hayashi, 2003). A brief outline as well as an introduction to a SAS IML program for robustifying a sample is provided in the appendix.

Note that many studies on T_{ML}, T_R, T_{ADF} and T_{CADF} in the first two classes only reported the rejection rates and their empirical means and standard deviations. Few studies examined their overall distributions (Curran, Bollen, Paxton, Kirby, & Chen, 2002; Yuan & Bentler, 1998b). By controlling the weights in some robust procedures, Yuan and Hayashi (2003) found that the distribution of T_{ML} applied to $\mathbf{S}_r$ can be well described by χ^2_{df}. But the distributions of T_R and T_{ADF} could not be well approximated by χ^2_{df} even when data were fairly normally distributed.

In additional to the above three classes of statistics, other statistics such as the normal theory generalized least squares statistic T_{GLS} and the heterogeneous kurtosis statistic T_{HK} (Kano, Berkane, & Bentler, 1990) are also available in standard software (Bentler, 1995). These are related to the normal theory based statistics of the first class. Yuan and Bentler (1998b, 1999a) also proposed two F-statistics T_F and T_{FR} based on T_{ADF} and T_{RADF}, respectively. They are asymptotically distribution free and thus belong to the second class. We will not specifically deal with these statistics in this article although preliminary work indicates that the latter especially might be very valuable in testing the significance of a model.

In EQS, when specifying `Crobust=x1, x2;`
T_ML is `the "CHI-SQUARE". This statistic need to adjust the two numbers x1 and x2. The other statistics in the output do not need to have optimal x1 and x2.`

Fit Indices

Fit indices can be classified into two categories, those that are defined explicitly through the overall test statistic T versus those that are not involving the statistic T directly. For example, the standardized root-mean-square residual (SRMR) is not defined through T but through residuals at the convergence of a model fitting procedure. Actually, when a model is approximately correct, all the overall statistics can be approximately regarded as a sum of weighted squares of residuals (Shapiro, 1985), and the weights are optimized according to the chosen estimation procedure. Thus, those defined through a T better utilize the residuals than SRMR. Fit indices that are explicitly defined through T also fall into three types (e.g., Hu & Bentler, 1998; Marsh, Balla, & McDonald, 1988): (a) Indices that do not need the involved statistic to follow any known distribution; (b) indices that assume the statistic T to satisfy $E(T|H_0) = df$, which is automatically satisfied when $T \sim \chi^2_{df}$; and (c) indices that assume $(T \mid H_0) \sim \chi^2_{df}$ and $(T \mid H_1) \sim \chi^2_{df}(\delta)$, where H_0 represents the null hypothesis, H_1 represents the alternative hypothesis and δ is the noncentrality parameter (NCP). Many fit indices are available in standard software (e.g., EQS, LISREL, SAS CALIS). The commonly reported ones are CFI, GFI, NFI, NNFI, and RMSEA, according to the review of McDonald and Ho (2002). We will mainly discuss the commonly used ones although our analysis and discussion also equally apply to other fit indices (e.g., Bollen, 1986, 1989; Hoelter, 1983; Steiger, 1989).

Let T_M and T_I be the chosen statistic T evaluated at the substantive model $\mathbf{\Sigma}_M = \mathbf{\Sigma}(\theta)$ and the independence model $\mathbf{\Sigma}_I = \text{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp})$, respectively. The normed fit index (Bentler & Bonett, 1980) is defined as

$$\text{NFI} = 1 - \frac{T_M}{T_I}.$$

As discussed by Hu and Bentler (1998), the T in NFI does not necessarily need to follow a particular distribution. Another widely used fit index is the nonnormed fit index (Bentler & Bonett, 1980; Tucker & Lewis, 1973)

$$\text{NNFI} = 1 - \frac{(T_M / df_M - 1)}{(T_I / df_I - 1)},$$

where df_M and df_I are the degrees of freedom in model $\mathbf{\Sigma}_M$ and $\mathbf{\Sigma}_I$, respectively. The difference between NFI and NNFI is that T is replaced by $T/df - 1$, which is motivated by $(T \mid H_0) \sim \chi^2_{df}$ (see Bentler, 1995). When $\mathbf{\Sigma}_M$ is correctly specified, $E(T_M) = df_M$ and NNFI ≈ 1. A more popular fit index is the comparative fit index (Bentler, 1990)

$$\text{CFI} = 1 - \frac{\max\left[(T_M - df_M), 0\right]}{\max\left[(T_I - df_I), (T_M - df_M)\right]}.$$

CFI is always within the range of [0,1]. Essentially equivalent to CFI but not necessarily within the range of [0,1] is the relative noncentrality index

$$\text{RNI} = 1 - \frac{T_M - df_M}{T_I - df_I},$$

independently proposed by Bentler (1990) and McDonald and Marsh (1990). Both CFI and RNI are motivated by $(T \mid H_1) \sim \chi^2_{df}(\delta)$ so that they measure the reduction of the NCP by $\mathbf{\Sigma}_M$ relative to that by $\mathbf{\Sigma}_I$ (see Bentler, 1990, 1995). Another commonly used index is the goodness of fit index, its general form was given in Bentler (1983, Equation 3.5) as

$$\text{GFI} = 1 - \frac{\mathbf{e}'\mathbf{We}}{\mathbf{s}'\mathbf{Ws}},$$

where $\mathbf{e} = \mathbf{s} - \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})$ with $\mathbf{s}$ and $\boldsymbol{\sigma}$ being vectors containing the nonduplicated elements of $\mathbf{S}$ and $\mathbf{\Sigma}$, respectively; the weight matrix $\mathbf{W}$ is subject to the estimation method. We can rewrite

$$\text{GFI} = 1 - \frac{T_M}{T_0},$$

where $T_M = n\mathbf{e}'\mathbf{We}$ is the normal theory based iteratively reweighted least squares (IRLS) statistic T_{IRLS} when the normal theory based $\mathbf{W}$ is used (see Bentler, 1995, p. 216) and $T_0 = \mathbf{s}'\mathbf{Ws}$ can also be regarded as a T_{IRLS} for testing $\mathbf{\Sigma}_0 = \mathbf{0}$. When the ADF weight matrix $\mathbf{W}$ is used, T_M and T_0 share similar meanings (Tanaka & Huba, 1985).

The last fit index we will discuss is the root-mean-square error of approximation (Browne & Cudeck, 1993; Steiger & Lind, 1980)

$$\text{RMSEA} = \sqrt{\max\left[(T_M - df_M)/(n \times df_M), 0\right]}.$$

The implicit assumption in RMSEA is that $(T \mid H_1) \sim \chi^2_{df}(\delta)$ and δ equals n times the "model misfit" so that it estimates the model misfit per degree of freedom. Notice that RMSEA only involves T_M. The literature on fit indices classifies it as an absolute fit index while those involving T_I are relative fit indices. There are also many other fit indices (see Hu & Bentler, 1998) in addition to the above six, but they are less frequently used.

MEAN VALUES OF FIT INDICES UNDER IDEALIZED CONDITIONS

A fit index F may aim to estimate a quantity measuring the model fit/misfit at the population level. F is always an unbiased estimate of $E(F)$ when the expectation exists. We need to consider how $E(F)$ is related to model fit/misfit. Analytically, this can only be studied when the involved statistics follow known distributions. Suppose

$$T_M \sim \chi^2_{df_M}(\delta_M) \text{ and } T_I \sim \chi^2_{df_I}(\delta_I). \tag{2}$$

Then

$$E(T_M) = df_M + \delta_M \text{ and } E(T_I) = df_I + \delta_I, \tag{3}$$

and

$$\hat{\delta}_M = T_M - df_M \text{ and } \hat{\delta}_I = T_I - df_I,$$

are the unbiased estimators of δ_M and δ_I, respectively. Most fit indices are functions of $\hat{\delta}_M$ and $\hat{\delta}_I$. It is interesting to see how the mean of a fit index is related to δ_M and δ_I. Suppose the fit index F is given by

$$F = g(\hat{\delta}_M, \hat{\delta}_I).$$

Because g is typically a nonlinear function, we might use a Taylor expansion to facilitate the study. Denote $\ddot{g}_{ij}$ as the second derivatives of g with respect to the parameters indicated in the subscripts. We have

$$\begin{aligned} E\left[g(\hat{\delta}_M, \hat{\delta}_I)\right] = {} & g(\delta_M, \delta_I) + E\left[\ddot{g}_{11}(\bar{\delta}_M, \bar{\delta}_I)(\hat{\delta}_M - \delta_M)^2/2\right] \\ & + E\left[\ddot{g}_{22}(\bar{\delta}_M, \bar{\delta}_I)(\hat{\delta}_I - \delta_I)^2/2\right] + E\left[\ddot{g}_{12}(\bar{\delta}_M, \bar{\delta}_I)(\hat{\delta}_M - \delta_M)(\hat{\delta}_I - \delta_I)\right], \end{aligned} \tag{4}$$

where $\bar{\delta}_M$ is a number between δ_M and $\hat{\delta}_M$, and $\bar{\delta}_I$ is a number between δ_I and $\hat{\delta}_I$. Most incremental fit indices can be expressed as $g(t_1, t_2) = 1 - (t_1 - c_1)/(t_2 - c_2)$, where c_1 and c_2 are constants. Then $\ddot{g}_{11} = 0, \ddot{g}_{22} = -2(t_1 - c_1)/(t_2 - c_2)^3$ can be either negative or positive, $\ddot{g}_{12} = 1/(t_2 - c_2)^2 > 0$. It follows from Equation 4 that $E\left[g(\hat{\delta}_M, \hat{\delta}_I)\right]$ can be greater than $g(\delta_M, \delta_I)$ or smaller than $g(\delta_M, \delta_I)$, depending on the last two terms in Equation 4. It is unlikely for $E\left[g\left(\hat{\delta}_M,\hat{\delta}_I\right)\right]$ to equal $g(\delta_M, \delta_I)$.

For indices that are only based on $\hat{\delta}_M$, there exists

$$E\left[g\left(\hat{\delta}_M\right)\right]=g\left(\delta_M\right)+E\left[\ddot{g}\left(\bar{\delta}_M\right)\left(\hat{\delta}_M-\delta_M\right)^2/2\right]. \tag{5}$$

When g is a strictly concave function, $\ddot{g}\left(\bar{\delta}_M\right)<0$ and it follows from Equation 5 that $E\left[g\left(\hat{\delta}_M\right)\right]<g\left(\delta_M\right)$. When g is a convex function, the opposite is true. For the index Mc proposed in McDonald (1989) for example, we have $g(t) = \exp[-t/(2n)]$, $\ddot{g}(t) = 0.25n^{-2}\exp[-t/(2n)]$, and

$$E(\text{Mc}) > \exp[-\delta_M/(2n)].$$

The difference between $E(\text{Mc})$ and $\exp[-\delta_M/(2n)]$ is proportional to the variance of T_M. The widely used RMSEA is a composite function $g\left(\hat{\delta}_M\right)=g_1\left[g_2\left(\hat{\delta}_M\right)\right]$, where $g_1(t) = \sqrt{t}$ is strictly concave with $\ddot{g}(t) = -0.25t^{-3/2} < 0$; and $g_2(t) = \max[t/(n \times df_M),0]$ is convex but not strictly. For bad models, T_M may never be below df_M so that $g_2(t)$ becomes an identity function, then

$$E(\text{RMSEA})<\sqrt{\delta_M/\left(n\times df_M\right)}.$$

For correct or nearly correct models, $E(\text{RMSEA})$ might be greater than $\sqrt{\delta_M/\left(n\times df_M\right)}$ due to the convexity of $g_2(t)$ (see Raykov, 2000).

Note that the assumptions in Equation 2 are unlikely to be true in practice. The above analysis implies that we do not know the mean values of the commonly used fit indices even under idealized conditions. Cutoff values such as 0.05 or 0.95 have little to do with the mean value of the fit indices or the NCPs of the idealized chi-squares. Of course, the chosen overall test statistic T maybe far from following a chi-square distribution. When a noncentral chi-square does not exist, then NCP is irrelevant. The rest of the article will discuss different aspects of F and T.

APPROXIMATING THE DISTRIBUTION OF *T* USING A LINEAR TRANSFORMATION

As reviewed in the previous section, Equations 2 and 3 are widely used in the context of fit indices (Bentler, 1990; Browne & Cudeck, 1993; Kim, 2003; MacCallum, Browne, & Sugawara, 1996; McDonald, 1989; McDonald & Marsh, 1990; Ogasawara, 2001; Steiger & Lind, 1980). Results in Equation 2 might be justified by the asymptotic result (see e.g., Bentler & Dijkstra, 1985; Browne, 1984; Satorra, 1989; Shapiro, 1983)

$$T = nD\left[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right] \xrightarrow{\mathscr{L}} \chi^2_{df}(\delta), \tag{6}$$

where $\hat{\boldsymbol{\theta}}$ is the parameter estimator, $\delta = n\tau$ and

$$\tau = D\left[\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}(\boldsymbol{\theta}^*)\right] = \min_{\boldsymbol{\theta}} D\left[\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right], \tag{7}$$

is the "model misfit" measured by a discrepancy function $D(\cdot,\cdot)$. Actually, Equation 6 is not true[1] with a given $\boldsymbol{\Sigma}_0$ (see Olsson, Foss, & Breivik, 2004; Yuan & Bentler, in press; Yuan & Chan, in press). The proof of Equation 6 needs the assumption

$$E(\mathbf{S}) = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{0,n} = \boldsymbol{\Sigma}(\boldsymbol{\theta}^*) + \frac{\boldsymbol{\Delta}}{\sqrt{n}}, \tag{8}$$

where $\boldsymbol{\Delta}$ is a constant matrix. Note that $\mathbf{S}$ contains sampling error ($\mathbf{S} - \boldsymbol{\Sigma}_0$) whose magnitude is of order $1/\sqrt{n}$. One might understand Equation 8 by thinking that the systematic error $\boldsymbol{\Delta}/\sqrt{n}$ approximately equals the sampling error. However, $\boldsymbol{\Sigma}_0$ is the population covariance matrix that should not depend on n while Equation 8 implies that the amount of misspecification in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ decreases as n increases. Such a condition has an obvious fault but it provides the mathematical convenience to allow one to show that Equation 6 holds with δ being a constant that does not depend on n. With fixed $\boldsymbol{\Sigma}_0$ and n, what the asymptotic statistical theory really tells us is that, when $\boldsymbol{\Sigma}_0$ is sufficiently close to $\boldsymbol{\Sigma}(\boldsymbol{\theta}\ )$, the distribution of T can be approximated by $\chi^2_{df}(\delta)$. But it says nothing about the goodness of the approximation.

Let's look at the T_{ML} with normally distributed data generated by the following confirmatory factor model (CFM)

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e} \quad \text{with} \quad \mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi},$$
$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\lambda} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\lambda} \end{pmatrix}, \qquad \boldsymbol{\Phi} = \begin{pmatrix} 1.0 & .30 & .40 \\ .30 & 1.0 & .50 \\ .40 & .50 & 1.0 \end{pmatrix}, \tag{9}$$

where $\boldsymbol{\lambda}' = (.70,.70,.75,.80,.80)$ and $\boldsymbol{\Psi}$ is a diagonal matrix chosen so that $\boldsymbol{\Sigma}_0$ is a correlation matrix. Suppose one fits the sample covariance matrix $\mathbf{S}$ by the independence model $\boldsymbol{\Sigma}_I$ using the normal theory maximum likelihood (ML). Then $\tau = 6.859$. With 500 replications, the top panel of Figure1 plots the quantiles of the sta-

[1] Curran et al.'s (2002) conclusion that Equation 6 approximately holds in some of their simulation conditions is based on 500 or few converged samples out of 650 replications. A different conclusion might be reached when all samples converge.

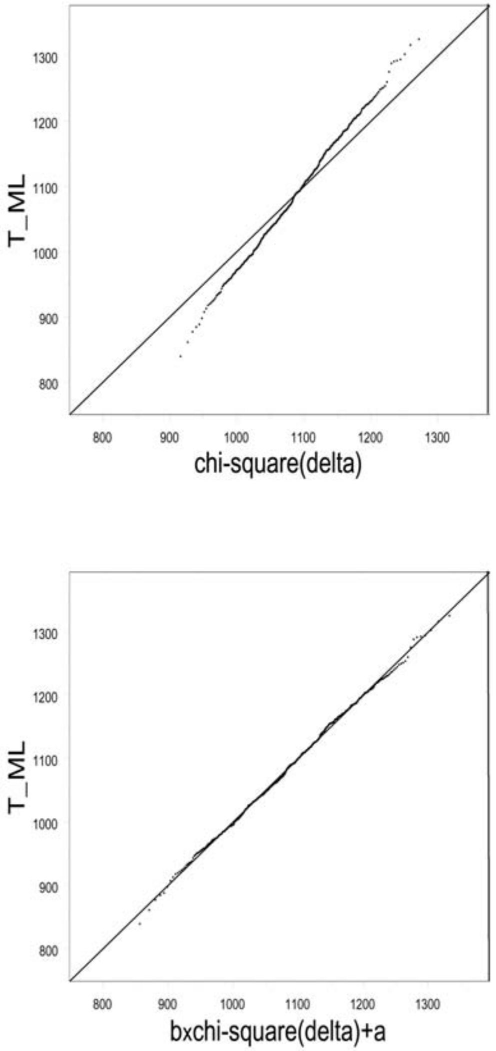

FIGURE 1 QQ plots of T_{ML} versus $\chi^2_{105}(\delta)$ and T_{ML} versus $b\chi^2_{105}(\delta)$ + a with $a = -368.000$ and $b = 1.337$, $N = 150$ and 500 replications.

tistic T_{ML} against the quantiles of $\chi^2_{105}(\delta)$ (QQ plot) at $N = 150$. It is obvious that the upper tail of the distribution of T_{ML} is much longer than that of $\chi^2_{105}(\delta)$ and the lower tail is much shorter than that of $\chi^2_{105}(\delta)$.

Motivated by Equations 1 and 3, we might describe the distribution of T by

$$T = b\chi^2_{df}(\delta) + a, \tag{10}$$

where the intercept a and slope b can be estimated by regression of the quantiles of T against those of $\chi^2_{df}(\delta)$. The bottom panel of Figure 1 plots the quantiles of T_{ML} against those of $b\chi^2_{105}(\delta)+a$ with $a=-368.000$ and $b=1.337$ being estimated based on the 500 replications. It is obvious that Equation 10 describes the distribution of T_{ML} very well. Actually, judged by a visual inspection, the two sets of quantiles in the bottom panel of Figure 1 match as close as a QQ plot (not presented here) of a sample of simulated chi-squares against its population quantiles. When each of the samples is fitted by the 3-factor model as that generated the sample in Equation 9, then H_0 holds with $df=87$. Figure 2 contains the QQ plots of T_{ML} for the correct model with 500 simulated normal samples of size $N=150$. It is obvious that T_{ML} is stochastically greater than χ^2_{87} and is described by $b\chi^2_{87}+a$ very well. Our limited evidence also implies that $b\chi^2_{df}(\delta)+a$ can also describe the distributions of the other statistics quite well. Notice that a and b depend on the involved statistic, the sample size, the distribution of the data, as well as the model itself.

Transformation Equation 10 can be extended to

$$(T_2|C_2) = b(T_1|C_1) + a, \qquad (11)$$

where T_2 and T_1 can be two different statistics or the same statistic evaluated at two different conditions C_1 and C_2. The conditions include the sample size, the distribution of $\mathbf{x}$, and the model specification. Equation 11 is quite useful in describing the distributional change of T when changing conditions. Of course, we would wish to have $b = 1$, $a = 0$ regardless of T_1, C_1 and T_2, C_2. However, a and b change with the sample size and the distribution of $\mathbf{x}$ even when T_1 and T_2 are the same statistic. To provide more information on a and b under different conditions, we let T_1 and T_2 be the same statistic while C_1 and C_2 have different sample sizes or distributions of $\mathbf{x}$. Using the same CFM as in Equation 9, the distribution of $\mathbf{x}$ in C_2 changes from elliptically distributed data to lognormal factors and errors, as further clarified in Table 1. The parameters are estimated by the ML. The statistics presented in Table 2 are T_{ML}, T_R, T_{RADF}, T_{CRADF}, as reviewed in the section entitled "Some Properties of Statistics and Rationales for Commonly Used Fit Indices." The statistics T_{AML} and T_{AR} are ad hoc corrections respectively to T_{ML} and T_R, which will be discussed further in a later section. The T_{IRLS} here is for the study of the properties of GFI. The T_0 corresponding to T_{RADF} is also for the properties of GFI, where $\mathbf{W}$ is just the ADF weight matrix due to the model $\boldsymbol{\sigma} = \mathbf{0}$ having all its derivatives equal to zero (see Browne, 1984). With 500 replications, Table 2 contains the a and b at $N = 150$ while the distribution of $\mathbf{x}$ changes; Table 3 contains the parallel results corresponding to $N = 500$. It is obvious that a and b change for all the statistics, especially at the misspecified independence model. A larger sample size does not help much even for the ADF type statistics. Table 4 contains similar result when T_1 and T_2 are the same statistic and C_1 and C_2 have the same distribution but different sample sizes. Compared to the distributional variation condition, sample size has a rela-

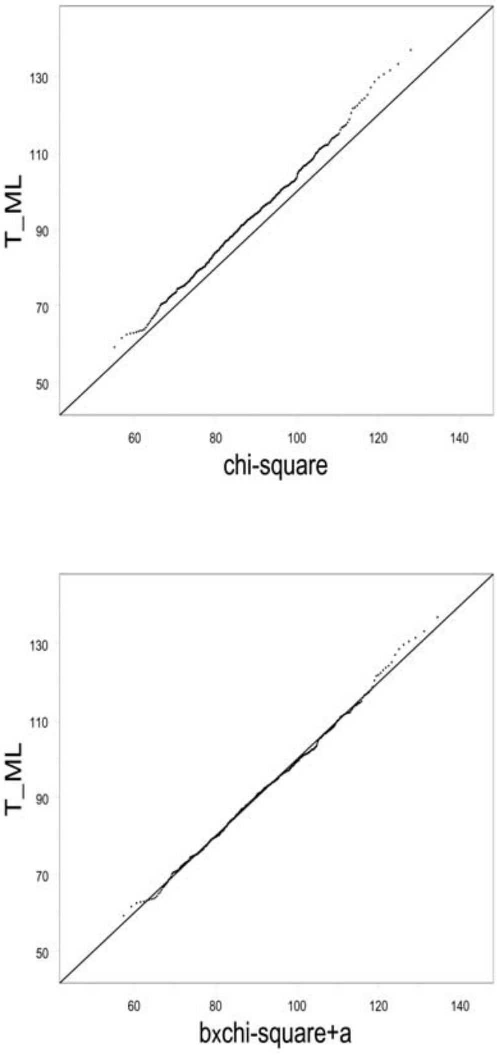

FIGURE 2 QQ plots of T_{ML} versus χ^2_{87} and T_{ML} versus $b\chi^2_{87} + a$ with $a = –0.555$ and $b = 1.054$, $N = 150$ and 500 replications.

tively smaller effect on *b*. But there is still a substantial difference among the *a*s, especially when the model is misspecified.

Notice that, for all the simulation conditions in Table 1, T_R asymptotically follows χ^2_{87} for the correctly specified model. Actually, Equation 1 holds with $\kappa = 3$ for all the non-normal distribution conditions in Table 1. A greater κ will have a

TABLE 1
Distribution Conditions of **x**

Normal **x**	$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e}, \mathbf{f} \sim N(\mathbf{0},\boldsymbol{\Phi}), \mathbf{e} \sim N(\mathbf{0},\boldsymbol{\Psi})$
Elliptical **x**	$\mathbf{x} = \boldsymbol{\mu} + (\boldsymbol{\Lambda}\mathbf{f} + \mathbf{e})/r, \mathbf{f} \sim N(\mathbf{0},\boldsymbol{\Phi}), \mathbf{e} \sim N(\mathbf{0},\boldsymbol{\Psi}), r \sim \sqrt{\chi_5^2/3}$
Skew **f** & normal **e**	$\mathbf{x} = \boldsymbol{\mu} + (\boldsymbol{\Lambda}\mathbf{f} + \mathbf{e})/r, \mathbf{f} \sim Lognormal(\mathbf{0},\boldsymbol{\Psi}), \mathbf{e} \sim N(\mathbf{0},\boldsymbol{\Phi}), r \sim \sqrt{\chi_5^2/3}$
Skew **f** & **e**	$\mathbf{x} = \boldsymbol{\mu} + (\boldsymbol{\Lambda}\mathbf{f} + \mathbf{e})/r, \mathbf{f} \sim Lognormal(\mathbf{0},\boldsymbol{\Phi}), \mathbf{e} \sim Lognormal(\mathbf{0},\boldsymbol{\Psi}), r \sim \sqrt{\chi_5^2/3}$

Note. **e**, **f,** and r are independent; $\mathbf{f} \sim Lognormal(\mathbf{0}, \boldsymbol{\Phi})$ is obtained by $\mathbf{f} = \boldsymbol{\Phi}^{1/2}\boldsymbol{\xi}$, where $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)'$ and ξ_i is to standardize $\exp(z)$ with $z \sim N(0, 1)$; and similarly for $\mathbf{e} \sim Lognormal(\mathbf{0}, \boldsymbol{\Psi})$.

TABLE 2
Intercept a and Slope b in $(T_2|C_2) = b(T_1|C_1) + a$, Where T_2 and T_1 Are the Same Statistic, C_1 and C_2 Have the Same Sample Size $N = 150$, C_1 Has Normally Distributed Data

	C_2	*Elliptical **x***		*Skew **f** & Normal **e***		*Skew **f** & **e***	
T		b	a	b	a	b	a
T_{ML}	T_M	3.166	–117.177	3.132	–113.210	3.347	–152.676
	T_I	1.492	–462.592	4.103	–3471.507	4.652	–3923.917
T_{AML}	T_M	3.166	–111.541	3.132	–107.765	3.347	–145.332
	T_I	1.492	–444.481	4.103	–3335.598	4.652	–3770.297
T_R	T_M	0.884	10.059	0.948	6.126	1.058	4.273
	T_I	1.472	–941.064	1.001	–725.069	1.231	–920.951
T_{AR}	T_M	0.884	9.575	0.948	5.831	1.058	4.067
	T_I	1.472	–904.222	1.001	–696.683	1.231	–884.896
T_{RADF}	T_M	0.799	38.717	0.825	31.263	0.710	54.620
	T_I	0.717	33.396	0.466	38.273	0.571	6.968
	T_O	0.317	235.940	0.312	201.524	0.239	–2.411
T_{CRADF}	T_M	0.884	9.361	0.911	6.666	0.806	15.779
	T_I	1.141	–23.544	1.477	–76.333	1.406	–63.747
T_{IRLS}	T_M	4.217	–201.608	4.107	–194.497	3.688	–186.975
	T_O	3.304	–2676.005	2.876	–2170.028	2.824	–2173.540

greater effect on a and b corresponding to T_{ML} in Tables 2 and 3. When Equation 1 does not hold or when κ changes from sample to sample, we would have more changes on a and b corresponding to T_{ML}, T_{AML}, T_R and T_{AR}. Of course, the a and b are also model dependent. When a different model other than that in Equation 9 is used, we will observe different patterns on a and b. Furthermore, a and b can only be estimated empirically. The estimators obtained from one sample may not be applied to a different sample for the purpose of correcting the performance of statistics. But Equations 10 and 11 do provide a simple procedure for comparing two distributions.

TABLE 3
Intercept a and Slope b in $(T_2|C_2) = b(T_1|C_1) + a$, Where T_2 and T_1 Are the Same Statistic, C_1 and C_2 Have the Same Sample Size $N = 500$, C_1 Has Normally Distributed Data

	C_2	*Elliptical* **x**		*Skew* ***f*** *& Normal* ***e***		*Skew* ***f*** *&* ***e***	
T		b	a	b	a	b	a
T_{ML}	T_M	4.580	–207.985	4.621	–214.494	3.980	–176.997
	T_I	1.536	–1783.119	4.855	–13638.862	5.291	–14887.005
T_{AML}	T_M	4.580	–204.998	4.621	–211.413	3.980	–174.455
	T_I	1.586	–1762.275	4.855	–13479.424	5.291	–14712.976
T_R	T_M	0.863	10.916	0.855	11.533	0.976	10.939
	T_I	2.097	–5135.660	1.438	–3916.342	1.620	–4440.145
T_{AR}	T_M	0.863	10.759	0.855	11.368	0.976	10.782
	T_I	2.097	–5075.624	1.438	–3870.560	1.620	–4388.239
T_{RADF}	T_M	0.899	11.116	0.876	13.179	0.725	28.562
	T_I	0.755	–120.164	0.430	–136.305	0.476	–163.075
	T_O	0.289	–259.937	0.285	–293.929	0.162	–254.850
T_{CRADF}	T_M	0.898	9.265	0.881	10.447	0.732	23.084
	T_I	1.337	–165.491	1.463	–302.667	1.557	–326.911
T_{IRLS}	T_M	6.124	–371.420	6.309	–395.120	4.496	–258.577
	T_O	2.951	–7377.125	2.432	–5408.609	2.182	–4572.810

In summary, chi-square distributions are generally not achievable even when data are normally distributed. Distribution shapes of the commonly used statistics vary substantially when conditions such as the sample size, the distribution of **x**, model size and model misspecification change. The results suggest that using (noncentral) chi-square distributions of T to describe the properties of a fit index F is inappropriate. In the following section we will discuss the properties of fit indices when T does not follow a chi-square distribution but can be approximated by Equations 10 or 11.

Note that all robust procedures are tailor-made. One has to properly choose the weight function in order to obtain the estimator or test statistic. For a given sample, one can find a proper weighting scheme so that $(T_{ML}|H_0)$ applied to the robustified sample approximately follows χ^2_{df} (Yuan & Hayashi, 2003). When T_{ML} approximately follows the same distribution across two robustified samples, then we will approximately have $a = 0$ and $b = 1$. But even if $(T_{ML}|H_0)$ approximately follows χ^2_{df}, this does not imply that $(T_{ML}|H_1)$ will approximately follow $\chi^2_{df}(\delta)$. We will further discuss whether $T \sim \chi^2_{df}(\delta)$ is achievable or necessary when interpreting fit indices involving T.

TABLE 4
Intercept a and Slope b in $(T_2|C_2) = b(T_1|C_1) + a$, Where T_2 and T_1 Are the Same Statistic and C_1 and C_2 Have the Same Distribution, but C_1 Has $N = 150$ and C_2 Has $N = 500$

	C_1 & C_2	*Normal* **x**		*Elliptical* **x**		*Skew* ***f*** *& Normal* ***e***		*Skew* ***f*** *&* ***e***	
T		b	a	b	a	b	a	b	a
T_{ML}	T_M	1.032	–5.963	1.599	–78.382	1.659	–92.813	1.299	–24.126
	T_I	1.830	1464.784	1.890	1333.560	2.077	1095.487	2.031	1101.599
T_{AML}	T_M	1.069	–5.877	1.656	–77.256	1.717	–91.480	1.345	–23.779
	T_I	1.883	1447.661	1.944	1317.970	2.136	1082.681	2.089	1088.722
T_R	T_M	1.008	–5.148	0.984	–3.406	0.910	1.476	0.932	1.712
	T_I	1.747	1358.086	2.514	37.271	2.502	–139.501	2.305	–124.598
T_{AR}	T_M	1.044	–5.074	1.019	–3.357	0.942	1.455	0.965	1.687
	T_I	1.797	1342.210	2.586	36.835	2.573	–137.871	2.371	–123.141
T_{RADF}	T_M	0.327	29.938	0.367	24.015	0.346	28.761	0.333	32.408
	T_I	0.300	526.297	0.311	270.135	0.277	79.972	0.250	85.748
	T_0	0.240	3827.699	0.220	793.140	0.221	750.715	0.164	363.810
T_{CRADF}	T_M	1.576	–54.287	1.585	–53.106	1.515	–46.771	1.416	–37.834
	T_I	2.437	–3.535	2.866	–103.960	2.399	–121.866	2.693	–159.363
T_{IRLS}	T_M	1.094	–1.971	1.824	–103.984	2.035	–141.218	1.402	–27.787
	T_0	2.589	735.523	2.342	942.115	2.328	954.490	1.913	1484.340

PROPERTIES OF FIT INDICES WHEN T DOES NOT FOLLOW A CHI-SQUARE DISTRIBUTION

To facilitate the discussion we use

$$T = bt + a, \tag{12}$$

where t might have a known distribution as in Equation 10 or might be the distribution of a given statistic under the normal sampling scheme with a fixed sample size, as in Equation 11. We denote the a and b at the independence model as a_I, b_I and at the substantive model as a_M and b_M, respectively. We will mainly discuss the stability of fit indices when the underlying distribution of **x** changes. Due to the uncritical use of T_{ML} in the practice of SEM, we will pay special attention to the properties of fit indices associated with T_{ML}. We cannot predict a_I, a_M or b_I, b_M in general. But in the special case when $\mathbf{\Sigma}_I$ is approximately correct, we might approximately have $a_I = a_M = 0$ and $b_I = b_M = b$ within the class of elliptical distributions, which will be specifically mentioned when a nice property holds.

The normed fit index NFI can be written as

$$\text{NFI} = 1 - \frac{b_M t_M + a_M}{b_I t_I + a_I},$$

which is invariant to $b_I = b_M = b$ and $a_I = a_M = 0$. When model $\mathbf{\Sigma}_I$ is approximately correct and $T = T_{ML}$, there may approximately exist $b_I = b_M = b$ and $a_I = a_M = 0$ within the class of elliptical distributions. However, when the off diagonal of $\mathbf{\Sigma}_0$ cannot be ignored, $b_I \neq b_M$ and a_I or a_M will not equal zero either. The distribution of NFI will change when $\mathbf{x}$ changes distributions or when $\mathbf{\Sigma}_M$ changes. Notice that the change happens not just in the mean value E(NFI) but also the distributional form of NFI. When n tends towards infinity,

$$\text{NFI} = 1 - \frac{\min_{\boldsymbol{\theta}} D\left[\mathbf{S}, \mathbf{\Sigma}_M(\boldsymbol{\theta})\right]}{\min_{\boldsymbol{\theta}} D\left[\mathbf{S}, \mathbf{\Sigma}_I(\boldsymbol{\theta})\right]} \to 1 - \frac{\min_{\boldsymbol{\theta}} D\left[\mathbf{\Sigma}_0, \mathbf{\Sigma}_M(\boldsymbol{\theta})\right]}{\min_{\boldsymbol{\theta}} D\left[\mathbf{\Sigma}_0, \mathbf{\Sigma}_I(\boldsymbol{\theta})\right]}, \tag{13}$$

which measures the reduction (explanation) of all the covariances by $\mathbf{\Sigma}_M$ relative to $\mathbf{\Sigma}_I$. However, the limit in Equation 13 still depends on the measure $D(\cdot,\cdot)$ of distance (see La Du & Tanaka, 1989, 1995; Sugawara & MacCallum, 1993; Tanaka, 1987) unless the model $\mathbf{\Sigma}_M$ is correctly specified (Yuan & Chan, in press). In practice, N is always finite,

$$E(\text{NFI}) \neq 1 - \frac{\min_{\theta} D\left[\mathbf{\Sigma}_0, \mathbf{\Sigma}_M(\boldsymbol{\theta})\right]}{\min_{\theta} D\left[\mathbf{\Sigma}_0, \mathbf{\Sigma}_I(\boldsymbol{\theta})\right]},$$

in general. Furthermore, the speed of convergence to the limit in Equation 13 might be very slow. So interpreting NFI according to its limit is not appropriate.

Very similar to NFI is GFI. Using the notation from the previous section, we can write GFI as

$$\text{GFI} = 1 - \frac{b_M t_M + a_M}{b_0 t_0 + a_0}.$$

When the model is correctly specified, $n\mathbf{e}'\mathbf{We}$ asymptotically follows a chi-square distribution, but $n\mathbf{s}'\mathbf{Ws}$ will not approximately follow a noncentral chi-square distribution. Based on the result from the previous section, GFI changes its value as well as its distribution when N, the distribution of $\mathbf{x}$, as well as the model specification and the magnitude of $\mathbf{\Sigma}_0$ change. As n tends to infinity, the limit of GFI measures the relative reduction of all the variances-covariances by model $\mathbf{\Sigma}_M$ relative to a model with all elements equal to zeros. Similar to NFI, the limit of GFI depends on the measure $D(\cdot,\cdot)$ which further depends on the weight $\mathbf{W}$. The speed of convergence can be very slow. Interpreting GFI as a reduction of all the variances and covariances in the population may not be appropriate.

Using Equation 12, the fit index NNFI can be rewritten as

$$\text{NNFI} = 1 - \frac{b_M t_M / df_M + a_M / df_M - 1}{b_I t_I / df_I + a_I / df_I - 1}.$$

When T_M and T_I are based on T_{ML} and $\boldsymbol{\Sigma}_I$ is approximately correct, then $b_I \approx b_M = b$ and $a_I \approx a_M \approx 0$ within the class of elliptical distributions. Thus, $\text{NNFI} \approx \left(t_I / df_I - t_M / df_M\right)\left(t_I / df_I - 1/b\right)$ is a decreasing function of b. However, the model may not be approximately correct in practice, and hence the value of NNFI as well as its distribution will be affected by the distribution of $\mathbf{x}$, the sample size N, the estimation method, as well as model misspecifications. Of course, using a noncentral chi-square distribution to interpret the value of NNFI is generally inappropriate.

For RNI, the nonnormed version of CFI, we have

$$\text{RNI} = 1 - \frac{b_M t_M + a_M - df_M}{b_I t_I + a_I - df_I}.$$

In practice, CFI and RNI will be affected by sample size, the distribution of $\mathbf{x}$ and the measure $D(\cdot,\cdot)$ used in constructing the fit indices. When T cannot be described by a noncentral chi-square distribution, interpreting CFI or RNI as a reduction of NCP may not be appropriate.

Using Equation 12 we can rewrite RMSEA as

$$\text{RMSEA} = \sqrt{\left(b_M t_M + a_M - df_M\right) / \left(n \times df_M\right)}.$$

It is obvious that RMSEA is an increasing function of a_M and b_M, which will change with the distribution of $\mathbf{x}$, the sample size, the model misspecification as well as the discrepancy function $D(\cdot,\cdot)$. The confidence interval for RMSEA, as printed in standard software, makes sense only when $b_M = 1$ and $a_M = 0$, which will not hold in any realistic situation. Notice that the larger the n, the smaller the effect of a_M; but the effect of b_M will not diminish even when n is huge.

In conclusion, when the t in Equation 12 is anchored at a chi-square distribution or at the distribution of a statistic under fixed conditions, fit indices F defined through T are functions of a and b, which change across conditions. Desirable properties of F are difficult to achieve even with the idealized elliptically distributed data.

MATCHING FIT INDICES WITH STATISTICS

For fit indices involving $T - df$, the rationale is $E(T|H_0) = df$ so that $\hat{\delta} = T - df$ is an unbiased estimate of the part of $E(T)$ due to model misspecifications while df is the part due to random sampling that is not related to model misspecification. When $E(T|H_0) \neq df$, the meaning of $T - df$ is not clear. One cannot use the "degrees of freedom" to justify the use of $\hat{\delta} = T - df$. For example, when $E(T|H_0) = 1.5df$, then $\hat{\delta} = T - df$ does not make more sense than $\tilde{\delta} = T - c$ with c being an arbitrary constant. Fortunately, there do exist a few statistics that approximately satisfy $E(T|H_0) = df$. Based on the simulation results of Hu et al. (1992), Chou et al. (1991), Curran et al. (1996), Yuan and Bentler (1998b), Fouladi (2000) and others, the statistic T_R approximately have a mean equal to df when n is relatively large. This is also implied by the asymptotic theory, as reviewed in the second section of this article. For small n, results in Bentler and Yuan (1999) indicate that $E(T_R)$ is substantially greater than df even when $\mathbf{x}$ follows a multivariate normal distribution. The results in Bentler and Yuan (1999) and Yuan and Bentler (1998b) indicate that $(T_{CRADF}|H_0)$ approximately has a mean equal to df regardless of the distribution of $\mathbf{x}$ and N. For convenience, some of the empirical results in Bentler and Yuan (1999) and Yuan and Bentler (1998b) are reproduced in Table 5, where similar data generation schemes as in Table 1 were used.

TABLE 5
Empirical Means[a] of T_R, T_{AR}, and T_{CRADF} ($df = 87$)

		Sample Size N									
		90	*100*	*110*	*120*	*150*	*200*	*300*	*500*	*1000*	*5000*
Normal **x**	T_R	96.33	95.70	94.95	94.16	90.89	89.75	88.24	87.72	87.91	87.29
	T_{AR}	88.57	88.77	88.71	88.49	86.52	86.52	86.12	86.46	87.28	87.16
	T_{CRADF}	87.42	88.90	89.90	89.66	90.36	90.50	90.22	89.81	88.31	87.44
Elliptical **x**	T_R	96.52	95.76	94.57	93.80	90.47	88.48	87.19	86.11	86.02	86.67
	T_{AR}	88.75	88.83	88.35	88.15	86.12	85.29	85.10	84.87	85.40	86.55
	T_{CRADF}	87.50	88.55	88.97	88.83	89.28	90.11	89.99	89.91	89.09	87.84
Normal **f** & skew **e**	T_R	98.80	98.30	96.66	95.75	92.12	90.64	88.93	88.19	88.24	87.19
	T_{AR}	90.84	91.18	90.30	89.98	87.69	87.38	86.80	86.92	87.61	87.07
	T_{CRADF}	87.39	88.21	88.76	89.07	88.57	89.04	88.39	88.06	88.81	88.08
Skew **f** & **e**	T_R	99.68	97.36	96.34	95.40	91.77	90.77	88.92	87.52	86.68	87.61
	T_{AR}	91.65	90.31	90.01	89.65	87.36	87.50	86.79	86.26	86.06	87.48
	T_{CRADF}	87.45	88.44	88.85	89.43	88.52	88.87	87.97	88.17	88.36	88.50

[a] All the three statistics asymptotically follow χ^2_{87}.

Notice that the empirical mean of T_{ML} is greater than *df* even when data are normally distributed, especially when N is small (see Bentler & Yuan, 1999). The well-known Bartlett correction is to make $E(T_{ML})$ more close to *df*. We might apply such a correction to the statistic T_R when data are not normally distributed. The Bartlett correction for exploratory factor model (EFM) with normal data already exists (Lawley & Maxwell, 1971, pp. 35–36). Yuan, Marshall and Bentler (2002) applied it to the rescaled statistic in exploratory factor analysis. The Bartlett correction for general covariance structures also exists for T_{ML} when data are normally distributed (see Wakaki, Eguchi, & Fujikoshi, 1990). However, the correction is quite complicated even for rather simple models. For nonnormal data, this correction is no longer the Bartlett correction. Here we propose an ad hoc correction to the statistic T_R. Consider the Bartlett correction with m factor in EFM,

$$T_{BML} = \left[N - 1 - (2p+5)/6 - 2m/3\right] D_{ML}\left[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right], \tag{14}$$

where the total number of free parameters is $p(m+1) - m(m-1)/2$. Notice that the CFM and EFM are identical when there is a single factor, then the coefficient in Equation 14 is $N - 5/3 - (2p + 5)/6$. In formulating a SEM model, unidimensional (univocal) measurements are recommended and typically used in practice (Anderson & Gerbing, 1988). When all the measurements are unidimensional, the number of factor loadings in the SEM model is equal to that of the 1-factor model. The factors can be correlated or predicted by each other. With m factors, there are $m(m - 1)/2$ additional parameters when the structural model is saturated. Considering that there may exist double or multiple loadings and the structural model may not be saturated, we propose the following ad hoc correction to T_R,

$$T_{AR} = n^{-1}\left[N - 5/3 - (2p+5)/6 - (m-1)/3\right] T_R. \tag{15}$$

Table 5 contrasts the empirical means of T_{AR} and T_R, based on the empirical results in Yuan and Bentler (1998b) and Bentler and Yuan (1999). Although many of the empirical means of T_{AR} are still greater than $df = 87$ when n is small, they are much nearer to the *df* than the corresponding ones of T_R. A parallel T_{AML} follows when applying the correction factor in Equation 15 to T_{ML}.

The statistic T_{CADF} also has empirical means very close to *df* for various distribution and sample size conditions (Yuan & Bentler, 1997b). However, it is easier to get a converged solution with minimizing D_{ML} than with D_{ADF}. The statistic T_{AML} can also be used when data are normally distributed. But $E(T_{AML})$ is unpredictable when the distribution of $\mathbf{x}$ is unknown as with practical data. A nonsignificant Mardia's multivariate kurtosis does not guarantee that $E(T_{ML}) \approx df$, as showed in Example 3 of Yuan and Hayashi (2003). So it is wise to avoid T_{ML} or even T_{AML} before an effective way of checking normality is performed. The means of T_{ADF} and T_{RADF} are much greater than *df* unless *N* is huge (see Hu et al., 1992; Yuan & Bentler, 1997b, 1998b), thus, they are not recommended to be used with fit indices involving *df*. Although both T_{CRADF} and T_R are now avaliable in EQS (Bentler, 2004), T_{CRADF} can only be applied when $n > df$. Similarly, T_R makes sense only when $n > df$ (see Bentler & Yuan, 1999) although it can still be numerically calculated when $n < df$.

Yuan and Hayashi (2003) showed that the statistic T_{ML} applied to a proper robustified sample can closely follow χ^2_{df}. When T approximately follows χ^2_{df}, the empirical mean of T also approximately equals *df*. Also, applying T_{ML} with a robust procedure does not need the specific condition $n > df$. Of course, one may not be able to apply the robust procedures to any data. We will further discuss this in the appendix.

In this section we argue that fit indices involving $T - df$ do not make much sense when $E(T) \neq df$. Because T_{AR} and T_{CRADF} approximately satisfy $E(T) = df$, they should be used in defining fit indices involving *df* instead of T_{ML} or T_{ADF}. Robust procedures with T_{ML} are preferred when data contain heavy tails or when $n < df$, but one needs to check that $E(T_{ML}) = df$ using resampling or simulation (see Yuan & Hayashi, 2003).

STABILITY OF FIT INDICES WHEN CONDITIONS CHANGE

Notice that the distributions of T_{AR} and T_{CRADF} still change substantially even though their empirical means approximately equal *df*. So we still need to study how the distributions of fit indices change when changing conditions. In this section, we mainly study the effect of the distribution of **x** with different sample sizes. This is partially because people seldom check the distribution of the data when using fit indices in practical data analysis. Previous studies on fit indices mainly focused on *T*s based on $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with different sample sizes or estimation methods (Anderson & Gerbing, 1984; Fan, Thompson, & Wang, 1999; La Du & Tanaka, 1989, 1995; Marsh et al., 1988, 2004), but not the statistic T_{AR} or T_{CRADF}. We will resort to Equation 12 and rewrite it as $F = bf + a$ for the current purpose, where f represents a fit index when $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and F represents a fit index when **x** follows a different distribution. Because it does not make sense to apply T_{ML} or T_{RADF} to fit indices involving *df* when data are nonnormally distributed or when sample size is not huge, we only study the performance of these fit indices when evaluated by T_{AR} and T_{CRADF}. Specifically, we will study the effect of distributional change of **x** on NFI, NNFI, GFI, CFI, and RMSEA. Because NFI does not involve *df*, we will report its distribution when evaluated using T_{ML}, T_{AR}, T_{RADF}, and

T_{CRADF}. Similarly, we will evaluate GFI when T_M and T_0 are T_{IRLS} or T_{RADF}, as previously discussed. Parallel to Tables 2 to 4 and with the same CFM as in Equation 9, we choose three distribution conditions for **x**. Table 6 contains the intercept a and slope b for the above designs when $N = 150$ and 500. It is obvious that most of the fit indices change distributions substantially when **x** changes its distribution. Sample size also has a big effect on a and b. Among these, RMSEA changes its overall distribution the least. Notice that both T_{AR} and T_{CRADF} change distributions substantially in Tables 2 to 4. RMSEA is a function of $\hat{\delta}$ and totally decided by T. The greater stability of RMSEA is due to the effect of a square root, while no other fit index studied here uses such a transformation. The same square root effect may attenuate its sensitivity when the model $\mathbf{\Sigma}_M$ changes. Hu and Bentler (1998) studied sensitivity of fit indices, but not when they are evaluated by T_{AR} and T_{CRADF}. Further study in this direction is needed.

TABLE 6
Intercept a and Slope b With $F = bf + a$, Where f Is the Distribution of F for Normal Data

			Elliptical **x**		*Skew* **f** *& Normal* **e**		*Skew* **f** *&* **e**	
F	*T*	*N*	*b*	*a*	*b*	*a*	*b*	*a*
NFI	T_{ML}	150	2.510	–1.443	3.902	–2.734	3.306	–2.152
		500	4.220	–3.160	5.766	–4.667	4.756	–3.673
	T_{AR}	150	3.676	–2.504	9.244	–7.715	9.795	–8.207
		500	5.780	–4.668	20.864	–19.394	21.538	–20.048
	T_{RADF}	150	1.257	–0.265	1.466	–0.576	1.453	–0.502
		500	1.692	–0.688	3.383	–2.442	3.312	–2.355
	T_{CRADF}	150	0.949	–0.009	0.901	–0.057	0.891	–0.027
		500	1.200	–0.215	1.816	–0.892	1.738	–0.815
GFI	T_{IRLS}	150	3.244	–2.137	3.246	–2.137	3.166	–2.046
		500	5.354	–4.274	5.482	–4.398	4.409	–3.344
	T_{RADF}	150	2.337	–1.351	2.288	–1.307	3.342	–2.367
		500	3.963	–2.968	4.012	–3.019	7.657	–6.680
CFI	T_{AR}	150	2.256	–1.251	9.524	–8.489	12.132	–11.103
		500	2.571	–1.567	11.760	–10.737	14.082	–13.060
	T_{CRADF}	150	1.122	–0.111	1.616	–0.663	1.423	–0.407
		500	1.346	–0.346	3.618	–2.626	3.001	–2.002
NNFI	T_{AR}	150	2.322	–1.313	280.605	–278.981	15.448	–14.396
		500	2.948	–1.939	21.691	–20.635	18.770	–17.729
	T_{CRADF}	150	1.044	–0.018	61.079	–47.357	32.034	–23.782
		500	1.301	–0.305	5.680	–4.590	3.439	–2.413
RMSEA	T_{AR}	150	0.932	–0.002	0.916	–0.002	1.017	0.006
		500	0.914	–0.003	0.871	–0.003	1.015	–0.001
	T_{CRADF}	150	0.930	–0.001	0.937	–0.001	0.864	–0.001
		500	0.964	0.000	0.937	0.000	0.861	–0.000

The results of this section tell us that fit indices change their distributions substantially when conditions change. Commonly used cutoff values or confidence intervals for fit indices do not reflect these changes and thus provide questionable values regarding model fit/misfit.

THE POWER OF A FIT INDEX

There exist two seemingly unrelated aspects of power based on fit indices. One is the ability for a fit index to distinguish a good model from a bad model. The other is to first use fit indices F to specify null and alternative hypotheses and then to analyze the power of T under these two hypothesis. We will discuss these two aspects separately.

Commonly used cutoff values such as 0.05 or 0.95 for fit indices do not directly relate the fit indices to any null hypothesis. So one cannot treat the cutoff value as the traditional critical value and use the corresponding rejection rate as power. Even when one specifies a null hypothesis, say the model is correctly specified or the distance between the model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and the population $\boldsymbol{\Sigma}_0$ is less than a small number, then 0.05 or 0.95 may have nothing to do with the upper 5% quantiles of the distribution of the fit index under the specified null hypothesis. As we have seen from the previous sections, all fit indices change distributions when N, $\mathbf{x}$ and $D(\cdot,\cdot)$ change; we cannot relate the means of fit indices to the NCP even under the idealized conditions in Equation 2. In practice, we generally do not know the distribution of fit indices even when $\boldsymbol{\Sigma}_M$ is correctly specified, and ignoring their relationship to model misspecifications.

With the above mentioned difficulties, a more sensible approach to study the sensitivity or power of fit indices is by simulation or bootstrap, as in Hu and Bentler (1998, 1999), Gerbing and Anderson (1993), Bollen and Stine (1993), Yung and Bentler (1996), Muthén and Muthén (2002), and Yuan and Hayashi (2003). In such a study, the N, $D(\cdot,\cdot)$, and the distribution of $\mathbf{x}$ are all given. We still need to specify interesting null H_0 and alternative H_1 hypotheses. The null and alternative hypotheses should include both $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_M$. Then the distributions of a fit index F under H_0 and H_1 can be estimated through simulations. The cutoff value can be obtained from a proper quantile of the estimated distribution $(\widehat{F \mid H_0})$. The power of F at the given conditions will be the proportion of $(\widehat{F \mid H_1})$ that fall into the rejection region decided by the cutoff value. We need to emphasize that simulation or bootstrap are tailor-made approaches and may not be generalizable when conditions change. As we have seen, in addition to model misspecifications, the distribution of F changes with the sample size N, the distribution of $\mathbf{x}$ as well as the chosen statistic T. The most sensitive F under one set of conditions may no longer be most sensitive under a different set of conditions.

When studying power of T one has to specify H_0 and H_1 (e.g., Satorra & Saris, 1985). Traditionally, H_0 represents a correct model and H_1 represents an interesting alternative model. Because substantive models are at most approximately correct, MacCallum et al. (1996), MacCallum and Hong (1997), and Kim (2003) proposed to let H_0 represent a misspecified model with misspecification measured by RMSEA, CFI, GFI, and other fit indices. They used $(T \mid H_0) \sim \chi^2_{df}(\delta_0)$ and $(T \mid H_1) \sim \chi^2_{df}(\delta_1)$ to calculate the critical value and power. Such a proposal would allow researchers to test interesting but not necessarily perfect models when both $(T \mid H_0) \sim \chi^2_{df}(\delta_0)$ and $(T \mid H_1) \sim \chi^2_{df}(\delta_1)$ are attainable. When they are not attainable, then the critical value obtained using $\chi^2_{df}(\delta_0)$ or power based on $\chi^2_{df}(\delta_1)$ may have little relationship with the model misspecification. As with distributions of fit indices, distributions of the statistics depend on sample size, the distribution of $\mathbf{x}$ as well as the choice of statistic T, in addition to model misspecification. Again, a more sensible approach to power might be by simulation or bootstrap (see Yuan & Hayashi, 2003).

In conclusion, power or sensitivity of fit indices and test statistics might need to be evaluated by simulation. In such a study, one has to control both type of errors[2] using simulated critical values rather than referring F to 0.05 or 0.95 or referring T to $\chi^2_{df}(\delta)$. However, the findings by simulation in one set of conditions may not be generalizable to a different set of conditions.

DISCUSSION

The analysis and empirical results in the previous sections should help to provide a clearer overall picture of the relationship between fit indices and test statistics. We will further discuss some issues related to their relationship, which may lead to a better understanding of the properties of the fit indices.

As indicators for model fit, fit indices should relate to model misspecification. The commonly used measure of model misspecification is given by the τ in Equation 7. However, the definition there only makes a clear sense when $D = D_{ML}$ corresponding to T_{ML} at the population level. At the sample level, $D_{ML}\left[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]$ can still be calculated but it depends on N and the distribution of $\mathbf{x}$. Unfortunately, current literature does not provide us an analytical formula relating $D_{ML}\left[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]$ to τ and other conditions such as N and the underlying distribution of $\mathbf{x}$. For example, at a given N we do not know how good the approximation in Equation 1 is even when $\mathbf{x}$ follows an elliptical distribution and the model is correctly specified. What we know is that $D_{ML}\left[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]$ is consistent for τ, as well as that $E\left\{D_{ML}\left[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]\right\}$ is generally increasing with τ and the population kurtosis κ when $\mathbf{x}$ follows an ellipti-

[2] Yuan and Bentler (1997b), Fouladi (2000), and Curren et al. (1996) studied the power of various statistics but none of these studies control type I errors properly.

cal distribution. But such knowledge is not enough to accurately estimate τ, establish reliable cutoff values, or compute meaningful confidence intervals for τ.

When using T_R, T_{AR}, T_{ADF}, T_{CADF}, T_{RADF}, T_{CRADF} for model evaluation, the measure corresponding to τ should be

$$\tau_{\Gamma} = \min_{\boldsymbol{\theta}} D^{(\boldsymbol{\Gamma})}\left[\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right],$$

where $\boldsymbol{\Gamma}$ is the fourth-order population covariance matrix approximately equal to $N\text{Cov}(\mathbf{s})$, $D^{(\boldsymbol{\Gamma})}$ is the population counterpart of the discrepancy function corresponding to each T. In Monte Carlo studies, $\boldsymbol{\Gamma}$ can be obtained if one generates random numbers following the procedure of Yuan and Bentler (1997a). One generally does not know $\boldsymbol{\Gamma}$ with real data. The problem is that all fit indices and test statistics are related to $\boldsymbol{\Gamma}$ at the population level, including $E\left\{D_{ML}\left[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right]\right\}$. But $\boldsymbol{\Gamma}$ itself is difficult to estimate due to its large dimension. The idea in robust procedures is to transform the sample so that the corresponding $\boldsymbol{\Gamma}$ is comparable to that in a normal population with the same $\boldsymbol{\Sigma}_0$. The sensitivity or power of a fit index is closely related to $\boldsymbol{\Gamma}$. It is well-known that, with given $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, the greater the $\boldsymbol{\Gamma}$ (in the sense of positive definiteness) the smaller the power of a test statistic (e.g., Shapiro & Browne, 1987; Tyler, 1983; Yuan & Hayashi, 2003; Yuan et al., 2004). Fit indices that only involve T_M will inherit such a property. It is unclear whether the sensitivity of a fit index involving T_I will also be inversely affected by a large $\boldsymbol{\Gamma}$ matrix.

In practice, we deal with τ or τ_{Γ} through T. When $T \sim \chi^2_{df}(\delta)$ holds, $\hat{\delta} = T - df$ contains valuable information for model misspecification. The confidence intervals for δ and monotonic functions of δ also make probabilistic sense, although we may not be able to relate δ to τ or to the expectation of a fit index $E(F)$ explicitly. We want to reemphasize that the result in Equations 6 and 7 is not correct when the population covariance matrix $\boldsymbol{\Sigma}_0$ is fixed.

When $E(T|H_0) = df$ is true but $T \sim \chi^2_{df}(\delta)$ does not hold, $\hat{\delta} = T - df$ also contains valuable information for model misspecification under the condition that $E(T|H_1)$ increases as the model deteriorates. We can interpret $\hat{\delta}$ as the systematic part of T that is beyond the effect of sampling errors corresponding to the correct model. But a confidence interval for δ based on $T \sim \chi^2_{df}(\delta)$ does not make sense.

When $E(T|H_0) \neq df$, we might modify $\hat{\delta} = T - df$ to

$$\hat{\delta}_D = T - E\left(\widehat{T \mid H_0}\right),$$

where the subscript D is used to denote the involved discrepancy measure in formulating T. The population counterpart of $\hat{\delta}_D$ is $\delta_D = E(T \mid H_1) - E(T \mid H_0)$, which is the systematic part of T that is beyond the effect of sampling errors corresponding to the correct model. It is obvious that $\delta_D = 0$ when $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is correctly specified, regardless of N, D or $\mathbf{x}$. The drawback of $\hat{\delta}_D$ is that $E\left(\widehat{T \mid H_0}\right)$ is not as easy to obtain as *df*. Instead, one has to use a resampling based procedure or other alternatives to obtain a $E\left(\widehat{T \mid H_0}\right)$. See Yuan and Marshall (2004) for further comparison of δ_D with δ and ways to obtain confidence intervals for δ_D.

With the distribution of $\mathbf{x}$ generally unknown and N uncontrollable, $E(T|H_0) \approx df$ is not difficult to achieve. When N is not too small, $(T|H_0)$ approximately following χ^2_{df} is also achievable by choosing a proper T. But $(T|H_1)$ approximately following $\chi^2_{df}(\delta)$ is generally not achievable. Actually, we cannot analytically relate $E(F)$ to δ even when $(T \mid H_1) \sim \chi^2_{df}(\delta)$. For fit indices that are monotonic functions of T without referring to a base model, a confidence interval for δ can be obtained by assuming $T \sim \chi^2_{df}(\delta)$, but not a confidence interval for τ or τ_Γ because the key equation

$$\tau = \delta/n,$$

still depends on unrealistic conditions (Yuan & Bentler, in press; Yuan & Marshall, 2004). For fit indices $g\left(\hat{\delta}_M, \hat{\delta}_I\right)$ that involve both T_M and T_I, assumptions in Equation 2 do not allow one to obtain even a confidence interval for $g(\delta_M, \delta_I)$. So, the noncentral chi-square assumption is neither realistic nor necessary in making inference for τ or τ_Γ. For NFI, the sample size N is cancelled by the ratio. We will have a consistent estimator for $\mathrm{NFI}_0 = 1 - \min_{\boldsymbol{\theta}} D\left[\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right] / \min_{\boldsymbol{\theta}} D\left[\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_I(\boldsymbol{\theta})\right]$. Similar conclusion can be reached for other fit indices when N is cancelled. Of course, consistency is a very minimal requirement, which does not tell us how far it is from the population quantity for a finite N.

We might need to distinguish the role of a T as a statistic to evaluate whether a model is correctly specified from its role in formulating a fit index. As a test statistic, T is well behaved when $(T|H_0)$ approximately overlaps with χ^2_{df} at the 95% quantile and the greater the $(T|H_1)$ the better the T. We do not really need for T to follow chi-square distributions.

CONCLUSION AND RECOMMENDATION

We wanted to identify some statistics or estimation methods coupled with fit indices that are relatively stable when sample sizes and sampling distributions change. We may also want a fit index F to perform stably when changing the base-statistic T. Since the choice of T is controllable, it is not too bad when F varies across different Ts. The problem is that most fit indices are not stable across the uncontrollable conditions N and the distribution of $\mathbf{x}$. RMSEA is relatively most stable among the commonly used fit indices. But it is totally decided by T and none of the Ts are stable when changing N or $\mathbf{x}$. Thus, the stability of RMSEA may be just due to the artificial effect of the square root, which will attenuate its sensitivity across model misspecifications.

Although our analysis of fit indices still leaves many uncertainties, some conclusions can nonetheless be reached.

1. Given the population covariance matrix and the model structure, the mean value as well as the distribution of fit indices change with the sample size, the distribution of the data as well as the chosen statistic. Fit indices also reflect these variables in addition to reflecting model fit. Thus, cutoff values for fit indices, confidence intervals for model fit/misfit, and power analysis based on fit indices are open to question.

2. Statistics T_{AR} and T_{CRADF} have means approximately equal to *df* and are recommend for calculating fit indices involving *df*. Other statistics should not be used unless certain conditions such as normally distributed data and a large enough sample size are realized.

3. The asymptotic distribution of T_{AR} is generally unknown. The distribution of T_{CRADF} may be far from χ^2_{df} when the sample size is not large enough. Assuming they follow noncentral chi-square distributions is even more farfetched. Confidence intervals for model fit/misfit are not justified even when T_{AR} or T_{CRADF} are used in the evaluation.

4. The statistic T_{ML} applied to a robustified sample is recommended when the majority of the data are symmetrically distributed. This procedure is better coupled with the bootstrap procedure so that one can check that $(T_{ML}|H_0)$ approximately follows χ^2_{df}. The noncentral chi-square distribution is neither realistic nor necessary for evaluating model fit in the bootstrap procedure.

5. Our conclusions are temporary, as further study is needed to understand how each fit index is related to model misfit when evaluated by the more reasonable statistics T_{AR}, T_{CRADF}, or T_{ML} based on a robust procedure.

We believe our analysis and discussion have effectively addressed the questions raised at the beginning of the article. Although most of our results are on the negative side of current practice, fit indices are still meaningful as relative measures of model fit/misfit. Specifically, for NFI, NNFI, CFI, and GFI there exists $E(F \mid C_1) > E(F \mid C_2)$ when C_1 contains the model corresponding to a smaller τ or τ_Γ than C_2 while all the other conditions such as N, the distribution of $\mathbf{x}$, and $D(\cdot,\cdot)$ in C_1 and C_2 are the same; the opposite inequality sign holds for RMSEA. But such a relative value of fit indices should be distinguished from relative fit indices that utilize a base model. The relative value should not be interpreted as cutoff values being relative, for example, some use 0.95 while others use 0.90 to decide the adequacy of a model. We want to emphasize that the purpose of the article is not to abandon cutoff values but to point out the misunderstanding of fit indices and their cutoff values. We hope the result of the article can lead to better efforts to establish a scientific norm on the application of fit indices.

REFERENCES

Amemiya, Y., & Anderson, T. W. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *Annals of Statistics*, *18*, 1453–1463.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness of fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155–173.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423.

Anderson, R. D. (1996). An evaluation of the Satorra-Bentler distributional misspecification correction applied to McDonald fit index. *Structural Equation Modeling*, *3*, 203–227.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika*, *48*, 493–517.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Bentler, P. M. (1995). EQS structural equations [Computer program and manual]. Encino, CA: Multivariate Software.

Bentler, P. M. (2004). EQS6 structural equations [Computer program and manual]. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.

Bentler, P. M., & Dijkstra, T. K. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (Ed.), *Multivariate analysis VI* (pp. 9–42). Amsterdam: North-Holland.

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*, 181–197.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika*, *51*, 375–377.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*, 303–316.

Bollen, K. A., & Stine, R. (1993). Bootstrapping goodness of fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 111–135). Newbury Park, CA: Sage.

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Browne, M. W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology*, *41*, 193–208.

Chou, C.-P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, *44*, 347–357.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research*, *37*, 1–36.

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

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, *6*, 56–83.

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, *7*, 356–410.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40–65). Newbury Park, CA: Sage.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods & Research*, *11*, 325–344.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under parameterized model misspecification. *Psychological Methods*, *3*, 424–453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

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

Huba, G. J., & Harlow, L. L. (1987). Robust structural equation models: Implications for developmental psychology. *Child Development*, *58*, 147–66.

Kano, Y. (1992). Robust statistics for test-of-independence and related structural models. *Statistics & Probability Letters*, *15*, 21–26.

Kano, Y., Berkane, M., & Bentler, P. M. (1990). Covariance structure analysis with heterogeneous kurtosis parameters. *Biometrika*, *77*, 575–585.

Kim, K. (2003). *The relationship among fit indices, power, and sample size in structural equation modeling*. Unpublished doctoral dissertation, UCLA.

La Du, T. J., & Tanaka, J. S. (1989). The influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology*, *74*, 625–636.

La Du, T. J., & Tanaka, J. S. (1995). Incremental fit index changes for nested structural equation models. *Multivariate Behavior Research*, *30*, 289–316.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). New York: American Elsevier.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.

MacCallum, R. C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research*, *32*, 193–210.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391–410.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in over-generalizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*, 97–103.

McDonald, R. P., & Ho, R. M. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107*, 247–255.

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

Mooijaart, A., & Bentler, P. M. (1991). Robustness of normal theory statistics in structural equation models. *Statistica Neerlandica*, *45*, 159–171.

Muthén, L. K., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620.

Ogasawara, H. (2001). Approximations to the distributions of fit indexes for misspecified structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 556–574.

Olsson, U. H., Foss, T., & Breivik, E. (2004). Two equivalent discrepancy functions for maximum likelihood estimation: Do their test statistics follow a non-central chi-square distribution under model misspecification? *Sociological Methods & Research*, *32*, 453–500.

Raykov, T. (2000). On the large-sample bias, variance, and mean squared error of the conventional noncentrality parameter estimator of covariance structure models. *Structural Equation Modeling*, *7*, 431–441.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*, 131–151.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology*, *22*, 249–278.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *American Statistical Association 1988 Proceedings of Business and Economics Sections* (pp. 308–313). Alexandria,VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis*, *10*, 235–249.

Satorra, A., & Saris, W. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90.

Shapiro, A. (1983). Asymptotic distribution theory in the analysis of covariance structures (a unified approach). *South African Statistical Journal*, *17*, 33–81.

Shapiro, A. (1985). Asymptotic equivalence of minimum discrepancy function estimators to GLS estimators. *South African Statistical Journal*, *19*, 73–81.

Shapiro, A., & Browne, M. (1987). Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association*, *82*, 1092–1097.

Steiger, J. H. (1989). *EZPATH: A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Sugawara, H. M., & MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement*, *17*, 365–377.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*, 134–146.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, *38*, 197–201.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Tyler, D. E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika*, *70*, 411–420.

Wakaki, H., Eguchi, S., & Fujikoshi, Y. (1990). A class of tests for a general covariance structure. *Journal of Multivariate Analysis*, *32*, 313–325.

Wang, L., Fan, X., & Willson, V. L. (1996). Effects of non-normal data on parameter estimates in covariance structure analysis: An empirical study. *Structural Equation Modeling*, *3*, 228–247.

Yuan, K.-H., & Bentler, P. M. (1997a). Generating multivariate distributions with specified marginal skewness and kurtosis. In W. Bandilla & F. Faulbaum (Eds.), *SoftStat'97– Advances in statistical software 6* (pp. 385–391). Stuttgart, Germany: Lucius & Lucius.

Yuan, K.-H., & Bentler, P. M. (1997b). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, *92*, 767–774.

Yuan, K.-H., & Bentler, P. M. (1998a). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, *51*, 63–88.

Yuan, K.-H., & Bentler, P. M. (1998b). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *51*, 289–309.

Yuan, K.-H., & Bentler, P. M. (1998c). Structural equation modeling with robust covariances. *Sociological Methodology*, *28*, 363–396.

Yuan, K.-H., & Bentler, P. M. (1999a). F-tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, *24*, 225–243.

Yuan, K.-H., & Bentler, P. M. (1999b). On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. *Statistica Sinica*, *9*, 831–853.

Yuan, K.-H., & Bentler, P. M. (2000). Robust mean and covariance structure analysis through iteratively reweighted least squares. *Psychometrika*, *65*, 43–58.

Yuan, K.-H., & Bentler, P. M. (in press). Mean comparison: Manifest variable versus latent variable. *Psychometrika*.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*, 21–436.

Yuan, K.-H., & Chan, W. (in press). On nonequivalence of several procedures of structural equation modeling. *Psychometrika*.

Yuan, K.-H., Chan, W., & Bentler, P. M. (2000). Robust transformation with applications to structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *53*, 31–50.

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *56*, 93–110.

Yuan, K.-H., & Marshall, L. L. (2004). A new measure of misfit for covariance structure models. *Behaviormetrika*, *31*, 67–90.

Yuan, K.-H., Marshall, L. L., & Bentler, P. M. (2002). A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika*, *67*, 95–122.

Yung, Y. F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Techniques and issues* (pp. 195–226). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Zhang, W. (2004). *Comparing RMSEA and chi-square/df ratio*. Unpublished manuscript.

APPENDIX

EQS 6.0 (Bentler, 2004) has the case-robust procedure based on Yuan and Bentler (1998c), which contains two tuning parameters in controlling the case weights. Here we give a brief introduction to the robust transformation proposed by Yuan et al. (2000), where Huber-type weights (see Tyler, 1983) that only contain one tuning parameter are used.

Let ρ be the percentage of influential cases one wants to control, and r be a constant decided by ρ through $P\left(\chi_p^2 > r^2\right) = \rho$. Denote the Mahalanobis distance as

$$d = \left[\left(\mathbf{x} - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)\right]^{1/2}.$$

The Huber-type weights are given by

$$u_1(d) = \begin{cases} 1, & \text{if } d \le r \\ r/d, & \text{if } d > r \end{cases},$$

and $u_2(d^2)$ = $[u_1(d)]^2/\varphi$, where φ is a constant decided by ρ through $E\left[\chi_p^2 u_2\left(\chi_p^2\right)\right] = p$. The purpose of φ is to make the resulting covariance matrix estimator unbiased when $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Robust mean vector and covariance matrix can be obtained by iteratively solving

$$\boldsymbol{\mu} = \sum_{i=1}^{N} u_1(d_i)\mathbf{x}_i / \sum_{i=1}^{N} u_1(d_i), \tag{16}$$

$$\boldsymbol{\Sigma} = \sum_{i=1}^{N} u_2\left(d_i^2\right)(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' / N, \tag{17}$$

where d_i is the M-distance based on the i^{th} case. Notice that the only tuning parameter in solving Equations 16 and 17 is ρ. It is obvious that the greater the d_i the smaller the weights $u_{1i} = u_1(d_i)$ and $u_{2i} = u_2\left(d_i^2\right)$. Denote the solution to Equations 16 and 17 as $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$. Yuan et al. (2000) proposed the transformation

$$\mathbf{x}_i^{(\rho)} = \sqrt{u_{2i}}\,(\mathbf{x}_i - \hat{\boldsymbol{\mu}}). \tag{18}$$

Yuan and Hayashi (2003) suggested applying the bootstrap procedure to $\mathbf{x}_i^{(\rho)}$ and verifying the distribution of $(T|H_0)$ across the bootstrap samples against the distribution of χ_{df}^2. When data contain heavy tails, by adjusting ρ, they showed that the empirical distribution of $(T_{ML}|H_0)$ can match χ_{df}^2 very well for different real data sets.

It is obvious that $\mathbf{S}_r = \hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of $\mathbf{x}_i^{(\rho)}$. In this setup, the population covariance matrix $\boldsymbol{\Sigma}_r$ corresponding to $\mathbf{S}_r$ may not equal $\boldsymbol{\Sigma}_0$. When the data are elliptically distributed, analyzing $\mathbf{S}_r$ and $\mathbf{S}$ leads to the same substantive conclusion. In practice, data might contain outliers which will make a true symmetric distribution skewed at the sample level. In such a situation, analyzing $\mathbf{S}_r$ is preferred. If one believes that the true distribution of $\mathbf{x}$ is skewed, then the results corresponding to $\boldsymbol{\Sigma}_r$ may not be substantively equivalent to those corresponding to $\boldsymbol{\Sigma}_0$. Hampel, Ronchetti, Rousseeuw, and Stehel's (1986, p. 401) discussion implies that analyzing $\mathbf{S}_r$ might still be preferred even when $\mathbf{x}$ has a skew distribution.

The SAS IML program at www.nd.edu/-kyuan/courses/sem/RTRANS.SAS performs the iterative procedure in Equations 16 and 17. The transformed sample $\mathbf{x}_i^{(\rho)}$ in Equation 18 is printed out at the end of the program. When applying this program, one needs to modify the program lines 2 and 3 so that a proper ASCII file is correctly read. The only other thing one needs to change is the tuning parameter ρ (rho) in the main subroutine. The program also includes Mardia's multivariate skewness and kurtosis for $\mathbf{x}_i$ and $\mathbf{x}_i^{(\rho)}$. Yuan et al. (2000) suggested using the standardized Mardia's kurtosis < 1.96 to select ρ. Examples in Yuan and Hayashi (2003) implies that even when Mardia's kurtosis is not significantly different from that of a multivariate normal distribution, $(T_{ML}|H_0)$ may not approximately follow χ^2_{df}.