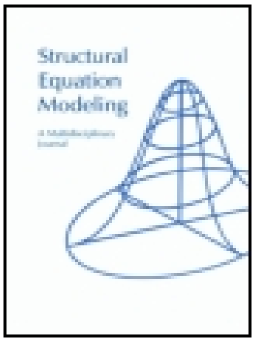

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

Reliable and More Powerful Methods for Power Analysis in Structural Equation Modeling

Ke-Hai Yuan, Zhiyong Zhang & Yanyun Zhao

To cite this article: Ke-Hai Yuan, Zhiyong Zhang & Yanyun Zhao (2017) Reliable and More Powerful Methods for Power Analysis in Structural Equation Modeling, Structural Equation Modeling: A Multidisciplinary Journal, 24:3, 315-330, DOI: 10.1080/10705511.2016.1276836

To link to this article: http://dx.doi.org/10.1080/10705511.2016.1276836

Structural Equation Modeling: A Multidisciplinary Journal, 24: 315–330, 2017
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2016.1276836

Reliable and More Powerful Methods for Power Analysis in Structural Equation Modeling

Ke-Hai Yuan,[1,2] Zhiyong Zhang,[2] and Yanyun Zhao[1]
[1]*Renmin University of China*
[2]*University of Notre Dame*

The normal-distribution-based likelihood ratio statistic $T_{ml} = nF_{ml}$ is widely used for power analysis in structural Equation modeling (SEM). In such an analysis, power and sample size are computed by assuming that T_{ml} follows a central chi-square distribution under H_0 and a noncentral chi-square distribution under H_a. However, with either violation of normality or not a large enough sample size, both empirical and analytical results indicate that the chi-square distribution assumptions are not realistic and consequently methods of power analysis based on such assumptions are not valid. This article describes a Monte Carlo (MC) method for power analysis. A measure of effect size for characterizing the power property of different rescaled statistics is also provided. Robust methods are proposed to increase the power of T_{ml} and other statistics. Simulation results show that the MC method reliably controls Type I errors and robust estimation methods effectively increase the power, and their combination is thus recommended for conducting power analysis in SEM.

Keywords: likelihood ratio statistic, Monte Carlo, power, robust estimation, Type I errors

Measurements in social and health sciences typically contain errors. By separating measurement errors from true scores or latent constructs, structural equation modeling (SEM) has become a major research methodology in many areas where the focus is on the relationships among the latent constructs as well as those between the latent constructs and their observed indicators. Like any statistical method, inference in SEM might suffer from Type I or Type II errors. Proper methods are needed in applications to minimize those errors. In particular, a power value itself does not have much meaning if the testing procedure cannot control Type I errors. Various developments have been made for power analysis in SEM. Most of them assume that the involved statistic follows a central chi-square distribution under the null hypothesis H_0, and a noncentral chi-square distribution under an alternative hypothesis H_a. The purpose of this article is to present methods for power analysis that do not rely on unrealistic chi-square distribution assumptions. In addition, we also discuss the relationship between the power of a statistical test and the underlying population distribution of the sample with respect to effect size in SEM, and propose to use robust methods for more powerful tests. A Monte Carlo (MC) study is also conducted to evaluate the proposed methods.

One of the earliest developments for power analysis in SEM was given by Satorra and Saris (1985). They considered the normal-distribution-based likelihood ratio statistic T_{ml} and assumed that $(T_{ml}|H_0) \sim \chi^2_{df}$ and $(T_{ml}|H_a) \sim \chi^2_{df}(\eta)$, where *df* is the nominal degrees of freedom, and the noncentrality parameter (ncp) η is obtained by fitting the proposed model to the population covariance matrix generated by a misspecified model under H_a. The approach to power analysis in MacCallum, Browne, and Sugawara (1996) is also based on T_{ml} following a central chi-square distribution under H_0 and a noncentral chi-square distribution under H_a, but the ncp η in their approach is specified via a hypothetical value of the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) rather than generated by a specific misspecified model. Power analysis based on T_{ml} following χ^2_{df} and $\chi^2_{df}(\eta)$ were further extended to incomplete data by Dolan, van der Sluis, and

---

Correspondence should be addressed to Ke-Hai Yuan, Department of Psychology, University of Notre Dame, Notre Dame, Indiana 46556. Email: kyuan@nd.edu

Grasman (2005) and Davey and Savla (2009). However, because normally distributed data are as rare as unicorns (Micceri, 1989), procedures based on T_{ml} following a chi-square distribution might not be valid with either complete or incomplete data in practice. In particular, with a non-normally distributed population, T_{ml} could reject a correct model 100% at the nominal level of 5% (e.g., Hu, Bentler & Kano, 1992). Actually, even for normally distributed data with a correctly specified model, T_{ml} might reject the correct model close to 100% at the nominal level of 5% when p (the number of variables) is large and N (the number of observations) is not large enough (Moshagen, 2012). Thus, more reliable methods for power analysis are needed for both normally and non-normally distributed data.

Satorra (2003) studied the power properties of three test statistics: the likelihood ratio statistic T_{ml}, the rescaled statistic T_{rml} (Satorra & Bentler, 1994), and the residual-based asymptotically distribution free statistic T_{radf} (Browne, 1984). Each statistic is assumed to follow a central chi-square distribution under H_0 and a noncentral chi-square distribution under H_a. Limited MC results in Satorra (2003) indicated that the empirical distribution of T_{rml} is reasonably approximated by its asymptotic distribution under certain conditions but not always, and the distribution of T_{radf} is markedly different from its nominal asymptotic chi-square distribution at smaller N. The distributions of T_{radf} and T_{rml} under H_0 were also examined in Nevitt and Hancock (2004), Yuan and Bentler (1998a), and Bentler and Yuan (1999). The results indicate that the nominal chi-square distribution χ^2_{df} cannot describe the empirical behavior of T_{rml} well unless the model structure and the population fourth-order moments satisfy certain conditions (Yuan & Bentler, 1999) together with a large enough sample size N; and T_{radf} deviates significantly from χ^2_{df} unless N is very large. Thus, statistics designed to account for non-normality are not good candidates for conducting power analysis in practice when they are compared against the nominal chi-square distributions.

Muthén and Muthén (2002), Schoemann, Miller, Pornprasertmanit, and Wu. (2014), and Zhang (2014) proposed using MC methods to evaluate power in testing the hypotheses on population parameters in SEM models (see also Wolf, Hamington, Clark, & Miller, 2013). A key element in their proposals is to estimate the power through MC simulation. The model is estimated for each sample generated under a given population distribution, and the z score or Wald statistic for the parameter under the hypothesis is then compared against $N(0,1)$ or a central chi-square distribution, yielding either a rejection or no rejection. The power is estimated as the proportion of rejections across many replications. The methods would be sound if the z or Wald statistic literally follows $N(0,1)$ or the central chi-square distribution under the null hypothesis. However, empirical results indicate that standard errors (SEs) obtained via the information matrix, sandwich-type covariance matrix, or other consistent formula can be a lot smaller than the corresponding empirical SEs when the sample size is not large enough, especially when the underlying population distribution is of heavy tails (see, e.g., MacKinnon & White, 1985; Yuan & Bentler, 1997b). Also, sandwich-type SEs as given in standard programs might not be consistent (Yuan & Hayashi, 2006) unless the model is correctly specified. The idea of estimating the power by comparing the values of z score against $N(0,1)$ parallels that in Yung and Bentler (1996), who proposed estimating power for testing the overall structure of an SEM model by comparing $(T_{ml}|H_a)$ against the nominal chi-square distribution, which would be a sound method if $(T_{ml}|H_0)\sim\chi^2_{df}$. However, many conditions can render the $N(0,1)$ or χ^2_{df} invalid, including non-normally distributed population, not large enough sample size, or incomplete data.

A power evaluation approach that does not rely on the assumptions of $(T|H_0)\sim\chi^2_{df}$ or $(T|H_a)\sim\chi^2_{df}(\eta)$ was given by Yuan and Hayashi (2003), where the critical value corresponding to the distribution of $(T|H_0)$ is estimated by the method of bootstrap, and power is also estimated by the bootstrap methodology in which the value of $(T|H_a)$ is estimated at each replication and compared against the estimated critical value. Because the bootstrap methodology is used, this approach is conditional on a given sample. Yuan and Hayashi (2003) described this approach using a real complete data set. In this article, we extend this approach to power analysis by MC simulation when a sample is not available, and call it the MC method. The MC method resembles the methodology of parametric bootstrap (Efron & Tibshirani, 1993, Section 6.5), and it allows us to include features of the expected population distribution to be studied. But we do not need to know the exact form of the distribution of the target population. In particular, for any statistic T with Type I error and power defined by

$$\alpha = P(T > c_{1-\alpha}|H_0) \text{ and power} = P(T > c_{1-\alpha}|H_a),$$

we estimate the critical value $c_{1-\alpha}$ and the value of power using MC simulation. Thus, we do not need to assume a distribution on either $(T|H_0)$ or $(T|H_a)$ and can still obtain a consistent estimate of power by controlling Type I error at level α. We conduct a simulation study of the MC method and contrast different approaches to power analysis in SEM.

The true power of a test statistic[1] is defined once the model, sample size, and a target population are given. The

---

[1] A test statistic is a numerical value aiming to optimally summarize the deviance in the data against the null hypothesis, and the statistic typically depends on the values of parameters computed by a particular estimation method (see, e.g., https://en.wikipedia.org/wiki/Test_statistic).

MC method allows us to consistently estimate the true power. However, the true power varies when a different statistic or an alternative estimation method is chosen. Methods for power analyses in the literature are mostly based on the analysis of sample means and covariances by assuming data are from normally distributed populations. For example, the *t*, and *F* statistics for analysis of vanance (ANOVA) and regression, the Fisher *z* statistic for the product-moment correlation, and the statistic T_{ml} for SEM are all based on the normality assumption, which is idealized rather than realistic (Micceri, 1989). When the underlying population distribution is of heavy tails, inaccuracy in power estimation for testing hypotheses on means mostly occurs at smaller sample sizes, because the *t* and *F* statistics are asymptotically robust to distribution violations. However, T_{ml} or other test statistics in the context of SEM could suffer from severe loss of true power if the underlying population distribution is of heavy tails. As a matter of fact, regardless of how wrong a model is, T_{ml} or the Fisher *z* statistic at a given sample size might have a power close to zero. Because real data tend to have heavier tails than that of the normal distribution, we propose to apply robust methods for achieving better power. We also use MC simulation to compare the power of test statistics following a robust method against that under the analysis of sample variances–covariances.

There are many test statistics in SEM, and we mainly study the power of T_{ml} and T_{rml}, but the methods and ideas described here can be equally applied to other statistics as well as to evaluating the power for testing parameters. We discuss power properties of T_{ml} and T_{rml} under different estimation methods[2] in the following section. A simulation study evaluating the validity of the MC method and comparing the power properties of different statistics is presented in the following section. Conclusions, discussion, and recommendations are provided in the concluding section.

Power analysis might be performed in the planning stage of a project before any data are collected or when data are partially or completely collected. The analysis in the planning stage is mostly for predicting the minimum sample size needed to achieve desired results, and is called proactive, whereas analysis after data were collected is called reactive and can also be useful (see Marcoulides & Chin, 2013). Although the focus of this article is the proactive approach for power analysis, our emphasis is on methods to control Type I error and increase statistical power. The procedures described are equally applicable to power analysis reactively. In the concluding section, we also discuss how to use the information from existing data sets in the planning stage of a related or a new project.

TEST STATISTICS, ROBUST METHODS, EFFECT SIZE, AND THE MONTE CARLO METHOD

In this section we first discuss properties of T_{ml} and T_{rml} defined via the method of normal-distribution-based maximum likelihood (NML). Power loss of the two statistics when the population distribution is of heavy tails is described next. We then introduce robust methods and discuss power properties of T_{ml} and T_{rml} defined via robust estimates. The concept of effect size is also introduced and compared across different estimation methods. Methods for handling incomplete data and how T_{ml} and T_{rml} are computed will also be reviewed, to set up the context for MC study with missing data in the following section. Because, beyond the class of elliptical distributions, the distribution of T_{ml} or T_{rml} does not have a simple form even asymptotically, we mostly refer to elliptical distributions when discussing the effect of heavy-tailed distribution on the power of the two statistics in this section. Our MC study in the following section includes both symmetric and skewed population distributions.

Test Statistics T_{ml} and T_{rml}

Many test statistics have been developed in the SEM literature (see, e.g., Yuan & Bentler, 2007). The most widely used one is the likelihood ratio statistic T_{ml} derived from NML. Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ be a random sample from a p-variate population represented by $\mathbf{x}$ with $E(\mathbf{x}) = \boldsymbol{\mu}$ and $Cov(\mathbf{x}) = \boldsymbol{\Sigma}$; $\mathbf{S}$ be the sample covariance matrix; and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be the structural model. The NML method for SEM is commonly presented via the ML discrepancy function

$$F_{ml}(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) = \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})) - \log|\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})| - p. \qquad (1)$$

Let $\hat{\boldsymbol{\theta}}$ be the value of $\boldsymbol{\theta}$ that minimizes $F_{ml}(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, then $T_{ml} = nF_{ml}(\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))$, where $n = N - 1$. Under the normality assumption and $H_0{:}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, T_{ml} converges to the nominal chi-square distribution χ^2_{df} as $N \to \infty$. However, T_{ml} does not approach χ^2_{df} when the normality assumption fails to hold. When data are elliptically distributed with relative kurtosis

$$\beta = E\{[(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})]^2\}/[p(p+2)], \qquad (2)$$

T_{ml} converges to $\beta\chi^2_{df}$ under H_0. This implies that Type I error rate of statistical inference, according to $T_{ml} \sim \chi^2_{df}$, is directly related to the value of β. When β is large enough, we reject a correct model 100% by comparing T_{ml} against the critical value of χ^2_{df} although the aim is at 5%. For example, if the population follows a multivariate *t*-distribution with degrees of freedom m (i.e., $\mathbf{x} \sim t(m, \boldsymbol{\mu}, \boldsymbol{\Sigma})$), then

---

[2] As the estimation method varies, the properties of T_{ml} and T_{rml} also vary, and we label the two statistics with additional notation when they are evaluated under robust methods.

its relative kurtosis is given by $\beta = (m-2)/(m-4)$. Thus, if the population distribution follows $t(m, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the statistic $T_{ml} \sim \chi^2_{df}$ will reject the correct model 100% when m is either less than 4 or greater than but close to 4. In such a case, results of power analysis based on $(T_{ml}|H_0) \sim \chi^2_{df}$ and $(T_{ml}|H_a) \sim \chi^2_{df}(\eta)$ are misleading because Type I error and power are confounded. A larger value of the β in Equation 2 is reflected by heavier tails in the observed data, and can be measured by multivariate kurtosis (Bentler, 2006, p. 106; Cain, Zhang, & Yuan, in press; Mardia, 1970.)

More generally, when the population **x** follows a distribution with finite fourth-order moments,

$$T_{ml} = \sum_{j=1}^{df} \beta_j z_j^2 + o_p(1), \tag{3}$$

where $o_p(1)$ is a term that approaches zero in probability, the values of β_j, $j = 1, 2, \ldots, df$ are determined by the model structure and the population fourth-order moments of **x**; and z_j are independent and each follows the standard normal distribution $N(0,1)$. Thus, the asymptotic mean of T_{ml} is given by $\sum_{j=1}^{df} \beta_j$. Consequently, T_{ml} tends to reject correct models more often when $\sum_{j=1}^{df} \beta_j$ is greater than df. Again, results of power analysis based on $(T_{ml}|H_0) \sim \chi^2_{df}$ and $(T_{ml}|H_a) \sim \chi^2_{df}$ are misleading.

Let $\hat{\beta}$ be a consistent estimator of the average $\bar{\beta} = \sum_{j=1}^{df} \beta_j / df$, then the rescaled statistic $T_{rml} = \hat{\beta}^{-1} T_{ml}$ converges to a distribution with mean equal to df. Consequently, T_{rml} is better approximated by χ^2_{df} than T_{ml} when $\bar{\beta} \neq 1.0$. However, it does not mean that T_{rml} literally follows a chi-square distribution. In particular, when there are substantial differences among the $\beta_j, j = 1, 2, \ldots, df$, then the distribution of T_{rml} can be far from χ^2_{df} even asymptotically (Yuan & Bentler, 1999). Power analysis based on $T_{rml} \sim \chi^2_{df}$ can also be misleading, especially when the coefficients β_j differ substantially or when the sample size is not large enough (Bentler & Yuan, 1999; Nevitt & Hancock, 2004; Yuan & Bentler, 1998a).

Power Loss of T_{ml} and T_{rml} with Heavy-Tailed Population Distributions

Because T_{ml} tends to reject a correct model when the underlying population distribution is of heavier tails, we might think that T_{ml} will have a desired power value in general. However, the story will change if we control Type I errors, because the power of T_{ml} is determined by the separation between $(T_{ml}|H_0)$ and $(T_{ml}|H_a)$, which further depends on the variance of T_{ml} in addition to the size of model misspecification. For a give model under H_a, the size of misspecification is measured by

$$F_{mla} = F_{ml}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_a)), \tag{4}$$

where $\boldsymbol{\theta}_a$ is the value of $\boldsymbol{\theta}$ that minimizes $F_{ml}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. Thus, the size of model misspecification is determined once the model is given, and the overlap between $(T_{ml}|H_0)$ and $(T_{ml}|H_a)$ is then determined by the distribution of the underlying population in addition to sample size. As an example, consider **x** following an elliptical distribution. Then the asymptotic mean and variance of $(T_{ml}|H_0)$ are given by βdf and $2\beta^2 df$, respectively; and those of $(T_{ml}|H_a)$ are given by $\beta df + \eta$ and $2\beta^2 df + 4\beta\eta$, respectively, where

$$\eta = nF_{mla} \tag{5}$$

is no longer the noncentrality parameter because $(T_{ml}|H_a)$ does not follow a noncentral chi-square distribution (see, e.g., Shapiro & Browne, 1987). Thus, the mean difference between $(T_{ml}|H_a)$ and $(T_{ml}|H_0)$ is affected little by β while their standard deviations increase as β increases. Consequently, there is less separation between $(T_{ml}|H_a)$ and $(T_{ml}|H_0)$ as β increases. The analysis implies that the true power of T_{ml} decreases as the kurtosis of the population distribution increases.

Power loss due to heavy-tailed distribution also occurs to the rescaled statistic T_{rml}. Consider again **x** following an elliptical distribution, then $(T_{rml}|H_0)$ and $(T_{rml}|H_a)$ asymptotically follow χ^2_{df} and $\chi^2_{df}(\eta/\beta)$, respectively (Shapiro & Browne, 1987). Consequently, the mean difference between $(T_{rml}|H_a)$ and $(T_{rml}|H_0)$ approximately equals η/β, which goes to zero as β increases. Thus, although the asymptotic variance of $(T_{rml}|H_a)$, equal to $2(df + 2\eta/\beta)$, decreases with β, there will be little separation between $(T_{rml}|H_0)$ and $(T_{rml}|H_a)$ if β is rather large. The statistic T_{rml} eventually lacks power in detecting any misspecified model as β increases.

A parameter test in SEM following the NML method faces the same issue of power loss. Although sandwich-type SEs provide consistent estimates of the variability of parameter estimates, the SEs increase as kurtosis increases. Consider testing $H_0 : \theta = \theta_0$ based on the NML estimate $\hat{\theta}$ of θ. Assume that the population counterpart of $\hat{\theta}$ is θ_a; the sandwich-type SE of $\hat{\theta}$ is estimated as $\hat{\tau}/\sqrt{n}$, and the population counterpart of $\hat{\tau}$ is τ. Then τ is proportional to $\sqrt{\beta}$ under an elliptically distributed population with kurtosis β (Tyler, 1983). If we define effect size as $|\theta_a - \theta_0|/\tau$, then it decreases to zero as β increases. Thus, a parameter test based on sandwich-type covariance matrix will lose its power when the population distribution is of heavy tails.

It is hard to give a precise quantification of power loss of T_{ml} and T_{rml} when the population distribution is not elliptical, as the β_js in Equation 3 are no longer equal in general. We would expect that their power values for a given η to decrease as the average $\bar{\beta}$ of the β_js increases. We use the MC method to evaluate their power in the next section.

Robust Methods and Statistical Power

Closely related to statistical power is efficiency of parameter estimates. A method yields more efficient parameter estimates also corresponds to higher statistical power in general when the estimates are used to construct test statistics. This is because the resulting test statistics $(T|H_0)$ and $(T|H_a)$ will have less variation and thus more separation if their mean difference remains about the same. It is known that the true ML estimates are asymptotically most efficient, and NML estimates can be very inefficient when data possess extra kurtoses than that of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In practice, with real data typically having heterogeneous marginal kurtoses (e.g., Blanca, Arnau, Lipez-Montiel, Bono, & Bondayan, 2013; Cain et al., in press; Micceri, 1989), it is unlikely for us to obtain the true ML estimates. Nevertheless, ML based on a distribution that can accommodate heavy tails in real data will lead to more efficient parameter estimates in general. A class of distributions that can account for heavy tails is multivariate t-distributions, and t-distribution-based ML (TML) has been shown to yield more reliable results than NML when modeling real multivariate data (see, e.g., Lange, Little, & Taylor, 1989; Little, 1988; Tong & Zhang, 2012; Wilcox, 2012; Yuan & Bentler, 1998b).

A more general method than TML is robust M-estimation, which aims to generalize ML methodology in general (Huber, 1967). The key idea in robust M-estimation is that observations geometrically sitting far from the center of the majority of the data are down-weighted, and different down-weighting schemes correspond to different M-estimators. Many weighting functions have been proposed according to the size of the Mahalanobis distance:

$$d_i^2 = (\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x_i} - \boldsymbol{\mu}).$$

In particular, the weight corresponding to TML is given by $w_i = w(d_i) = (m + p)/(m + d_i^2)$, where m is the degrees of freedom of the t-distribution and p is the number of variables. The contribution of $\mathbf{x}_i$ to the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ or the structural parameter in TML is proportional to w_i. Clearly, the weight w_i assigned to the ith case becomes smaller as d_i increases. In particular, once w_i is given, the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be regarded as the weighted average of $\mathbf{x}_i$ and $(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x_i} - \boldsymbol{\mu})'$, respectively. In the estimation process, the value of w_i will be assigned automatically and iteratively in the process of estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ or the parameters in the structural model (Yuan & Bentler, 1998b). The degrees of freedom, m, can be either chosen or estimated. Unless the population is truly a multivariate t-distribution, fixing m at a given value might be preferred because the resulting estimation is a lot easier (Little, 1988). Also, empirical studies suggested that using a weight function to down-weight extreme cases is more important than its precise form (Little, 1988; Wilcox, 2012; Yuan & Zhong, 2008).

Another popular weight is the Huber-type weights (Huber, 1981). Let r^2 be the quantile of the distribution of χ_p^2 corresponding to probability $(1 - \alpha)$. The Huber-type weights are given by

$$\begin{aligned} w_{i1} &= w_1(d_i) \\ &= \begin{cases} 1, & \text{if } d_i \le r, \\ r/d_i, & \text{if } d_i > r, \end{cases} \quad \text{and } w_{i2} = w_{i1}^2/\kappa \end{aligned} \qquad (6)$$

where κ is a constant such that $E[\chi_p^2 w_{i1}^2(\chi_p)/\kappa] = p$, with κ being determined by α so that the resulting estimates of means and covariances or the structural parameters are consistent when $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Different from weights derived from a multivariate t-distribution, weight w_{i1} is applied to case $\mathbf{x}_i$ in estimating the means $\boldsymbol{\mu}$, and weight w_{i2} is applied to $(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'$ in estimating the covariance matrix $\boldsymbol{\Sigma}$. Under Huber-type weights, only cases having $d_i > r$ are down-weighted. A larger value of α corresponds to a smaller r, and consequently more cases are down-weighted. In the estimation process, one only needs to choose a value of α and then the weights w_{i1} and w_{i2} as well as κ will be adjusted automatically in the estimation process (Yuan & Zhong, 2008).

After robust estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ being obtained, SEM can be performed by replacing the $\mathbf{S}$ in Equation 1 with $\hat{\boldsymbol{\Sigma}}$ and proceeding to estimate the structural parameters $\boldsymbol{\theta}$ by minimizing the corresponding discrepancy function F_{ml}. Similarly, one can also substitute $\hat{\boldsymbol{\mu}}$ for $\bar{\mathbf{x}}$ when a mean structure is involved. Yuan and Bentler (1998b) showed that the resulting robust estimates $\hat{\boldsymbol{\theta}}$ will inherit the robust properties of $\hat{\boldsymbol{\Sigma}}$. In particular, a robust version of the rescaled statistic, $T_{rml}^{(r)}$, can be obtained so that its asymptotic mean equals df. Yuan, Chan, and Bentler (2000) also showed that the test statistic $T_{ml}^{(r)}$ by treating $\hat{\boldsymbol{\Sigma}}$ as $\mathbf{S}$, or a robust version of T_{ml}, also more closely follows χ_{df}^2 than T_{ml} for typical real data with heavy tails. Because the power of a test statistic is closely related to the efficiency of the parameter estimates and robust methods tend to yield more efficient estimates, both $T_{ml}^{(r)}$ and $T_{rml}^{(r)}$ are expected to have better power than T_{ml} and T_{rml} (see Tyler, 1983; Yuan, Bentler, & Chan, 2004; Yuan & Hayashi, 2003). We use the MC method to evaluate the power properties of $T_{ml}^{(r)}$ and $T_{ml}^{(r)}$ and contrast them against those of T_{ml} and T_{rml}. Again, with the MC method, we do not need to make a central or noncentral chi-square distribution assumption on either $T_{ml}^{(r)}$ or $T_{rml}^{(r)}$.

Athough robust M-estimation has been shown to yield more reliable results than NML in analyzing real data in previous studies, the advantage of a robust method also depends on the population distribution underlying the sample. If the population distribution is multivariate normal, then a robust method might not perform as well as NML although the difference is typically minor. Differences between NML and robust methods can be characterized asymptotically when **x** follows an elliptical distribution. A model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is called invariant under a constant scaling factor (ICSF) if for any parameter vector $\boldsymbol{\theta}$ and a constant c there exists an $\boldsymbol{\theta}_c$ such that $c^2\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}_c)$. As noted in Browne (1984), essentially all interesting models are ICSF. Yuan et al. (2004) showed that, for a model that is ICSF, the rescaled statistic $T_{rml}^{(r)}$ asymptotically follows $\chi^2_{df}(\eta/\beta_r)$, where the value of β_r is determined by the estimation method and the underlying elliptical distribution. When data are normally distributed and NML is chosen, $\beta_r = \beta = 1$. Tyler (1983) studied robust M-estimation for covariance matrices and examined the properties of a rescaled statistic for testing hypothesis H_0: $\boldsymbol{h}(\boldsymbol{\Sigma}) = \mathbf{0}$, where $\boldsymbol{h}(\cdot)$ is a vector of smooth functions defined over the covariance matrix $\boldsymbol{\Sigma}$. For an SEM model that is ICSF, the scaling factor used in Tyler (1983) has the same asymptotic value as β_r that governs the asymptotic distribution of $T_{rml}^{(r)}$ for testing the structural model $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ (Yuan et al. 2004). Part of the results of Tyler (1983) for the value of β_r is shown in the Table 1 for us to see the effect of robust methods, where the populations are 10-variate t-distribution with 1 and 5 degrees of freedom, and normal distribution, respectively; and the estimation methods are TML based on 1 and 5 degrees of freedom, robust M-estimation using Huber-type weights with the tuning parameter r in Equation 6 corresponding to $\alpha = .1$, .5, and 0 (NML), respectively. When the underlying population distribution is normal, NML yields most efficient estimates with $\beta_r = 1$, which has the largest ncp possible for the given size of model misspecification as measured by the η in Equation 5. Under this ideal condition, the ncp corresponding to Huber (.1) is only 1% smaller $(1/1.01 = .99)$ than the η defined in Equation 5. When the underlying population is multivariate t-distribution with $m = 5$, TML(5) performs best with $\beta_r = 1.13$ and the resulting ncp is 12% smaller $(1/1.13 = .88)$ than η; Huber(.5) corresponds to $\beta_r = 1.15$ and the resulting ncp is 87% of η $(1/1.15 = .87)$. In contrast, NML corresponds to $\beta_r = 3.0$, which makes the ncp only 33% of the η defined in Equation 5. When the underlying population is multivariate t-distribution with $m = 1$ degree of freedom, NML corresponds to $\beta_r = \infty$ and consequently T_{ml} or T_{rml} becomes powerless regardless of sample size or the size of misspecification as measured by η. In contrast, $T_{rml}^{(r)}$ following TML(1) corresponds to $\beta_r = 1.18$, and the resulting ncp is still 85% of η $(1/1.18 = .85)$, and Huber (.5) corresponds to an ncp that is 81% of η $(1/1.23 = .81)$. Note that, with $\mathbf{x} \sim t(1, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, ncp $= \eta/1.18$ is the largest possible value of ncp a method can achieve because, TML (1) yields asymptotically most efficient estimates. Athough Huber-type weights are not optimal, the ncp under Huber (.5) is only 4% $(1.18/1.23 = .96)$ smaller than the largest possible value of ncp under one of the worst conditions.

TABLE 1
The Value of β_r With Three Population Distributions and Five Estimation Methods ($p = 10$, Tyler, 1983)

	Estimation Method				
Population Distribution	TML(1)	TML(5)	Huber(.1)	Huber(.5)	NML
$t(1, \boldsymbol{\mu}, \boldsymbol{\Sigma})$	1.18	1.28	1.48	1.23	∞
$t(5, \boldsymbol{\mu}, \boldsymbol{\Sigma})$	1.17	1.13	1.21	1.15	3.00
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	1.16	1.09	1.01	1.08	1.00

Note. TML= t-distribution-based maximum likelihood; NML=normal-distribution-based maximum likelihood.

Effect Size in SEM

As is well known, the power of a statistical test is closely related to sample size and effect size. Effect size in ANOVA, regression, or analysis of correlation has been well documented, but not in SEM. We give a definition of effect size for SEM in this subsection and discuss its value by relating it to the material presented in the previous subsections.

For testing $H_0: \mu = 0$ based on a sample of size n from a normally distributed population $N(\mu, \sigma^2)$ with a known σ^2, the effect size is defined as $\delta = \mu/\sigma$ (assuming μ>0). The test statistic $z = \sqrt{n}\bar{x}/\sigma$ follows $N(\sqrt{n}\delta, 1)$ or equivalently z^2 follows $\chi^2_1(n\delta^2)$, where $n\delta^2$ is the ncp. Thus, the effect size is simply $(\text{ncp}/n)^{1/2}$. For SEM with normally distributed data, both T_{ml} and T_{rml} approximately follow $\chi^2_{df}(\eta)$. When the population distribution is normal, a direct imitation of the effect size from the literature on mean comparison leads us to defining effect size as

$$\text{ES}_{ml} = (\text{ncp}/n)^{1/2} = F_{mla}^{1/2}, \tag{7}$$

where F_{mla} is defined in Equation 4. Note that the effect size in Equation 7 is defined at the population level, not related to the sample size $N = n + 1$.

The material in the previous subsections implies that the definition of effect size in Equation 7 is applicable only for T_{ml} and T_{rml} with a normally distributed population. When **x** follows a distribution other than normal, $(T_{ml}|H_a)$ does not follow a noncentral chi-square distribution in general, and the definition of effect size in Equation 7 is no longer valid. When **x** follows an elliptical distribution with relative kurtosis β, and $(T_{rml}|H_0)$ and $(T_{rml}|H_a)$ still asymptotically follow chi-square distributions, we can define effect size as

$$\mathrm{E}S_{rml} = (F_{mla}/\beta)^{1/2}, \quad (8)$$

which is inversely proportional to the relative kurtosis. When the population distribution of $\mathbf{x}$ is multivariate t with less than 4 degrees of freedom, then the effect size in Equation 8, associated with T_{rml}, would be zero regardless of the value of F_{mla}. In contrast, the effect size associated with $T_{rml}^{(r)}$ following a robust method would be

$$\mathrm{E}S_{rml}^{(r)} = (F_{mla}/\beta_r)^{1/2}. \quad (9)$$

As listed in Table 1, even if $\mathbf{x} \sim t(1, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the effect size corresponding to Huber(.5) is still 90% $(= 1/\sqrt{1.23})$ of the effect size corresponding to a normally distributed population. Thus, effect sizes following a robust method can be substantially greater than those following the widely used NML with real data typically having heavy tails.

When the population is not elliptically distributed, T_{rml} or $T_{rml}^{(r)}$ does not follow a central or noncentral chi-square distribution in general. However, the values of the power of T_{rml} and $T_{rml}^{(r)}$ are still inversely affected by the heaviness of the tails in the population distribution. Considering that T_{rml} and $T_{rml}^{(r)}$ are obtained by rescaling T_{ml} and $T_{ml}^{(r)}$, respectively, we can generalize the effect sizes in Equations 8 and 9 to

$$\mathrm{E}S_{rml} = (F_{mla}/\bar{\beta})^{1/2} \text{ and } \mathrm{E}S_{rml}^{(r)} = (F_{mla}/\bar{\beta}_r)^{1/2}, \quad (10)$$

where $\bar{\beta}$ is the average of the β_js in Equation 3 and $\bar{\beta}_r$ is the counterpart of $\bar{\beta}$ corresponding to $T_{ml}^{(r)}$. Note that the power of T_{rml} or $T_{rml}^{(r)}$ might only be approximately determined by the effect size in Equation 10 and sample size when the two statistics do not follow chi-square distributions. However, we expect that $\bar{\beta}_r$ corresponding to $T_{rml}^{(r)}$ to be much smaller than $\bar{\beta}$ corresponding to T_{rml} under NML, and thus $T_{rml}^{(r)}$ to be a more powerful test statistic in general when the underlying population distribution of $\mathbf{x}$ is of heavy tails. We use MC simulation to study the power properties of T_{rml} and $T_{rml}^{(r)}$ in the next section.

Incomplete Data

We have discussed the issue of statistical power in SEM with complete data via the test statistics T_{ml}, T_{rml}, $T_{ml}^{(r)}$, and $T_{rml}^{(r)}$. Each of the statistics also has a corresponding version when a sample contains missing values. In particular, estimates of $\boldsymbol{\Sigma}$ can be obtained using the normal-distribution-based expectation-maximization (EM) algorithm (Dempster, Laird & Rubin, 1977; Enders & Peugh, 2004) or an expectation-robust (ER) algorithm (Little & Smith, 1987; Yuan, Chan, & Tian, 2016). The incomplete-data version of test statistics T_{ml} and $T_{ml}^{(r)}$ are obtained when the $\mathbf{S}$ in Equation 1 is replaced by the corresponding NML and robust estimates $\hat{\boldsymbol{\Sigma}}$, respectively. The advantage of such a two-stage approach is that rescaled versions of T_{ml} and $T_{ml}^{(r)}$ have been developed for data with missing values (Yuan & Bentler, 2000; Yuan & Zhang, 2012), and they perform well in practice (Tong, Zhang & Yuan, 2014). We expect that the power of each of the test statistics will be inversely affected by heavy-tailed distribution, and will use MC simulation to study them in the next section.

An alternative test statistic is the likelihood ratio statistic assuming the observed marginal variables of each incomplete case follow a normal distribution corresponding to a subset of the p complete variables, called direct ML or full information maximum likelihood (FIML) in the SEM literature (Savalei & Bentler, 2009; Savalei & Falk, 2014). However, such an obtained statistic cannot be expressed in the form of the discrepancy function in Equation 1. Consequently, it is not clear how to define effect size for such a statistic or its rescaled version. Also, with normally distributed data as rare as unicorns in practice (Micceri, 1989), direct ML might not have any advantage over the two-stage approach. In addition, the two-stage approach facilitates the inclusion of auxiliary variables, whereas it is not clear how to include auxiliary variables in direct ML without affecting the evaluation of the model for the substantive variables (Savalei & Bentler, 2009; Yuan & Lu, 2008). Thus, although our proposed method also applies to statistics evaluated following direct ML estimation, we do not study the direct likelihood ratio statistic with incomplete data in this article.

The Monte Carlo Method for Estimating Power and Sample Size

Let T be a statistic defined under NML, robust M-estimation, or any other estimation method. For a sample of size N drawn from a target population under H_0 with complete or incomplete data, a value of the statistic T is obtained at the end of the parameter estimation. We next obtain N_r independent replications of the process of computing T, and then order the obtained values of T from small to large, and denote them as

$$T_{(1)} \le T_{(2)} \le \cdots \le T_{(N_r)}.$$

For a given level α, let the integer part of $N_r(1-\alpha)$ be denoted as $N_{r(1-\alpha)}$. For example, with $\alpha = .05$ and $N_r = 1{,}000$, $N_{r(1-\alpha)} = 950$. The estimated critical value for T under the condition (e.g., sample size, population distribution, missing data scheme, etc.) is

$$\hat{c}_{1-\alpha} = T_{(N_{r(1-\alpha)})}.$$

We then draw an independent sample of the same size from the target population under H_a, compute the value of T under the same estimation method by fitting an incorrectly specified model, and compare the obtained value of T against $\hat{c}_{1-\alpha}$. The statistic is counted as non-significant if $T \leq \hat{c}_{1-\alpha}$ and significant otherwise. Repeating the process N_a times, the percentage of significant T will be the estimated value of power under the MC method. In particular, the same process will generate an estimate of Type I error if we draw independent samples from the same population as that under H_0.

The value of power by the MC method, denoted as $\hat{p}_a$, might be smaller or greater than the desired value p_a. If $\hat{p}_a < p_a$, then increase the sample size N and repeat the processes of computing $\hat{c}_{1-\alpha}$ and $\hat{p}_a$. Otherwise, repeat the process by decreasing the sample size N. The estimated sample size N is obtained when $\hat{p}_a$ is close enough to p_a, say $|\hat{p}_a - p_a| < .01$. In the implementation of the MC method, the sample size might be increased or decreased by a step size of 5 initially, and then by a step size of 1 when $\hat{p}_a$ becomes closer to p_a. The iterative procedure of estimating N can be automatic and researchers just need to provide the desired value of p_a, the estimation method, and features of the population distribution (e.g., marginal skewnesses and kurtoses, or the multivariate kurtosis) in addition to the structural model.

SIMULATION STUDY

In this section, we use MC simulation to examine the MC method and the properties of test statistics T_{ml}, T_{rml}, $T_{ml}^{(r)}$, and $T_{rml}^{(r)}$. Our main interests are (a) how reliable the MC method is in controlling Type I errors when compared to the conventional methods, and (b) which statistic is most powerful with typically non-normally distributed data in practice.

Design

The population is generated according to a confirmatory factor model with nine variables and three factors. The path diagram of the model is given in Figure 1, where factor loadings λ_2, λ_3, λ_5, λ_6, λ_8, and λ_9, the nine error variances, and the three factor variances are all set at 1.0; and the three factor covariances are $\phi_{12} = .5$, $\phi_{13} = .3$, and $\phi_{23} = .4$. The particular values of the parameters in the population are not material because, for a given N, the power of a test statistic is mainly determined by the size of misspecification and the population distribution as measured by the effect sizes in Equation 7 to 10. The three loadings represented by the dashed lines in Figure 1 are all equal ($a = b = c$), and are set as 0 for the condition of H_0, and are set as .2, .4, and .6, respectively for three conditions of H_a. Thus, there are four different population covariance matrices $\mathbf{\Sigma}$ in the study.

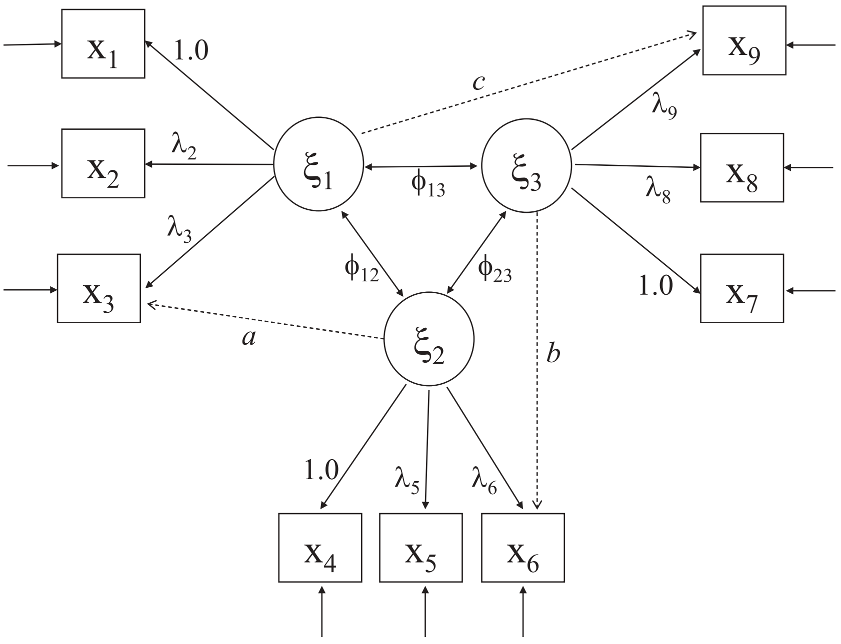

FIGURE 1 A path diagram for the model that generated the population.

Two population distributions are included in the study. The first is multivariate normal, and the second is generated according to

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}/u,$$

where $\boldsymbol{\Sigma}^{1/2}$ is a symmetric matrix such that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$; $\mathbf{z} = (z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8, z_9)'$ with z_1, z_2, z_4, z_5, z_7, z_8 following the standard normal distribution $N(0,1)$; z_3 following standardized χ^2_1, z_6 following standardized χ^2_3, and z_9 following standardized χ^2_5; $u \sim \sqrt{\chi^2_5/3}$; and z_1 to z_9 and u are all independent. Because $E(3/\chi^2_5) = 1$ and $Cov(\mathbf{x}) = E(1/u^2)\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$, the normal and non-normal population distributions have the same covariance matrix. With additional statistical and algebraic computation following Equation 2, it can be shown that the relative kurtosis for the non-normal population is $\beta \approx 3.558$. Thus, we would expect T_{ml} to have a higher Type I error rate than the nominal level when compared against χ^2_{df} for inference. We would also expect the power values of T_{ml} and T_{rml} to be smaller than those of $T^{(r)}_{ml}$ and $T^{(r)}_{rml}$.

At each of the population conditions (four conditions of $\boldsymbol{\Sigma}$ and two conditions of population distribution), we have eight conditions on sample size $N = 100$, 150, 200, 300, 500, 1,000, 2,000, 3,000, which can be regarded as from small to large.

A condition of incomplete data is also included. For each sample, variables x_1, x_2, x_4, x_5, x_7, and x_8 are always observed. Variables x_3, x_6, and x_9 contain missing values that are generated according to: x_3 is missing if $(x_1 + x_2) < c_1$, x_6 is missing if $(x_4 + x_5) < c_2$, and x_9 is missing if $(x_7 + x_8) < c_3$, where the values of c_1, c_2, and c_3 are controlled so that x_3, x_6, and x_9 each is missing 10%. Thus, there are eight observed patterns, and all the non-observed values are missing at random (MAR; Rubin, 1976).

The same nine-variable, three-factor model, as implied by the path diagram in Figure 1, is used to fit all the samples in our study, where the three dashed lines are not included in the model. Thus, the nominal degrees of freedom for the four statistics are $df = 24$. Both NML and robust methods are used to estimate the model. For each complete sample, NML is carried out via the Fisher-scoring (FS) algorithm in estimating the parameters $\boldsymbol{\theta}$ in the structural model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, and statistics T_{ml} and T_{rml} are subsequently obtained. The robust method using Huber-type weights in Equation 6 with r^2 being set at the 95th quantile of χ^2_9 is used to obtain $\hat{\boldsymbol{\Sigma}}$, the FS algorithm is then used to obtain the estimate of the parameters in the structural model, and statistics $T^{(r)}_{ml}$ and $T^{(r)}_{rml}$ are subsequently computed at the robust $\hat{\boldsymbol{\theta}}$. For each incomplete sample, NML is carried out first by the EM algorithm to obtain $\hat{\boldsymbol{\Sigma}}$ and then by the FS algorithm to obtain $\hat{\boldsymbol{\theta}}$, and subsequently T_{ml} and T_{rml}; robust estimation is obtained similarly to that for complete data by the Huber-type weights, where the tuning parameter r^2 in (6) is set at the 95th quantile of $\chi^2_{p_i}$, with p_i being the number of observed variables in the ith case (Yuan et al., 2016).

With 1,000 replications[3] of a statistic T ($= T_{ml}$, T_{rml}, $T^{(r)}_{ml}$ or $T^{(r)}_{rml}$) under H_0, our estimated critical value is

$$\hat{c}_{.95} = T_{(950)},$$

aiming to control Type I error at 5%. For each condition of H_0 or H_a, we independently draw another 1,000 replications of a sample at the same size and consequently 1,000 replications of the statistic T. Each replicated value is compared against the estimated critical value $\hat{c}_{.95}$. For the statistic T replicated under H_0, the proportion that is greater than $\hat{c}_{.95}$ is our estimated Type I error. For the T replicated under an H_a, the proportion that is greater than $\hat{c}_{.95}$ is our estimated value of power.

Our results are arranged in eight tables: two for complete and normally distributed data, two for complete and non-normally distributed data, two for incomplete and normally distributed data, and two for incomplete and non-normally distributed data. The first table in each set of two contains the estimated critical values and Type I errors, and the second table contains the estimated values of power corresponding to the three alternative covariance matrices with $a = b = c = .20$, .40, and .60, respectively. For the estimated power in Tables 3, 5, 7, and 9, we did not include the conditions when estimated values of power for all the statistics are 1.0.

Results

The upper panel of Table 2 contains the estimated critical values of the statistics T_{ml}, T_{rml}, $T^{(r)}_{ml}$, and $T^{(r)}_{rml}$ with complete and normally distributed data. The values of $\hat{c}_{.95}$ tend to be greater at smaller N and become close to the value $c_{.95} = 36.415$ corresponding to the nominal chi-square distribution χ^2_{24}. The averages of the estimated critical values across the eight conditions of sample sizes are also reported in Table 2, and all are slightly greater than 36.415.

The lower panel of Table 2 contains the estimated Type I errors when compared each statistic against $c_{.95}$ (left panel) and against the estimated critical values $\hat{c}_{.95}$ (right panel). Type I errors under $c_{.95}$ in the left panel tend to be greater than those under $\hat{c}_{.95}$ in the right panel, especially for the statistic T_{rml} at smaller N. In contrast, there is little variation among the Type I errors under $\hat{c}_{.95}$, although Type I errors at $N = 3{,}000$ tend to be smaller for all the statistics.

---

[3] One could choose a larger number of replications if the range of the involved statistic or its df is large.

TABLE 2
Estimated Critical Values $\hat{c}_{.95}$ (Upper Panel) and Type I Errors (Lower Panel), Complete Data and Normally Distributed Population ($df = 24$, $c_{.95} = 36.415$)

N	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
100	38.242	39.651	39.020	37.693
150	37.501	37.915	37.538	36.546
200	38.089	38.378	37.944	37.460
300	37.011	37.488	37.596	37.148
500	35.890	36.046	36.364	36.056
1,000	36.632	37.033	37.164	36.704
2,000	37.112	36.956	37.402	37.164
3,000	37.066	36.992	37.212	36.924
Average	37.193	37.557	37.530	36.962

	Compared Against $c_{.95}$				*Compared Against* $\hat{c}_{.95}$			
N	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
100	0.077	0.094	0.079	0.069	0.053	0.050	0.048	0.054
150	0.076	0.092	0.079	0.069	0.062	0.072	0.060	0.068
200	0.072	0.081	0.071	0.060	0.046	0.050	0.053	0.045
300	0.062	0.070	0.066	0.060	0.054	0.056	0.049	0.051
500	0.050	0.051	0.051	0.048	0.056	0.059	0.052	0.050
1,000	0.045	0.044	0.050	0.040	0.042	0.038	0.035	0.035
2,000	0.052	0.050	0.054	0.052	0.043	0.047	0.047	0.047
3,000	0.034	0.036	0.043	0.041	0.030	0.032	0.034	0.033
Average	0.059	0.065	0.062	0.055	0.048	0.051	0.047	0.048

All the Type I errors in Table 2 are reasonably close to .05, due to a normally distributed population.

With complete and normally distributed data, Table 3 contains the estimated values of power when each statistic is compared against the critical values $c_{.95}$ and $\hat{c}_{.95}$ in Table 2, respectively. The values of power under the classical method determined by the separation between χ^2_{24} and $\chi^2_{24}(\eta)$ are also included under χ^2 in Table 3. Except at the smallest N, the power values by all the methods are rather close. When N is small, the values of power under χ^2 tend to be greater than those under $\hat{c}_{.95}$ but smaller than those under $c_{.95}$, especially with the statistic T_{rml}. Such a phenomenon is closely related to Type I errors in Table 2. It follows from the averaged values of power in the last row of Table 3 that $T_{ml}^{(r)}$ and $T_{rml}^{(r)}$ are slightly less powerful than T_{ml} and T_{rml} due to a normally distributed population.

Table 4 contains the estimated critical values and Type I errors for complete data but with a non-normally distributed population. The estimated critical values of T_{ml} are about three times that of $c_{.95} = 36.415$, and the average Type I error of T_{ml} under $c_{.95} = 36.415$ is .787! The estimated critical value for T_{rml} at $N = 100$ is also way above $c_{.95} = 36.415$, reflecting its greater variation at small N. Regardless of the size of the estimated critical values, estimated Type I errors of all the statistics under $\hat{c}_{.95}$ are rather close to .05, although not exactly.

The results on power with complete data and the nonnormally distributed population are given in Table 5. Because we cannot properly control Type I errors by comparing a statistic against $c_{.95}$, as reflected in Table 4, the estimated values of power in the left panel of Table 5 are not trustworthy. We thus only discuss the results in the right panel of Table 5, where substantial differences exist between NML and the robust method. Although Huber-type weights by setting the r^2 in Equation 6 at the 95th

TABLE 3
Estimated Values of Power by Classical (Left Panel) and Monte Carlo Methods (Right Panel), Complete Data and Normally Distributed Population

			Compared Against $c_{.95}$				*Compared Against* $\hat{c}_{.95}$			
H_a	N	χ^2	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
$a = .2$	100	0.141	0.161	0.188	0.166	0.146	0.118	0.123	0.109	0.124
	150	0.204	0.238	0.259	0.246	0.210	0.198	0.207	0.202	0.208
	200	0.275	0.279	0.286	0.290	0.254	0.230	0.235	0.238	0.223
	300	0.432	0.436	0.447	0.444	0.421	0.414	0.410	0.391	0.398
	500	0.715	0.712	0.716	0.715	0.702	0.729	0.725	0.716	0.709
	1,000	0.980	0.981	0.981	0.983	0.979	0.979	0.978	0.978	0.976
$a = .4$	100	0.514	0.529	0.561	0.529	0.478	0.451	0.433	0.428	0.417
	150	0.749	0.760	0.785	0.765	0.728	0.729	0.730	0.729	0.722
	200	0.892	0.893	0.890	0.896	0.869	0.857	0.847	0.863	0.834
	300	0.987	0.985	0.985	0.982	0.979	0.982	0.981	0.979	0.978
$a = .6$	100	0.884	0.873	0.882	0.870	0.829	0.831	0.814	0.811	0.796
	150	0.985	0.987	0.987	0.987	0.981	0.987	0.984	0.985	0.981
	200	0.999	0.998	0.998	0.998	0.997	0.996	0.996	0.996	0.996
Average		0.674	0.679	0.690	0.682	0.659	0.654	0.651	0.648	0.643

Note. The values of RMSEA = $(F_{mla}/df)^{1/2}$ corresponding to $a = .2$, .4, and .6 are .040, .076, and .106, respectively.

quantile of χ^2_9 might be far from optimal, the estimated values of power for statistics $T^{(r)}_{ml}$ and $T^{(r)}_{rml}$ are way above those of T_{ml} and T_{rml} for many conditions, including $a = .2$ and $N = 200$, 300, and 500; $a = .4$ and $N = 100$, 150; and $a = .6$ and $N = 100$. There is little difference between NML and the robust method in statistical power when the values of power for $T^{(r)}_{ml}$ and $T^{(r)}_{rml}$ are close to Type I errors or when the values of power for T_{ml} and T_{rml} are close to .90. This is expected because no method is able to tell the difference between two distributions with a tiny separation due to a small sample size or small effect size, and all methods would have enough power if the separation between two distributions is large enough.

Table 6 contains the estimated critical values and Type I errors with incomplete data from the normally distributed population, where all the critical values are above $c_{.95} = 36.415$, especially those corresponding to T_{ml} and $T^{(r)}_{ml}$. Except for $T^{(r)}_{rml}$, Type I errors for the other three statistics under $c_{.95}$ at $N = 100$ and 150 are double the nominal level of .05. In contrast, the estimated Type I errors under $\hat{c}_{.95}$ are much closer to .05, although the smallest one is .028.

Table 7 contains the estimated values of power with incomplete data and the normally distributed population. The values of power among the four statistics under $\hat{c}_{.95}$ are

TABLE 4
Estimated Critical Values $\hat{c}_{.95}$ (Upper Panel) and Type I Errors (Lower Panel), Complete Data and Non-normally Distributed Population ($df = 24$, $c_{.95} = 36.415$)

N	T_{ml}	T_{rml}	$T^{(r)}_{ml}$	$T^{(r)}_{rml}$
100	123.962	52.982	46.861	32.472
150	77.429	39.269	45.057	33.574
200	86.916	39.528	45.206	34.759
300	96.830	36.948	43.022	33.971
500	103.913	36.945	45.076	36.386
1,000	108.315	36.926	45.452	37.522
2,000	105.350	35.683	44.058	36.755
3,000	107.399	36.410	44.211	37.052
Average	101.264	39.336	44.868	35.311

	Compared Against $c_{.95}$				*Compared Against* $\hat{c}_{.95}$			
N	T_{ml}	T_{rml}	$T^{(r)}_{ml}$	$T^{(r)}_{rml}$	T_{ml}	T_{rml}	$T^{(r)}_{ml}$	$T^{(r)}_{rml}$
100	0.667	0.120	0.205	0.012	0.035	0.033	0.036	0.034
150	0.706	0.097	0.198	0.018	0.079	0.061	0.049	0.051
200	0.741	0.068	0.191	0.024	0.047	0.037	0.038	0.043
300	0.762	0.066	0.209	0.030	0.050	0.057	0.066	0.064
500	0.795	0.054	0.179	0.039	0.037	0.045	0.039	0.039
1,000	0.869	0.064	0.179	0.047	0.034	0.058	0.033	0.033
2,000	0.874	0.049	0.179	0.051	0.053	0.055	0.046	0.048
3,000	0.881	0.054	0.187	0.056	0.068	0.054	0.047	0.049
Average	0.787	0.072	0.191	0.035	0.050	0.050	0.044	0.045

TABLE 5
Estimated Values of Power by Classical (Left Panel) and Monte Carlo Methods (Right Panel), Complete Data and Non-Normally Distributed Population

			Compared Against $c_{.95}$				*Compared Against* $\hat{c}_{.95}$			
H_a	N	χ^2	T_{ml}	T_{rml}	$T^{(r)}_{ml}$	$T^{(r)}_{rml}$	T_{ml}	T_{rml}	$T^{(r)}_{ml}$	$T^{(r)}_{rml}$
$a = .2$	100	0.141	0.761	0.201	0.378	0.045	0.049	0.049	0.093	0.094
	150	0.204	0.831	0.195	0.459	0.098	0.118	0.132	0.168	0.162
	200	0.275	0.861	0.196	0.533	0.150	0.082	0.120	0.215	0.205
	300	0.432	0.924	0.251	0.670	0.336	0.091	0.233	0.438	0.440
	500	0.715	0.965	0.369	0.862	0.620	0.097	0.343	0.627	0.620
	1,000	0.980	1.000	0.690	0.993	0.971	0.287	0.666	0.962	0.963
	2,000	1.000	1.000	0.942	1.000	1.000	0.796	0.951	1.000	1.000
	3,000	1.000	1.000	0.982	1.000	1.000	0.973	0.982	1.000	1.000
$a = .4$	100	0.514	0.901	0.472	0.716	0.202	0.107	0.132	0.382	0.346
	150	0.749	0.962	0.518	0.857	0.491	0.293	0.422	0.625	0.606
	200	0.892	0.981	0.622	0.939	0.724	0.266	0.501	0.794	0.792
	300	0.987	0.997	0.800	0.993	0.966	0.385	0.787	0.981	0.979
	500	1.000	1.000	0.943	1.000	1.000	0.679	0.938	1.000	1.000
	1,000	1.000	1.000	0.997	1.000	1.000	0.994	0.996	1.000	1.000
	2,000	1.000	1.000	0.999	1.000	1.000	1.000	0.999	1.000	1.000
	3,000	1.000	1.000	0.999	1.000	1.000	1.000	0.999	1.000	1.000
$a = .6$	100	0.884	0.978	0.780	0.944	0.559	0.189	0.320	0.748	0.727
	150	0.985	0.999	0.858	0.987	0.888	0.629	0.790	0.944	0.938
	200	0.999	1.000	0.923	0.998	0.974	0.687	0.873	0.978	0.980
	300	1.000	1.000	0.986	1.000	0.998	0.872	0.985	0.999	0.999
	500	1.000	1.000	0.995	1.000	1.000	0.992	0.995	1.000	1.000
Average		0.798	0.960	0.701	0.873	0.715	0.504	0.629	0.760	0.755

Note. The values of RMSEA = $(F_{mla}/df)^{1/2}$ corresponding to $a = .2$, .4, and .6 are .040, .076, and .106, respectively.

TABLE 6
Estimated Critical Values $\hat{c}_{.95}$ (Upper Panel) and Type I Errors (Lower Panel), Incomplete Data and Normally Distributed Population ($df = 24$, $c_{.95} = 36.415$)

N	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
100	42.587	40.698	42.983	38.057
150	40.394	37.728	41.001	36.985
200	41.443	38.291	41.345	37.479
300	40.505	37.798	40.779	37.257
500	39.573	37.013	39.947	36.497
1,000	40.539	37.564	40.881	37.346
2,000	40.026	36.945	40.560	37.174
3,000	40.074	36.730	40.328	36.947
Average	40.643	37.846	40.978	37.218

	Compared Against $c_{.95}$				*Compared Against $\hat{c}_{.95}$*			
N	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
100	0.139	0.108	0.145	0.078	0.050	0.049	0.050	0.052
150	0.135	0.101	0.143	0.078	0.074	0.074	0.063	0.069
200	0.119	0.080	0.127	0.065	0.049	0.055	0.053	0.054
300	0.109	0.069	0.102	0.057	0.053	0.051	0.053	0.050
500	0.086	0.050	0.084	0.049	0.050	0.047	0.047	0.048
1,000	0.091	0.045	0.094	0.040	0.032	0.030	0.028	0.027
2,000	0.097	0.045	0.102	0.047	0.039	0.040	0.039	0.041
3,000	0.092	0.047	0.089	0.045	0.036	0.041	0.036	0.036
Average	0.109	0.068	0.111	0.057	0.048	0.048	0.046	0.047

comparable, with T_{ml} and T_{rml} being slightly more powerful than $T_{ml}^{(r)}$ and $T_{rml}^{(r)}$ on average.

Table 8 contains the estimated critical values and Type I errors with incomplete data from the non-normally distributed population, where the estimated Type I errors for all the statistics under $c_{.95}$ are affected by the sample size, and are not trustworthy when N is small. In contrast, Type I errors under $\hat{c}_{.95}$ are rather close to the nominal level, regardless of the values of $\hat{c}_{.95}$ in the upper panel.

Table 9 contains the estimated values of power when the four statistics are compared against $c_{.95}$ and $\hat{c}_{.95}$, respectively. Again, we cannot trust the estimated values of power in the left panel of Table 9 because they are confounded with Type I errors. Similar to the results in Table 5, statistics $T_{ml}^{(r)}$ and $T_{rml}^{(r)}$ are much more powerful than T_{ml} and T_{rml} for many conditions, including $a = .2$ and $N = 200$, 300, 500, 1,000; $a = .4$ and $N = 100$, 150, 200, 300; and $a = .6$ and $N = 100$, 150, 200. Again, there is little difference between the robust method and NML when T_{ml} and T_{rml} have enough power or when $T_{ml}^{(r)}$ and $T_{rml}^{(r)}$ have little power.

In summary, the MC method developed in this article allows us to control Type I errors very well for all the test statistics, regardless of whether the population distribution is normally distributed or non-normally distributed or whether the data are complete or incomplete. With complete or incomplete data from the normal distribution, the robust method performs essentially the same as NML. In contrast, statistics under the robust method are much more powerful than those under NML for the non-normally distributed population, regardless of whether data are complete or incomplete. Compared across the tables, $T_{ml}^{(r)}$ is slightly more powerful than $T_{rml}^{(r)}$ across all the conditions on average. In contrast, T_{ml} is slightly more powerful than T_{rml} for the normal population and somewhat less powerful than T_{rml} for the non-normally distributed population.

Note that evaluating power by comparing each statistic against $c_{.95}$ is parallel to the bootstrap and MC methods

TABLE 7
Estimated Values of Power by Classical (Left Panel) and Monte Carlo Methods (Right Panel), Incomplete Data and Normally Distributed Population

			Compared Against $c_{.95}$				*Compared Against $\hat{c}_{.95}$*			
H_a	N	χ^2	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
$a = .2$	100	0.141	0.256	0.193	0.257	0.149	0.107	0.105	0.102	0.105
	150	0.204	0.337	0.252	0.348	0.220	0.214	0.218	0.198	0.203
	200	0.275	0.378	0.286	0.385	0.251	0.217	0.230	0.225	0.226
	300	0.432	0.535	0.431	0.543	0.406	0.372	0.374	0.370	0.365
	500	0.715	0.768	0.679	0.764	0.657	0.663	0.656	0.655	0.656
	1000	0.980	0.984	0.967	0.986	0.966	0.956	0.959	0.955	0.953
$a = .4$	100	0.514	0.602	0.522	0.605	0.436	0.388	0.367	0.373	0.382
	150	0.749	0.790	0.730	0.791	0.682	0.690	0.693	0.675	0.665
	200	0.892	0.922	0.865	0.918	0.835	0.815	0.815	0.822	0.814
	300	0.987	0.987	0.975	0.988	0.971	0.967	0.967	0.964	0.967
$a = .6$	100	0.884	0.896	0.860	0.900	0.799	0.780	0.760	0.763	0.752
	150	0.985	0.989	0.984	0.989	0.975	0.975	0.979	0.969	0.972
	200	0.999	1.000	0.997	0.999	0.996	0.994	0.993	0.994	0.992
Average		0.674	0.726	0.672	0.729	0.642	0.626	0.624	0.620	0.619

Note. The values of RMSEA = $(F_{mla}/df)^{1/2}$ corresponding to $a = .2$ .4 and .6 are .040, .076, and .106, respectively.

TABLE 8
Estimated Critical Values $\hat{c}_{.95}$ (Upper Panel) and Type I Errors (Lower Panel), Incomplete Data and Non-Normally Distributed Population ($df = 24$, $c_{.95} = 36.415$)

N	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
100	120.963	49.944	50.800	32.333
150	77.787	38.532	48.456	33.768
200	76.664	36.440	47.809	34.273
300	72.499	35.127	46.350	33.629
500	72.899	35.271	47.874	35.771
1,000	72.633	35.540	48.730	37.771
2,000	70.664	36.238	46.818	36.463
3,000	70.972	35.800	46.754	36.184
Average	79.385	37.862	47.949	35.024

	Compared Against $c_{.95}$				*Compared Against $\hat{c}_{.95}$*			
N	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
100	0.756	0.141	0.320	0.015	0.038	0.043	0.050	0.050
150	0.790	0.075	0.293	0.018	0.047	0.049	0.058	0.045
200	0.781	0.046	0.278	0.030	0.040	0.046	0.048	0.052
300	0.763	0.033	0.289	0.032	0.054	0.044	0.066	0.072
500	0.753	0.032	0.234	0.035	0.046	0.040	0.042	0.044
1,000	0.789	0.059	0.257	0.051	0.047	0.063	0.037	0.033
2,000	0.764	0.048	0.237	0.052	0.052	0.051	0.046	0.050
3,000	0.784	0.054	0.263	0.045	0.058	0.061	0.043	0.047
Average	0.773	0.061	0.271	0.035	0.048	0.050	0.049	0.049

proposed in Yung and Bentler (1996), Muthén and Muthén (2002), Schoemann et al. (2014), and Zhang (2014). According to Tables 2 and 3, the method controls Type I error reasonably well when the underlying population is multivariate normal and data are complete. However, the method does not work well when data are incomplete, especially when the underlying population is not multivariate normal.

DISCUSSION AND RECOMMENDATION

In this article we proposed using the MC method for power analysis in SEM, and using robust M-estimation to deal with non-normally distributed data in practice. The most notable feature of the MC method is that it allows us to reliably control Type I errors for any statistic for which the distribution might not be known. The advantage of a robust method is that it programmatically increases the effect size when the population distribution contains heavy tails. Compared to conventional methods in which power value is determined by the separation between χ^2_{df} and $\chi^2_{df}(\eta)$, the MC method needs a population from which we simulate our data. In practice, a researcher might not know what the true population distribution is, but can only specify marginal skewnesses and kurtoses. Then one can simulate data

TABLE 9
Estimated Values of Power by Classical (Left Panel) and Monte Carlo Methods (Right Panel), Incomplete Data and Non-Normally Distributed Population

			Compared Against $c_{.95}$				*Compared Against $\hat{c}_{.95}$*			
H_a	N	χ^2	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$	T_{ml}	T_{rml}	$T_{ml}^{(r)}$	$T_{rml}^{(r)}$
$a = .2$	100	0.141	0.814	0.219	0.482	0.053	0.055	0.066	0.106	0.102
	150	0.204	0.856	0.144	0.527	0.073	0.095	0.106	0.157	0.139
	200	0.275	0.877	0.122	0.599	0.122	0.109	0.121	0.207	0.184
	300	0.432	0.921	0.149	0.701	0.286	0.198	0.182	0.376	0.394
	500	0.715	0.956	0.281	0.860	0.536	0.286	0.327	0.546	0.563
	1,000	0.980	0.999	0.693	0.993	0.935	0.665	0.720	0.924	0.925
	2,000	1.000	1.000	0.973	1.000	1.000	0.977	0.974	1.000	1.000
	3,000	1.000	1.000	0.999	1.000	1.000	0.999	0.999	1.000	1.000
$a = .4$	100	0.514	0.923	0.461	0.756	0.180	0.115	0.160	0.330	0.327
	150	0.749	0.968	0.403	0.873	0.407	0.292	0.336	0.569	0.529
	200	0.892	0.981	0.443	0.939	0.657	0.405	0.443	0.746	0.721
	300	0.987	0.996	0.657	0.993	0.925	0.727	0.694	0.962	0.969
	500	1.000	0.999	0.931	1.000	0.999	0.936	0.942	0.999	0.999
$a = .6$	100	0.884	0.979	0.756	0.949	0.511	0.218	0.375	0.679	0.664
	150	0.985	0.997	0.697	0.989	0.829	0.643	0.643	0.902	0.892
	200	0.999	1.000	0.766	0.994	0.954	0.819	0.766	0.972	0.968
	300	1.000	1.000	0.956	1.000	0.997	0.982	0.965	0.997	0.997
	500	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000
Average		0.764	0.959	0.592	0.870	0.637	0.529	0.546	0.693	0.687

Note. The values of RMSEA = $(F_{mla}/df)^{1/2}$ corresponding to $a = .2$, .4, and .6 are .040, .076, and .106, respectively.

using the method as described in Vale and Maurelli (1983), Nagahara (2004), or Auerswald and Moshagen (2015). In the Appendix, we describe a method that allows the control of the relative multivariate kurtosis in addition to marginal skewnesses and kurtoses.

When nonnormally distributed data are expected and a researcher can approximately specify the marginal skewnesses and kurtoses or multivariate kurtosis, the MC method including the features of non-normality is recommended. In practice, information on skewness and kurtosis can be obtained from meta-analysis on distributions of real data (e.g., Blanca et al., 2013; Cain et al., in press) or by searching the Web or contacting authors for data that are collected using the same measurement scales or on the same research topic. When a sample with a moderate size has already been collected, then one could execute the MC method by drawing samples from the empirical distribution of the sample (nonparametric bootstrap). If a researcher has no idea of the population distribution, then use the chi-square-based method when p is relatively small, and the MC method with normally distributed population when p is relatively large. However, power and sample size determined by such methods might not be reliable when the data to be collected turn out to be non-normally distributed with heavy tails. As for test statistics, we recommend using $T_{ml}^{(r)}$ (T_{ml} following a robust method) with non-normally distributed data, and use T_{ml} when one has confidence that the population is normally distributed. The MC methodology as developed in this article together with the robust methods will be available via an online statistical package, which allows researchers to conduct power analysis by specifying marginal skewness and kurtoses as well as the relative multivariate kurtosis.

For the simulation study in the previous section, we used the robust method with Huber-type weights corresponding to down-weighting 5% of a normally distributed population. The robust method might be refined by adjusting the tuning parameter r in Equation 6. In particular, one can choose r corresponding to the empirically most efficient parameter estimates using MC or the bootstrap methodology (Yuan et al., 2004). The most efficient parameter estimates correspond to the smallest possible β_r in Equation 9 and consequently most powerful test statistic.

Because the values of effect sizes defined in Equations 7 to 10 are simply measures of model misspecification, there might be a conflict of interest between researchers who like to have a test statistic with better power and those who are interested in a model that might not fit their data as well but allows them to elaborate on the substantive variables. A robust method might yield more efficient parameter estimates but at the same time the test statistics $T_{ml}^{(r)}$ or $T_{rml}^{(r)}$ will make an already poor model even poorer, due to greater effect sizes. Although researchers might not choose a test statistic with a better power, we still recommend using a robust method in real data analysis. This is because a seemingly poor model might be due to outliers or a small proportion of observations that do not fit the model well, and robust methods will down-weight the influence of these observations and return the model with the credit it deserves.

FUNDING

This research was supported by a grant from the Institute of Education Sciences (R305D140037), and by the National Science Foundation under Grant No. SES-1461355.

REFERENCES

Auerswald, M., & Moshagen, M. (2015). Generating correlated, non-normally distributed data using a non-linear structural model. *Psychometrika*, *80*, 920–937.

Bentler, P. M. (2006). *EQS 6 structural Equations program manual*. Encino, CA: Multivariate Software.

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

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*, 78–84.

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

Cain, M., Zhang, Z., & Yuan, K. H. (in press). Univariate and multivariate skewness and kurtosis for measuring non-normality: Prevalence, influence and estimation. *Behavior Research Methods*.

Davey, A., & Savla, J. (2009). Estimating statistical power with incomplete data. *Organizational Research Methods*, *12*, 320–346.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, *39*, 1–38.

Dolan, C., & van der Sluis, S., & Grasman, R. (2005). A note on normal theory power calculation in SEM with data missing completely at random. *Structural Equation Modeling*, *12*, 245–262.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural Equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling*, *11*, 1–19.

Fleishman, A. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532.

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

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 221–233. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512988.

Huber, P. J. (1981). *Robust statistics*. New York, NY: Wiley.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the distribution. *Journal of the American Statistical Association*, *84*, 881–896.

Little, R. J. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, *37*, 23–38.

Little, R. J. A., Smith, P. J. (1987). Editing and imputing for quantitative survey data. *Journal of the American Statistical Association*, *82*, 58–68.

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.

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*, 305–325.

Marcoulides, G. A., & Chin, W. (2013). You write, but others read: Common methodological misunderstandings in PLS and related methods. In H. Abdi, W. Chin, V. E. Vinzi, G. Russolillo, & L. Trinchera (Eds.), *New perspectives in partial least squares and related methods* (pp. 31–64). Berlin, Germany: Springer-Verlag.

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

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

Moshagen, M. (2012). The model size effect in SEM: inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling*, 19, 86–98.

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

Nagahara, Y. (2004). A method of simulating multivariate non-normal distributions by the Pearson distribution system and estimation. *Computational Statistics & Data Analysis*, *47*, 1–29.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural Equation modeling. *Multivariate Behavioral Research*, *39*, 439–478.

Rubin, D. B. (1976). Inference and missing data (with discussions). *Biometrika*, *63*, 581–592.

Satorra, A. (2003). Power of goodness-of-fit tests in structural Equation models: The case of non-normal data. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments of psychometrics* (pp. 57–68). Tokyo Japan: Springer Verlag.

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

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

Savalei, V., & Bentler, P. M. (2009). A two-stage ML approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling*, *16*, 477–497.

Savalei, V., & Falk, C. (2014). Robust two-stage approach outperforms robust FIML with incomplete non-normal data. *Structural Equation Modeling*, *21*, 280–302.

Schoemann, A. M., Miller, P. J., Pornprasertmanit, S., & Wu, W. (2014). Using Monte Carlo simulations to determine power and sample size in planned missing data designs. *International Journal of Behavioral Development*, *38*, 471–479.

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

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

Tadikamalla, P. R. (1980). On simulating non-normal distributions. *Psychometrika*, *45*, 273–279.

Tong, X., & Zhang, Z. (2012). Diagnostics of robust growth curve modeling using student's t distribution. *Multivariate Behavioral Research*, *47*, 493–518.

Tong, X., Zhang, Z., & Yuan, K.-H. (2014). Evaluation of test statistics for robust structural Equation modeling with non-normal missing data. *Structural Equation Modeling*, *21*, 553–565.

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

Vale, D., & Maurelli, V. A. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, *48*, 465–471.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Amsterdam The Netherlands: Academic Press.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural Equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*, 913–934.

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). Improving parameter tests in covariance structure analysis. *Computational Statistics and Data Analysis*, *26*, 177–198.

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

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

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with non-normal missing data. *Sociological Methodology*, *30*, 167–202.

Yuan, K.-H., & Bentler, P. M. (2007). Structural Equation modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26: Psychometrics* (pp. 297–358). Amsterdam The Netherlands: North-Holland.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural Equation modeling with heavy tailed distributions. *Psychometrika*, *69*, 421–436.

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., Chan, W., & Tian, Y. (2016). Expectation-robust algorithm and estimating Equations for means and dispersion matrix with missing data. *Annals of the Institute of Statistical Mathematics*, *68*, 329–351.

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., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology*, *59*, 397–417.

Yuan, K.-H., & Lu, L. (2008). SEM with missing data and unknown population using two-stage ML: Theory and its application. *Multivariate Behavioral Research*, *62*, 621–652.

Yuan, K.-H., & Zhang, Z. (2012). Robust structural Equation modeling with missing data and auxiliary variables. *Psychometrika*, *77*, 803–826.

Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations and influential cases in factor analysis: Minimizing their effect using robust procedures. *Sociological Methodology*, *38*, 329–368.

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). Hillsdale, NJ: Erlbaum.

Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods*, *46*, 1184–1198.

APPENDIX

Simulation Data with Desired Multivariate Kurtosis

This appendix describes a method for simulating non-normally distributed data from a distribution with given marginal skewnesses and kurtoses as well as multivariate kurtosis. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the target population mean vector and covariance matrix, respectively. There is a symmetric matrix $\mathbf{A} = (a_{ij}) = \boldsymbol{\Sigma}^{1/2}$ such that $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$. Our model for simulating non-normally distributed data is

$$\mathbf{x} = \boldsymbol{\mu} + u\mathbf{A}\mathbf{z}, \tag{A.1}$$

where $\mathbf{z} = (z_1, z_2, \cdots, z_p)'$ with $E(z_j) = 0$, $E(z_j^2) = 1.0$, $E(z_j^3) = skew_j$, and $E(z_j^4) = kurt_j + 3$, and u is a standardized random variable with $E(u^2) = 1$, $E(u^3) = \gamma_3$, $E(u^4) = \gamma_4$, and is independent with $\mathbf{z}$. The way of simulating data according to Equation A.1 is a special case of a method proposed in Yuan and Bentler (1997a). It can be shown that the marginal skewness and kurtosis of the $\mathbf{x}$ in Equation A.1 are given by

$$\text{skew}(x_i) = \gamma_3 \sum_{j=1}^{p} a_{ij}^3 \ \text{skew}_j/\sigma_{ii}^{3/2}, \quad i = 1, 2, \ldots, p; \tag{A.2}$$

$$\text{kurt}(x_i) = \gamma_4 \sum_{j=1}^{p} a_{ij}^4 \ \text{kurt}_j/\sigma_{ii}^2 + 3(\gamma_4 - 1), \quad i = 1, 2, \ldots, p; \tag{A.3}$$

and the relative multivariate kurtosis is given by

$$\beta = \gamma_4 \left\{ 1 + \sum_{j=1}^{p} \text{kurt}_j/[p(p+2)] \right\}. \tag{A.4}$$

Note that there are $2p + 1$ Equations in A.2, A.3 and A.4. The coefficients a_{ij} are determined via the covariance matrix $\boldsymbol{\Sigma}$. The left sides of Equation A.2, A.3, and A.4 are the marginal skewnesses, kurtoses, and relative multivariate kurtosis to be specified, and the right sides of A.2 to A.4 contain $2p + 2$ unknown quantities: $(skew_j, kurt_j)$, $j = 1, 2, \ldots, p$, γ_3, and γ_4. Because the number of unknowns is more than the number of Equations, there can be more than one set of $skew_j$ and $kurt_j$, $j = 1, 2, \ldots, p$; and γ_3 and γ_4 that satisfies A.2 to A.4, which can be solved numerically.

Once $skew_j$ and $kurt_j$, $j = 1, 2, \ldots, p$ are obtained, standardized z_j can be obtained by the power transformation method of Fleishman (1978; see also Tadikamalla, 1980) and so is the random variable u.

Note that the values of skewness and kurtosis of $\mathbf{x}$ cannot be arbitrarily chosen because moments of random variables need to satisfy certain conditions (see, e.g., Yuan & Bentler, 1997a). If Equations A.2, A.3, or A.4 do not have a set of solutions for the required values of $skew(x_i)$ and $kurt(x_i)$, then it is most likely that the specified values are inadmissible in the sense that no pupulation distribution satisfies such conditions. Similarly, even if $skew_j$ and $kurt_j$ are obtained, there is still a possibility that the method in Fleishman (1978) or Tadikamalla (1980) cannot generate random z_i that satisfy such desired values. In such a case, there is a need to respecify these values.