*Multivariate Behavioral Research, 37* (1), 1-36


# The Noncentral Chi-square Distribution in Misspecified Structural Equation Models: Finite Sample Results from a Monte Carlo Simulation

Patrick J. Curran
University of North Carolina, Chapel Hill

Kenneth A. Bollen
Department of Sociology
University of North Carolina, Chapel Hill

Pamela Paxton
Department of Sociology
Ohio State University

James Kirby
Agency for Health Care Policy and Research

Feinian Chen
Department of Sociology
University of North Carolina, Chapel Hill

The noncentral chi-square distribution plays a key role in structural equation modeling (SEM). The likelihood ratio test statistic that accompanies virtually all SEMs asymptotically follows a noncentral chi-square under certain assumptions relating to misspecification and multivariate distribution. Many scholars use the noncentral chi-square distribution in the construction of fit indices, such as Steiger and Lind's (1980) Root Mean Square Error of Approximation (RMSEA) or the family of baseline fit indices (e.g., RNI, CFI), and for the computation of statistical power for model hypothesis testing. Despite this wide use, surprisingly little is known about the extent to which the test statistic follows a noncentral chi-square in applied research. Our study examines several hypotheses about the suitability of the noncentral chi-square distribution for the usual SEM test statistic under conditions commonly encountered in practice. We designed Monte Carlo computer simulation experiments to empirically test these research hypotheses. Our experimental

---

This work was funded in part by grant DA13148 awarded by the National Institute on Drug Abuse to the first two authors. The authors would like to thank Steve Gregorich for several helpful discussions on this topic. Correspondence should be addressed to Patrick Curran, Department of Psychology, University of North Carolina-Chapel Hill, Chapel Hill, NC, 27599-3270. Electronic mail may be sent to curran@unc.edu.

conditions included seven sample sizes ranging from 50 to 1000, and three distinct model types, each with five specifications ranging from a correct model to the severely misspecified uncorrelated baseline model. In general, we found that for models with small to moderate misspecification, the noncentral chi-square distribution is well approximated when the sample size is large (e.g., greater than 200), but there was evidence of bias in both mean and variance in smaller samples. A key finding was that the test statistics for the uncorrelated variable baseline model did not follow the noncentral chi-square distribution for any model type across any sample size. We discuss the implications of our findings for the SEM fit indices and power estimation procedures that are based on the noncentral chi-square distribution as well as potential directions for future research.

## *Introduction*

Structural equation modeling (SEM) represents a broad class of models that allows simultaneous estimation of the relations between observed and latent variables and among the latent variables themselves (Bollen, 1989). The SEM framework subsumes a remarkable variety of analytic methods including the simple *t*-test, ANOVA, regression, confirmatory factor analysis and beyond (Bentler, 1980, 1983; Jöreskog, 1971a, 1971b; Jöreskog & Sörbom, 1978). Most of the statistical estimators for SEMs share the goal of minimizing the difference between the covariance matrix observed in the sample and the covariance matrix implied by the model parameters, where the minimization is with respect to a "fitting function," $F$. If we denote $\hat{F}$ as the value of the sample fitting function at its minimum, then we have a scalar that ranges from 0 to infinity and equals 0 only when the estimated implied covariance matrix exactly reproduces the sample covariance matrix. Larger values of $\hat{F}$ reflect greater discrepancies between the observed and implied matrices.

The maximum likelihood fitting function leads to a test statistic $T$ formed by multiplying $\hat{F}$ by $N - 1$, where $N$ represents sample size.[1] This test statistic $T$ asymptotically follows a central chi-square distribution under a set of standard assumptions. Key among these is that the specified model is correct. That is, the covariance matrix implied by the model exactly reproduces the observed variables' population covariance matrix. However, researchers have long recognized that no model is without error and all models are misspecified to some unknown degree (e.g., Cudeck & Browne, 1983; Meehl, 1967). In their seminal early work on this topic, both Steiger and Lind (1980) and Browne (1984) demonstrated that in the typical case of a misspecified model, the test statistic $T$ does not follow a central chi-square distribution. Instead, under certain known conditions $T$ asymptotically

[1] The test statistic $T$ is commonly referred to as the "model $\chi^2$ test" both in the literature and in nearly all SEM computer packages. However, we will refer to this as $T$ throughout because this test statistic may or may not actually follow a chi-square distribution.

follows a noncentral chi-square distribution defined by degrees of freedom *df* and noncentrality parameter $\lambda$. The noncentrality parameter $\lambda$ carries important information about the degree of model misspecification, and thus the noncentral chi-square distribution has come to play an important role in structural equation modeling.

Despite the prominence of the noncentral chi-square distribution in structural equation modeling, little empirical work has examined the extent to which the test statistic $T$ follows the expected distribution in applied research. The purpose of this article is to empirically evaluate the appropriateness of using a noncentral chi-square distribution for $T$ under a range of model misspecifications and sample sizes commonly encountered in practice. We test three key research hypotheses using data generated from Monte Carlo simulations and compare the obtained $T$ statistics both to the population chi-square distributions and to a large set of random draws from the known population distributions. Prior to presenting the specifics of our study, we will first review the important role of the noncentral chi-square distribution in structural equation modeling.

### *The Noncentral Chi-Square Distribution in SEM*

Evidence of the ubiquitous role of the noncentral chi-square distribution in SEM is reflected in the development of numerous measures of overall model fit. For example, Steiger and Lind (1980) originally proposed the RMSEA (Root Mean Square Error of Approximation) to calibrate the omnibus fit of a SEM. Not only does the computation of the point estimate of the RMSEA depend on the sample estimate of $T$, but a critical feature that they introduced was the ability to form confidence intervals for the RMSEA directly based on the noncentral chi-square distribution. Extending this work, Browne and Cudeck (1993) proposed using the RMSEA and the noncentral chi-square distribution to form hypothesis tests of approximate fit rather than the traditional tests of exact fit. Steiger, Shapiro, and Browne (1985) use the noncentral chi-square distribution in their analysis of test statistics for stand alone factor analysis models and comparisons of nested model fit. Steiger (1989) and Maiti and Mukherjee (1990) apply the noncentral chi-square distribution to develop the sampling distribution of the GFI fit statistic. Bentler (1990), McDonald and Marsh (1990), and others form fit indices that compare a "baseline" model to a specific hypothesized model, and underlying their proposal is treating the test statistics from both models as if they follow noncentral chi-square distributions. It is clear that assuming that $T$ follows a noncentral chi-square distribution is critical to the computation and interpretation of all of these measures of fit.

Another important application of the noncentral chi-square distribution is in the study of statistical power in SEM. For instance, Satorra and Saris (1985) and Matsueda and Bielby (1986) used the noncentral chi-square distribution to determine the power of the usual chi-square test statistic for a hypothesized model when a specified alternative model actually holds in the population. These methods have been extended in a variety of directions over the past 15 years. For example, MacCallum, Browne, and Sugawara (1996) rely on the noncentral chi-square distribution when computing the power of ‘close’ and ‘exact’ fit based upon the RMSEA, and Muthén and Curran (1997) extended the methods of Satorra and Saris (1985) to compute statistical power for a broad class of longitudinal models. Taken together, all of these techniques are based on the premise that the test statistic $T$ for a misspecified model follows an underlying noncentral chi-square distribution.

### *The Validity of the Distributional Assumptions for T*

Whether it is the development of new fit indices or the study of statistical power, the noncentral chi-square distribution has moved from a little used statistical distribution in SEM to a key feature of contemporary applications. Given this prominence, it is surprising that there is so little work on whether, and under what conditions, the test statistic $T$ does and does not follow a noncentral chi-square distribution. Some suggest that the test statistic follows a noncentral chi-square distribution whenever a model is incorrect while others claim that the asymptotic noncentral chi-square distribution holds only if certain conditions are met. For example, Steiger et al. (1985) note “...the noncentral Chi-square approximation will be reasonably effective so long as the noncentrality parameter is not ‘too large’” (p. 259, quotes in original). And when discussing the role of the noncentral chi-square distribution of $T$ for his proposed comparative fit index, Bentler (1990) notes “It is possible that the null model of independence may be so different from the true model that another distribution could be more appropriate at times” (p. 245).

Specifically, a noncentral chi-square distribution for $T$ rests on a series of assumptions. Chief among these is that “systematic errors due to lack of fit of the model to the population covariance matrix are not large relative to random sampling errors in **S**” (where **S** represents the sample covariance matrix) (Browne, 1984, p. 66). See also Satorra (1989), Steiger et al. (1985), and Browne and Cudeck (1993) for additional details on this issue. However, it is difficult to know when the systematic errors or the misspecifications are mild enough to justify this assumption. Furthermore, these are asymptotic or large sample results so it is unclear as to how large $N$ must practically be for this approximation to hold.

A number of simulation studies examined the empirical sampling distributions of $T$ for correct model specification. A typical finding is that the value of $T$ tends to be higher than it should be for a chi-square variable at smaller sample sizes, but that this bias disappears as the sample size grows (e.g., Anderson & Gerbing, 1984; Boomsma, 1982; Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992). A smaller number of studies have examined the $T$ statistic under various misspecified models, and results have indicated similar patterns of findings to those under proper model specification (e.g., Curran et al., 1996; Fan, Thompson, & Wang, 1999). Finally, Rigdon (1998) presented the only published study of which we are aware that provides an example of the empirical distribution of $T$ for the uncorrelated variable model that is part of baseline fit indices. Although his results indicated that the distribution of $T$ for the uncorrelated variable model may not follow the noncentral chi-square distribution, the external validity of this finding is limited given the consideration of a single model and a single sample size.

The small amount of existing research of the empirical distribution of $T$ under proper and improper specification tends to be hampered by two key limitations. The first is that, with few exceptions, researchers only compare the means of the empirical distributions of $T$ to that expected for the corresponding population distributions. Rarely are measures of dispersion compared, and this could be critical when using the noncentral chi-square to compute confidence intervals. Second, studies of misspecified models have not considered *severe* misspecifications, the condition under which $T$ is least likely to follow the noncentral distribution. More specifically, almost nothing is known about the distribution of $T$ for the uncorrelated variable model that is commonly used in the computation of many baseline fit indices (e.g., TLI, IFI, or CFI). Thus the validity of treating the test statistic $T$ as if it follows a central or noncentral chi-square distribution in situations commonly encountered in applied research is open to question. The purpose of our article is to empirically evaluate the validity of employing this distribution in practice.

### *Proposed Research Hypotheses*

We use extensive Monte Carlo computer simulations to empirically evaluate hypotheses based on statistical theory and prior research. To isolate the impact of misspecification and sample size from problems caused by the distribution of variables, all observed variables are generated from multinormal distributions. Our three key research hypotheses are as follows.

1. Drawing on Browne (1984) and others, we propose that under proper model specification, the test statistic *T* will follow a central chi-square distribution with mean *df* and variance 2*df*, but only at moderate to large sample sizes; *T* will follow some other (unknown) distribution at smaller sample sizes.

2. Drawing on Steiger et al. (1985) and others, we propose that under small to moderate model misspecification, the test statistic *T* will follow a noncentral chi-square distribution with noncentrality parameter $\lambda$, mean $df + \lambda$ and variance $2df + 4\lambda$. However, this will only hold at moderate to large sample sizes; *T* will follow some other (unknown) distribution at smaller sample sizes.

3. Also drawing on Steiger et al. (1985) and others, we propose that under severe model misspecification, especially the uncorrelated variable model, the test statistic *T* will not follow either the central nor the noncentral chi-square distribution, and this will occur across all sample sizes; *T* will follow some other (unknown) distribution regardless of sample size.

To maximize the external validity of our study, we utilized 15 separate specifications of three general model types that represent a broad sampling of common models. Further, we evaluate these models using sample sizes ranging from very small to very large to further understand these issues across a spectrum of applied research settings. Finally, we test both the mean and the variance of the empirical distribution of *T* relative to the population distributions to evaluate the implications of potential bias in the calculation of both point estimates and confidence intervals. Taken together, we believe our methodological design and analytic strategy provide a rigorous empirical evaluation of our proposed research hypotheses.

### *Technical Background*

Prior to presenting the design of the simulation study, we will briefly review some basic technical issues to provide background context and to concretely define terms and clarify notation.

*The Central and Noncentral Chi-Square Distribution*. The central chi-square ($\chi^2$) distribution is a common distribution in inferential statistics. The central $\chi^2$ distribution is defined by a single parameter *df*, or degrees of freedom, and is a special case of the broader family of gamma distributions (Freund, 1992). We can express a random variable that is distributed as a central chi-square as the sum of *df* squared random normal deviates *z* such that

(1) $$\chi^2_{df} = \sum_{j=1}^{df} z_j^2$$

where $df$ = degrees of freedom. The mean of $\chi^2_{df}$ is $df$ and the variance is $2df$. A less widely utilized variant of the central $\chi^2$ distribution is the *noncentral* chi-square distribution (commonly denoted $\chi'^2$). Whereas the central $\chi^2$ is the sum of one or more squared normal deviates, the noncentral $\chi'^2$ is the sum of one or more squared normal deviates plus a constant $c$ such that

(2) $$\chi'^2_{df} = \sum_{j=1}^{df} (z_j + c_j)^2$$

The noncentral chi-square is defined by two parameters, $df$ and the noncentrality parameter $\lambda$ (where $\lambda = \Sigma c_j^2$). The mean of $\chi'^2_{df}$ is $df + \lambda$ and the variance is $2df + 4\lambda$.

*Structural Equation Modeling.* Within the SEM framework, $\mathbf{\Sigma}$, the population covariance matrix of the observed variables, equals an implied covariance matrix, $\mathbf{\Sigma}(\theta)$ where the values of $\theta$ represent the regression coefficients, factor loadings, and covariance matrices of the specified model [e.g., for further details see Chapter 2 of Bollen, 1989, for notation, and Chapter 8 for $\mathbf{\Sigma}(\theta)$]. This covariance structure is fitted to the observed covariance matrix $\mathbf{S}$ by means of minimizing a given fit function $F[\mathbf{S}, \mathbf{\Sigma}(\theta)]$ with respect to $\theta$. This minimization results in $\hat{\theta}$ which is a vector of model parameter estimates, and $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}\left(\hat{\theta}\right)$ which is the covariance structure implied by the parameter estimates. The goal of the estimation procedure is to select values for $\hat{\theta}$ that minimize the difference between $\mathbf{S}$ and $\hat{\mathbf{\Sigma}}$. The discrepancy function $F[\mathbf{S},\mathbf{\Sigma}(\theta)]$ is thus a scalar value that ranges from 0 to $\infty$ and is equal to 0 when $\mathbf{S} = \mathbf{\Sigma}\left(\hat{\theta}\right)$. There are several discrepancy functions from which to choose (see, e.g., Browne, 1984), but the most widely used in applied research is maximum likelihood estimation.

*Maximum Likelihood Estimation.* The maximum likelihood fitting function is:

(3) $$\hat{F}_{ML} = \log|\mathbf{\Sigma}(\hat{\theta})| + tr[S\mathbf{\Sigma}^{-1}(\hat{\theta})] - \log|S| - p$$

where $p$ represents the total number of observed measured variables. Assuming no excessive kurtosis, adequate sample size, and proper model specification, ML parameter estimates are asymptotically unbiased, consistent, and efficient (see, e.g., Bollen, 1989). Further, a test statistic is

(4) $$T = \hat{F}_{ML}(N-1)$$

which, given the above assumptions, is asymptotically distributed as a central chi-square with $df = 1/2(p)(p + 1) - t$ where $t$ is the number of parameters to be estimated. This test statistic and corresponding $df$ permit tests of the null hypothesis $H_0$: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\theta)$. Under the assumptions of no excessive kurtosis, adequate sample size, and improper model specification (but not severely so), the test statistic $T$ instead follows a noncentral chi-square distribution defined by $df$ and noncentrality parameter $\lambda$. The noncentrality parameter $\lambda$ provides a basis for evaluating the degree of model misfit.

## *Method*

Given space limitations, we provide a general summary of the simulation design and methods here. A comprehensive presentation of this information is available in Paxton, Curran, Bollen, Kirby and Chen (2001). As will be described below, we generated two sets of data to test our hypotheses. The first data set was comprised of the $T$ statistics estimated from the SEM simulations across a variety of experimental conditions. The second data set was comprised of random draws from a known central and noncentral chi-square distribution, the generation of which was entirely independent of the SEM simulations. A key component of our analytic strategy is to compare the distribution of the simulated $T$ statistics with (a) the population moments of the known underlying distribution, and (b) the sample moments of the variates randomly drawn from the same known underlying distribution. This second comparison was necessary given that the parametric tests comparing the $T$ test statistics to the underlying population distribution parameters assumes normality, and we know a priori that the $T$ statistics will not follow a normal distribution. We thus combine parametric and nonparametric comparisons to allow for a comprehensive evaluation of the proposed research hypotheses.

We will now describe the selection of the target models and the method used to generate the two simulated data sets.

### *Model Types and Experimental Conditions*

Drawing both on a review of the social science literature over the previous five years and on our combined modeling experience, we selected three general model types for study: Model 1 (see Figure 1) contains nine measured variables and three latent factors with three to four indicators per factor, Model 2 (see Figure 2) has 15 measured variables and three latent factors with five to six indicators per factor, and Model 3 (see Figure 3) consists of 13 measured variables with the same form as Model 1 but with the addition of four measured and correlated exogenous variables. We designed these models to represent features that are commonly encountered in social science research. Furthermore, for each model we use one correct and four incorrect specifications, resulting in a total of 15 target models.

*Model 1.* Specification 1 is a properly specified model such that the estimated model matches the population model; Specification 2 omits the complex loading linking item 7 with factor 2; Specification 3 additionally omits the complex loading linking item 6 with factor 3; Specification 4 additionally removes the complex loading linking item 4 with factor 1; and finally, Specification 5 is the standard uncorrelated variables model where variances are estimated but all covariances are fixed at zero.

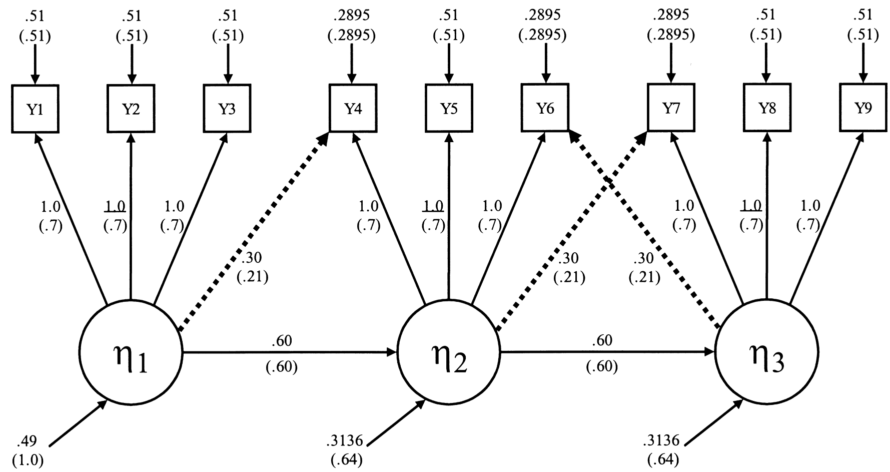


**Figure 1**

Target Population Model 1

*Note*: numbers shown are unstandardized parameter values with standardized values in parenthesis; solid and dashed lines represent the population model structure, and dashed lines represent omitted parameters under model misspecification.

*Model 2*. Specification 1 is properly specified; Specification 2 omits the complex loading linking item 11 with factor 2; Specification 3 additionally omits the complex loading linking item 10 with factor 3; Specification 4 additionally removes the complex loading linking item 6 with factor 1; and Specification 5 is the standard uncorrelated variables model.

*Model 3*. Specification 1 is properly specified; Specification 2 jointly omits the set of three complex factor loadings (item 7 with factor 2, item 6 with factor 3, and item 4 with factor 1); Specification 3 jointly omits the set of four regression parameters (factor 2 on predictor 1, factor 3 on predictor 1, factor 2 on predictor 3, and factor 3 on predictor 3); Specification 4 jointly combines the omissions of Specifications 2 and 3 (omission of the set of three factor loadings and the set of four regression parameters); and Specification 5 is the standard uncorrelated variables model.

*Model Parameterization*. For all three model types, parameter values were carefully selected to result in a range of effect sizes (e.g., communalities and $R^2$ values ranging from 49% to 72%), and for the misspecified conditions to lead to both a wide range of power to detect the misspecifications (e.g., power ranging from .07 to 1.0 across all sample sizes) and a range of bias in parameter estimates (e.g., absolute bias ranging from 0 to 37%). See Paxton et al. (2001) for a comprehensive description of our model parameterization. We believe this parameterization reflects values commonly encountered in applied research and that the omission of

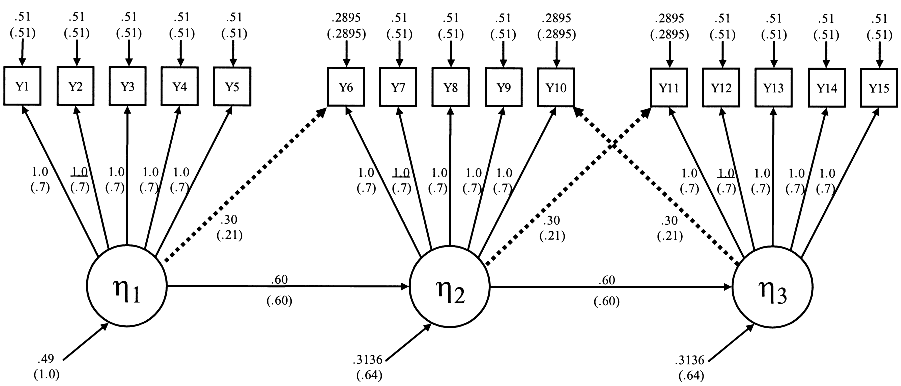


**Figure 2**

Target Population Model 2

*Note*: numbers shown are unstandardized parameter values with standardized values in parenthesis; solid and dashed lines represent the population model structure, and dashed lines represent omitted parameters under model misspecification.

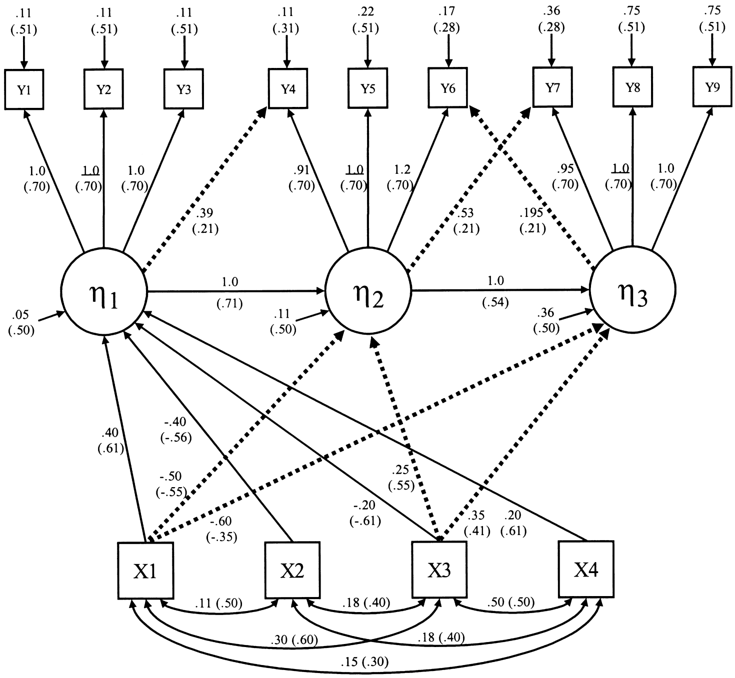


**Figure 3**
Target Population Model 3
*Note*: numbers shown are unstandardized parameter values with standardized values in parenthesis; solid and dashed lines represent the population model structure, and dashed lines represent omitted parameters under model misspecification.

one or more parameters would result in meaningful impacts on parameter estimation and overall model fit.

*Sample Size*. We chose seven sample sizes to represent those commonly encountered in applied research and these range from very small to large: 50, 75, 100, 200, 400, 800, and 1000.

*Data Generation and Estimation*. We used the simulation feature in Version 5 of EQS (Bentler, 1995) to generate the data and EQS's maximum likelihood estimation to estimate the model. The data generation and estimation procedure was comprised of three basic steps. First, the population covariance matrix was computed to correspond to the parameterization of each of the three target models. Second, raw data were randomly generated from a multivariate normal distribution to correspond to the structure of the population covariance matrix. Finally, the particular specification within each target model was fit to the simulated raw data using

ML estimation. We used the population values for each parameter as initial start values, and we allowed a maximum of 100 iterations to achieve convergence. This method permitted us to fit a model that differed in structure from the model that generated the data . See Bentler (1995) for further details about EQS data generation procedures.

*Distribution.* We generated data from a multivariate normal distribution.

*Replications.* There were a total of 105 experimental conditions (three models, five specifications, and seven sample sizes), and we generated up to 500 replications for each condition.

*Convergence.* We eliminated any replication that failed to converge within 100 iterations, or did converge but resulted in an out of bounds parameter estimate (e.g., "Heywood Case") or a linear dependency among parameters. To maintain 500 replications per condition, we generated an initial set of up to 650 replications. We then fit the models to the generated data and selected the first 500 proper solutions, or selected as many proper solutions as existed when the total number of replications was reached. This resulted in 500 proper solutions for all properly specified and most misspecified experimental conditions, but there were several misspecified conditions that resulted in fewer than 500 proper solutions. Of the 105 experimental conditions, 82 (78%) contained 500 replications and 23 (22%) contained fewer than 500 replications. Of those 23 conditions containing fewer than 500 replications, the number of replications ranged from 443 to 499 with a median of 492, and the smallest number of 443 replications was for Model 3, Specification 4, $N = 50$. Whether improper solutions should be excluded or removed from the simulation design is a debatable issue. We chose to exclude improper solutions to mimic the lowered chance of results with improper solutions being reported. Fortunately, no differences were found in any results when including or excluding improper solutions (see Anderson & Gerbing, 1984, for further discussion of this topic). Elsewhere we have examined the causes and consequences of improper solutions in more detail ( Chen, Bollen, Paxton, Curran, & Kirby, 2001).

*Outcome Measures.* The outcome measures of key interest here is the likelihood ratio test statistic $T$ (commonly referred to as the "model $\chi^2$") estimated for each replicated model and the corresponding degrees of freedom for the estimated model. We obtained these values directly from EQS in which the $T$ statistic is computed as the product of the minimum of the *ML* fit function and $N$ - 1, and the *df* is computed as the difference between the total number of unique variances and covariances minus the total number of estimated parameters.

*Simulated Data Drawn from the Noncentral Chi-Square Distribution*

*Data Generation*. A key research question posed here is whether the model test statistics *T* computed from the simulated structural equation models follow the expected underlying central or noncentral chi-square distribution. As part of the empirical evaluation of this question, we generated additional simulated data drawn directly from the *known* population chi-square distributions using Version 7.0 of the SAS data system (SAS Inc., 1999). As mentioned earlier, we did this to allow for nonparametric tests that do *not* assume that the test statistics are normally distributed, a condition that likely does not hold here. We generated random variates from expected population distributions using a combination of the gamma and normal distribution functions in SAS. When $\lambda$ was zero, this resulted in random draws from the central chi-square distribution; when $\lambda$ was greater than zero, this resulted in random draws from the noncentral chi-square distribution.

*Experimental Conditions*. The mean and variance of the central and noncentral chi-square distributions vary as a function of degrees of freedom and the noncentrality parameter $\lambda$. Thus, we drew 105 separate samples from 105 *different* population distributions, one corresponding to each SEM experimental condition under study.

*Replications*. To achieve stable sample estimates of the underlying population distributions, we made 5000 draws for each of the 105 experimental conditions. Thus, all means and variances reported below are based on 5000 independent draws for each experimental condition.

*Summary*

In sum, we generated two complete sets of data to empirically evaluate our proposed research hypotheses. The first set was comprised of up to 500 test statistics *T* (one drawn from each SEM replication) estimated within each of 105 experimental conditions; we refer to these data as the SEM *simulations*. The second set was comprised of 5000 random draws from 105 different population central (for properly specified models) and noncentral (for misspecified models) chi-square distributions, one distribution corresponding to each SEM experimental condition under study; we refer to these data as the *chi-square simulations*. The core of our data analytic strategy is (a) the parametric comparison of the sample means and variances of the *T* statistics from the SEM simulations with the known population counterparts, and (b) the nonparametric comparison of the means and variances of the *T* statistics from the SEM simulations with the random chi-square variates drawn from the known population distributions.

## *Results*

We empirically evaluated the proposed research hypotheses using three related methods. First, we computed one-sample *z*-tests of the sample mean of the SEM test statistics to the corresponding population mean with known population variance, and we used one-sample $\chi^2$ tests of the sample variance of the SEM test statistics to the corresponding population variance. These are parametric tests of the null hypothesis that the sample mean and variance of the SEM test statistics equals the population mean and variance of the expected underlying distributions, and both tests assume that the population is normally distributed (Kanji, 1993). Second, because of the assumption of normality associated with the parametric tests, a condition that is not expected to hold here,[2] we used the nonparametric Wilcoxon Rank-Sum test of means and Siegel-Tukey test of variances to compare the empirical distributions of the SEM statistics with the corresponding empirical distributions from the chi-square simulations. The Wilcoxon Rank-Sum test evaluates the hypothesis that two random samples came from two populations with the same mean, and the Siegel-Tukey test evaluates the hypothesis that two random samples came from two populations with the same variance. Both of these nonparametric tests only assume that the two populations have continuous frequency distributions (Kanji, 1993, p. 86). Finally, to augment the parametric and nonparametric statistical tests, we computed effect sizes based on absolute relative bias (observed value minus expected value divided by expected value) and considered values of 5% or greater to indicate meaningful bias. In sum, we utilized parametric tests, nonparametric tests, and measures of effect size to evaluate our research hypotheses.

### *Tests of Central Tendency*

*One Sample z-Test.* Table 1 presents all summary statistics and the results of the parametric (*z*) and nonparametric (Wilcoxon) tests for the *means* of the population distribution, the SEM simulations, and the chi-square simulations. For each of the 105 experimental conditions, we compared the sample mean of the SEM test statistic *T* to the mean of the population

---

[2] We do not expect the assumption of normality to hold here because the *T* statistics are expected to follow a central or noncentral chi-square distribution which itself is not normal, at least under the conditions studied here. However, 103 of the 105 measures of univariate kurtosis of the sample distributions of *T* were below 1.0, and the largest value of kurtosis was 1.1 (Model 3, Specification 1, $N = 100$). Based on these empirical results, it does not appear that the assumption of normality is excessively violated here.

distribution that it is expected to follow given a known population variance. To control for inflated familywise error rate stemming from the 105 mean comparisons, we set the per comparison rate to $\alpha$ = .001 to maintain a familywise rate of approximately $\alpha$ = .10. Using this significance criterion,[3] a rather clear pattern of results emerges across all conditions: the mean of the SEM test statistics systematically overestimated the mean of the expected underlying population distributions at the smaller sample sizes across all three model types. For Specifications 1 through 4 (the one proper and three improper specifications) of Model 1, this tended to occur at sample sizes of 100 and below, and for Specifications 1 through 4 of Models 2 and 3, this tended to occur at sample sizes of 200 and below. Interestingly, there was little evidence of bias in the mean estimate for Specification 5 (the uncorrelated variables model) for Model 1, and there was modest overestimation for Specification 5 of Models 2 and 3, but only at the smallest sample size of 50.

*Relative Bias*. To further understand these relations in terms of effect sizes, relative bias was computed for all conditions. For Model 1, significant relative bias in the means (e.g., greater than 5%) was observed at sample sizes of $N$ = 100 and below for the proper specification, at $N$ = 50 for the improper specifications, and no bias was observed for the uncorrelated variables null model. For Model 2, the significant relative bias in the means was observed for both the proper and improper specifications at $N$ = 100 and below, and again there was no appreciable bias for the null model. This pattern was also found for Model 3 but only at sample sizes of $N$ = 75 and below. In general, the experimental conditions associated with significant relative bias closely corresponded with those conditions identified using the parametric $z$-test, but the relative bias results were somewhat more conservative compared to the parametric results. Thus, based on the 5% relative bias criterion, the means were systematically overestimated at the smaller sample sizes for the proper and improper model specifications, and the mean for the null model was unbiased across all sample sizes.

*Wilcoxon Rank-Sum Test*. The Wilcoxon Rank Sum test compared the mean of the SEM test statistics with the mean of the 5000 random variates drawn from the corresponding population distribution that the test statistics are expected to follow. Again, we used this method of comparison given the

---

[3] It could be argued that there are actually 210 total tests (e.g., 105 tests of mean and 105 tests of variance), or even 420 total tests given the inclusion of the nonparametric mean and variance comparisons. We chose to correct for 105 tests because these were the total number of comparisons that focused on one particular parameter within one particular statistical test. However, we present exact $p$-values for each individual test so that the reader may make any correction they so desire.

Table 1
Parametric and Nonparametric Tests of Mean Differences in Test Statistic *T*

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Mean of *T* | Mean of SEM Simulated *T* | *z*-test probability | Mean of CHI Simulated *T* | Wilcoxon Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 50 | 22 | 0.0000 | 22.0000 | 23.6400 | 0.0000 | 22.0400 | 0.0000 | 7.4685 |
| 1 | 1 | 75 | 22 | 0.0000 | 22.0000 | 23.0100 | 0.0003 | 21.9600 | 0.0002 | 4.5920 |
| 1 | 1 | 100 | 22 | 0.0000 | 22.0000 | 23.3200 | 0.0000 | 21.9500 | 0.0001 | 6.0201 |
| 1 | 1 | 200 | 22 | 0.0000 | 22.0000 | 22.5800 | 0.0254 | 21.9200 | 0.0349 | 2.6374 |
| 1 | 1 | 400 | 22 | 0.0000 | 22.0000 | 22.1800 | 0.2777 | 21.9400 | 0.2961 | 0.7960 |
| 1 | 1 | 800 | 22 | 0.0000 | 22.0000 | 21.8900 | 0.6394 | 22.0200 | 0.8597 | -0.4815 |
| 1 | 1 | 1000 | 22 | 0.0000 | 22.0000 | 21.7400 | 0.8121 | 22.0900 | 0.3244 | -1.1953 |
| 1 | 2 | 50 | 23 | 0.8100 | 23.8100 | 25.7000 | 0.0000 | 23.7800 | 0.0000 | 7.9291 |
| 1 | 2 | 75 | 23 | 1.2300 | 24.2300 | 25.3000 | 0.0004 | 24.2700 | 0.0015 | 4.4280 |
| 1 | 2 | 100 | 23 | 1.6400 | 24.6400 | 25.9600 | 0.0000 | 24.8300 | 0.0030 | 5.3420 |
| 1 | 2 | 200 | 23 | 3.3000 | 26.3000 | 26.6900 | 0.1258 | 26.2500 | 0.2447 | 1.5014 |
| 1 | 2 | 400 | 23 | 6.6100 | 29.6100 | 30.1400 | 0.0798 | 29.6700 | 0.1469 | 1.8094 |
| 1 | 2 | 800 | 23 | 13.2300 | 36.2300 | 36.5900 | 0.2082 | 36.4800 | 0.7618 | 0.9986 |
| 1 | 2 | 1000 | 23 | 16.5400 | 39.5400 | 38.4900 | 0.9867 | 39.6300 | 0.0240 | -2.6593 |
| 1 | 3 | 50 | 24 | 1.8400 | 25.8400 | 27.5000 | 0.0000 | 25.9500 | 0.0001 | 6.4011 |
| 1 | 3 | 75 | 24 | 2.7800 | 26.7800 | 27.6000 | 0.0089 | 26.6000 | 0.0028 | 3.0457 |
| 1 | 3 | 100 | 24 | 3.7200 | 27.7200 | 28.9600 | 0.0003 | 27.6200 | 0.0003 | 4.4413 |
| 1 | 3 | 200 | 24 | 7.4900 | 31.4900 | 31.8400 | 0.1859 | 31.4000 | 0.2594 | 1.1208 |
| 1 | 3 | 400 | 24 | 15.0100 | 39.0100 | 39.3200 | 0.2542 | 39.1400 | 0.5489 | 0.7888 |
| 1 | 3 | 800 | 24 | 30.0600 | 54.0600 | 54.2600 | 0.3683 | 54.0700 | 0.6700 | 0.3614 |
| 1 | 3 | 1000 | 24 | 37.5900 | 61.5900 | 60.1900 | 0.9864 | 61.4500 | 0.0477 | -2.2619 |
| 1 | 4 | 50 | 25 | 4.6300 | 29.6300 | 31.3200 | 0.0000 | 29.7200 | 0.0000 | 5.7103 |

Table 1 (*cont.*)

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Mean of $T$ | Mean of SEM Simulated $T$ | $z$-test probability | Mean of CHI Simulated $T$ | Wilcoxon Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 75 | 25 | 6.9900 | 31.9900 | 32.4200 | 0.1357 | 32.1200 | 0.2717 | 1.3592 |
| 1 | 4 | 100 | 25 | 9.3500 | 34.3500 | 35.8300 | 0.0002 | 34.4500 | 0.0041 | 4.3067 |
| 1 | 4 | 200 | 25 | 18.7900 | 43.7900 | 44.1700 | 0.2205 | 43.7000 | 0.5109 | 0.8813 |
| 1 | 4 | 400 | 25 | 37.6700 | 62.6700 | 63.5000 | 0.0961 | 62.7100 | 0.1765 | 1.3197 |
| 1 | 4 | 800 | 25 | 75.4400 | 100.4400 | 100.0100 | 0.6919 | 100.5600 | 0.5367 | -0.4189 |
| 1 | 4 | 1000 | 25 | 94.3200 | 119.3200 | 117.5600 | 0.9716 | 119.5800 | 0.0472 | -1.4774 |
| 1 | 5 | 50 | 36 | 174.8600 | 210.8600 | 212.5600 | 0.0855 | 210.6400 | 0.2460 | 0.8072 |
| 1 | 5 | 75 | 36 | 264.0700 | 300.0700 | 301.5200 | 0.1675 | 300.3500 | 0.6139 | 0.4831 |
| 1 | 5 | 100 | 36 | 353.2800 | 389.2800 | 391.1100 | 0.1447 | 389.7600 | 0.4539 | 0.4694 |
| 1 | 5 | 200 | 36 | 710.1300 | 746.1300 | 746.8800 | 0.3792 | 747.4200 | 0.7739 | 0.0996 |
| 1 | 5 | 400 | 36 | 1423.8300 | 1459.8300 | 1455.0600 | 0.9198 | 1459.2800 | 0.3298 | -0.3269 |
| 1 | 5 | 800 | 36 | 2851.2400 | 2887.2400 | 2880.4600 | 0.9213 | 2889.3100 | 0.0412 | -0.2348 |
| 1 | 5 | 1000 | 36 | 3564.9400 | 3600.9400 | 3594.5200 | 0.8845 | 3597.8900 | 0.4085 | -0.1783 |
| 2 | 1 | 50 | 85 | 0.0000 | 85.0000 | 98.7900 | 0.0000 | 84.8900 | 0.0000 | 16.2202 |
| 2 | 1 | 75 | 85 | 0.0000 | 85.0000 | 93.2300 | 0.0000 | 85.0400 | 0.0000 | 9.6850 |
| 2 | 1 | 100 | 85 | 0.0000 | 85.0000 | 91.4000 | 0.0000 | 84.9700 | 0.0000 | 7.5334 |
| 2 | 1 | 200 | 85 | 0.0000 | 85.0000 | 88.4600 | 0.0000 | 85.2900 | 0.0000 | 4.0757 |
| 2 | 1 | 400 | 85 | 0.0000 | 85.0000 | 86.0300 | 0.0381 | 85.3100 | 0.1707 | 1.2176 |
| 2 | 1 | 800 | 85 | 0.0000 | 85.0000 | 85.3600 | 0.2667 | 85.0700 | 0.5346 | 0.4277 |
| 2 | 1 | 1000 | 85 | 0.0000 | 85.0000 | 85.8600 | 0.0711 | 85.0300 | 0.1496 | 1.0077 |
| 2 | 2 | 50 | 86 | 1.9400 | 87.9400 | 101.8400 | 0.0000 | 87.8400 | 0.0000 | 15.8099 |
| 2 | 2 | 75 | 86 | 2.9300 | 88.9300 | 97.0600 | 0.0000 | 89.0500 | 0.0000 | 9.1429 |
| 2 | 2 | 100 | 86 | 3.9200 | 89.9200 | 96.1700 | 0.0000 | 89.6900 | 0.0000 | 6.9527 |

Table 1 (*cont.*)

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Mean of *T* | Mean of SEM Simulated *T* | *z*-test probability | Mean of CHI Simulated *T* | Wilcoxon Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 200 | 86 | 7.8800 | 93.8800 | 97.1900 | 0.0000 | 93.7100 | 0.0000 | 3.5210 |
| 2 | 2 | 400 | 86 | 15.8100 | 101.8100 | 102.4300 | 0.1812 | 101.9000 | 0.4779 | 0.6144 |
| 2 | 2 | 800 | 86 | 31.6500 | 117.6500 | 118.6300 | 0.1026 | 117.2700 | 0.1185 | 0.8331 |
| 2 | 2 | 1000 | 86 | 39.5800 | 125.5800 | 126.3000 | 0.1890 | 125.3900 | 0.4919 | 0.5713 |
| 2 | 3 | 50 | 87 | 4.0500 | 91.0500 | 105.1300 | 0.0000 | 90.9600 | 0.0000 | 15.4536 |
| 2 | 3 | 75 | 87 | 6.1200 | 93.1200 | 101.1700 | 0.0000 | 93.4200 | 0.0000 | 8.6425 |
| 2 | 3 | 100 | 87 | 8.1900 | 95.1900 | 101.6700 | 0.0000 | 95.3300 | 0.0000 | 6.8097 |
| 2 | 3 | 200 | 87 | 16.4700 | 103.4700 | 106.6500 | 0.0000 | 103.4800 | 0.0001 | 3.0729 |
| 2 | 3 | 400 | 87 | 33.0200 | 120.0200 | 120.2800 | 0.3672 | 120.1900 | 0.7530 | 0.2214 |
| 2 | 3 | 800 | 87 | 66.1200 | 153.1200 | 154.1100 | 0.1457 | 153.0800 | 0.2725 | 0.6460 |
| 2 | 3 | 1000 | 87 | 82.6700 | 169.6700 | 169.7100 | 0.4814 | 170.1300 | 0.6981 | 0.0276 |
| 2 | 4 | 50 | 88 | 6.9000 | 94.9000 | 108.7800 | 0.0000 | 95.0600 | 0.0000 | 14.6254 |
| 2 | 4 | 75 | 88 | 10.4200 | 98.4200 | 106.4500 | 0.0000 | 98.4900 | 0.0000 | 8.1602 |
| 2 | 4 | 100 | 88 | 13.9400 | 101.9400 | 108.3100 | 0.0000 | 102.2500 | 0.0000 | 6.2489 |
| 2 | 4 | 200 | 88 | 28.0200 | 116.0200 | 118.5000 | 0.0005 | 116.0000 | 0.0039 | 2.1392 |
| 2 | 4 | 400 | 88 | 56.1800 | 144.1800 | 144.7500 | 0.2638 | 144.3400 | 0.5801 | 0.3927 |
| 2 | 4 | 800 | 88 | 112.5000 | 200.5000 | 201.2600 | 0.2506 | 199.6300 | 0.1769 | 0.3757 |
| 2 | 4 | 1000 | 88 | 140.6700 | 228.6700 | 229.3200 | 0.2963 | 229.1600 | 0.9196 | 0.2848 |
| 2 | 5 | 50 | 105 | 328.0000 | 433.0000 | 449.3800 | 0.0000 | 433.0500 | 0.0000 | 3.7841 |
| 2 | 5 | 75 | 105 | 495.3400 | 600.3400 | 606.5100 | 0.0016 | 598.9300 | 0.0164 | 1.0276 |
| 2 | 5 | 100 | 105 | 662.6900 | 767.6900 | 774.1900 | 0.0033 | 768.7500 | 0.0141 | 0.8469 |
| 2 | 5 | 200 | 105 | 1332.0700 | 1437.0700 | 1437.1300 | 0.4928 | 1438.0400 | 0.7900 | 0.0042 |
| 2 | 5 | 400 | 105 | 2670.8400 | 2775.8400 | 2779.0200 | 0.2481 | 2777.1000 | 0.7097 | 0.1145 |

Table 1 (*cont.*)

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Mean of *T* | Mean of SEM Simulated *T* | *z*-test probability | Mean of CHI Simulated *T* | Wilcoxon Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 5 | 800 | 105 | 5348.3800 | 5453.3800 | 5439.6900 | 0.9813 | 5455.4400 | 0.1898 | -0.2510 |
| 2 | 5 | 1000 | 105 | 6687.1500 | 6792.1500 | 6762.7300 | 1.0000 | 6791.7900 | 0.0004 | -0.4331 |
| 3 | 1 | 50 | 50 | 0.0000 | 50.0000 | 57.6200 | 0.0000 | 50.0000 | 0.0000 | 15.2404 |
| 3 | 1 | 75 | 50 | 0.0000 | 50.0000 | 54.7100 | 0.0000 | 50.0400 | 0.0000 | 9.4135 |
| 3 | 1 | 100 | 50 | 0.0000 | 50.0000 | 52.8100 | 0.0000 | 49.8700 | 0.0000 | 5.6112 |
| 3 | 1 | 200 | 50 | 0.0000 | 50.0000 | 52.2200 | 0.0000 | 49.7200 | 0.0000 | 4.4418 |
| 3 | 1 | 400 | 50 | 0.0000 | 50.0000 | 50.2700 | 0.2714 | 49.9200 | 0.2255 | 0.5450 |
| 3 | 1 | 800 | 50 | 0.0000 | 50.0000 | 50.0000 | 0.4957 | 50.0900 | 0.9396 | 0.0097 |
| 3 | 1 | 1000 | 50 | 0.0000 | 50.0000 | 50.3000 | 0.2518 | 49.9400 | 0.4103 | 0.5990 |
| 3 | 2 | 50 | 53 | 6.1900 | 59.1900 | 67.4100 | 0.0000 | 59.2100 | 0.0000 | 13.8866 |
| 3 | 2 | 75 | 53 | 9.3400 | 62.3400 | 66.8500 | 0.0000 | 62.5500 | 0.0000 | 7.2365 |
| 3 | 2 | 100 | 53 | 12.5000 | 65.5000 | 68.3300 | 0.0000 | 65.2500 | 0.0000 | 4.3299 |
| 3 | 2 | 200 | 53 | 25.1200 | 78.1200 | 80.0900 | 0.0011 | 78.1100 | 0.0058 | 2.5117 |
| 3 | 2 | 400 | 53 | 50.3700 | 103.3700 | 102.5100 | 0.8649 | 103.2500 | 0.2933 | -0.8372 |
| 3 | 2 | 800 | 53 | 100.8700 | 153.8700 | 155.7200 | 0.0337 | 153.9800 | 0.1840 | 1.2008 |
| 3 | 2 | 1000 | 53 | 126.1200 | 179.1200 | 177.7700 | 0.8892 | 179.4900 | 0.2302 | -0.7546 |
| 3 | 3 | 50 | 54 | 18.8400 | 72.8400 | 81.6200 | 0.0000 | 72.9300 | 0.0000 | 12.0477 |
| 3 | 3 | 75 | 54 | 28.4600 | 82.4600 | 87.8700 | 0.0000 | 82.5300 | 0.0000 | 6.5649 |
| 3 | 3 | 100 | 54 | 38.0700 | 92.0700 | 94.3800 | 0.0007 | 91.9100 | 0.0012 | 2.5074 |
| 3 | 3 | 200 | 54 | 76.5300 | 130.5300 | 132.5600 | 0.0128 | 130.5200 | 0.0264 | 1.5570 |
| 3 | 3 | 400 | 54 | 153.4400 | 207.4400 | 207.7600 | 0.3953 | 207.8700 | 0.8725 | 0.1539 |
| 3 | 3 | 800 | 54 | 307.2700 | 361.2700 | 359.7000 | 0.8324 | 361.9300 | 0.4340 | -0.4367 |
| 3 | 3 | 1000 | 54 | 384.1900 | 438.1900 | 439.4700 | 0.2407 | 437.9700 | 0.3512 | 0.2917 |

Table 1 (*cont.*)

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Mean of $T$ | Mean of SEM Simulated $T$ | $z$-test probability | Mean of CHI Simulated $T$ | Wilcoxon Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 4 | 50 | 57 | 26.3100 | 83.3100 | 92.1000 | 0.0000 | 83.4900 | 0.0000 | 10.5594 |
| 3 | 4 | 75 | 57 | 39.7300 | 96.7300 | 101.6000 | 0.0000 | 96.7700 | 0.0000 | 5.0354 |
| 3 | 4 | 100 | 57 | 53.1500 | 110.1500 | 112.4800 | 0.0020 | 109.8600 | 0.0015 | 2.1179 |
| 3 | 4 | 200 | 57 | 106.8400 | 163.8400 | 165.3500 | 0.0733 | 164.2900 | 0.3195 | 0.9227 |
| 3 | 4 | 400 | 57 | 214.2200 | 271.2200 | 270.1000 | 0.7875 | 270.9900 | 0.4441 | -0.4103 |
| 3 | 4 | 800 | 57 | 428.9700 | 485.9700 | 486.4400 | 0.4022 | 485.8500 | 0.5837 | 0.0976 |
| 3 | 4 | 1000 | 57 | 536.3400 | 593.3400 | 592.8300 | 0.5954 | 593.1200 | 0.7477 | -0.0866 |
| 3 | 5 | 50 | 78 | 268.8100 | 346.8100 | 361.2000 | 0.0000 | 346.9100 | 0.0000 | 4.1472 |
| 3 | 5 | 75 | 78 | 405.9600 | 483.9600 | 492.1800 | 0.0000 | 484.2200 | 0.0018 | 1.6975 |
| 3 | 5 | 100 | 78 | 543.1100 | 621.1100 | 623.7800 | 0.1082 | 621.0800 | 0.3759 | 0.4299 |
| 3 | 5 | 200 | 78 | 1091.7100 | 1169.7100 | 1174.3800 | 0.0606 | 1168.6900 | 0.1011 | 0.3989 |
| 3 | 5 | 400 | 78 | 2188.9100 | 2266.9100 | 2271.5500 | 0.1364 | 2267.2100 | 0.3924 | 0.2044 |
| 3 | 5 | 800 | 78 | 4383.3100 | 4461.3100 | 4468.9000 | 0.1014 | 4460.5300 | 0.1779 | 0.1700 |
| 3 | 5 | 1000 | 78 | 5480.5100 | 5558.5100 | 5567.3400 | 0.0924 | 5560.0000 | 0.2206 | 0.1587 |

likely failure of the *T* statistics to meet the assumption of normality required by the *z*-test. As expected, the nonparametric tests tended to demonstrate lower power relative to the parametric counterparts. However, nearly without exception, every condition for which a meaningful difference results using the parametric tests, this same condition is identified in the nonparametric tests (based on the corrected $\alpha = .001$).

*Summary of Tests of Central Tendency*. Based on the corrected significance levels of the parametric and nonparametric tests as well as the magnitude of relative bias, we concluded that the mean of the SEM test statistics consistently overestimated the mean of the expected underlying population distribution for Specifications 1 through 4 for all three model types (the properly specified and three misspecified conditions), but only at the smallest sample sizes (e.g., 100 to 200 and below). At samples above 200, we found no appreciable bias across any condition. Further, we found no significant overestimation of the population mean for Specification 5 (the uncorrelated variables model) for any model type at any sample size.

## *Tests of Dispersion*

*One Sample* $\chi^2$*-Test*. Table 2 presents all summary statistics, parametric ($\chi^2$) and nonparametric (Siegel-Tukey) tests, and relative bias for the *variances* of the population distribution, the SEM simulations, and the chi-square simulations. Again using a per comparison rate of $\alpha = .001$ to control for multiple testing, the variance of the *T* statistics for Model 1 only significantly varied from the expected population value at the smallest sample size for the properly specified and the most minor improper specification (Specifications 1 and 2); in contrast, the variance was significantly overestimated across *all* sample sizes for the uncorrelated variable baseline model (Specification 5). A similar pattern of results was found for both Model 2 and Model 3. Thus, although there was evidence of significant overestimation of the sample variance of *T* relative to the expected underlying distribution at $N = 50$ for the proper and minor improper specifications, the variance of the uncorrelated variable baseline model was significantly overestimated at every sample size for every model type even using the adjusted $\alpha = .001$ per comparison error rate.

*Relative Bias*. Unlike the tests of central tendency that indicated a larger number of biased conditions based on the parametric test results compared to the relative bias results, for the tests of dispersion a larger number of biased conditions were identified based upon the relative bias results compared to the parametric test results. For Specifications 1 through 4 for all three model types, the variance of the SEM test statistics overestimated the expected variance of the population distributions at the smaller sample sizes. Relative bias exceeded 5% at samples of $N = 100$ and below for Model 1,[4] at about $N = 200$ and below for Model 2, and at about $N = 75$ and below for Model 3. Indeed, at the smallest sample size of $N = 50$, relative bias ranged from 12% up to 39% indicating substantial overestimation of the corresponding population parameter.[5] Of key interest is the finding that for Specification 5 (the uncorrelated variables model), the sample variance of the SEM test statistics significantly overestimated the population counterpart at every sample size across all three model types with relative bias ranging from a minimum of 32% to a remarkable 164%. Indeed, for Model 2, bias was 120% or higher across all sample sizes. Thus, for the uncorrelated variables model, there was not one instance in which the sample estimates of the SEM test statistics showed evidence of following the dispersion of the expected noncentral chi-square population distribution.

*Siegel-Tukey Test*. As with the *z*-test of the means, the parametric $\chi^2$ test of the variances also assumes normality thus necessitating the use of the nonparametric equivalents. The Siegel-Tukey nonparametric test of variance (Kanji, 1993, p. 87) compared the variance of the SEM test statistics and the variance of the $N = 5000$ random variates drawn from the expected underlying population distribution, and this test only assumes that the population distributions are continuous. In general, as was found with the Wilcoxon Rank-Sum test of central tendency, the results of the Siegel-Tukey test of dispersion closely corresponded with those of the $\chi^2$ test, although again there is some evidence of the lower power of the nonparametric test. In general, the Siegel-Tukey results indicated overestimation of the variance at the smallest sample size for nearly all of the properly and improperly specified models, and indicated significant overestimation at all sample sizes for all three model types for the uncorrelated variable baseline model.

*Summary of Tests of Dispersion*. Results from the parametric and nonparametric tests in combination with the magnitude of the absolute relative bias lead us to two key patterns of results. First, for Specifications

---

[4] An odd pattern of findings was evident for the first four Specifications of Model 1 at $N = 75$ in which the estimated variance of the SEM test statistics was smaller at $N = 75$ compared to the $N = 50$ and $N = 100$ conditions. We suspected that this was an error in data generation, but extensive exploration of these conditions coupled with the generation of additional data revealed no errors. We found that there is much sampling variability in the estimation of the variance of the SEM test statistics at the smaller sample sizes, and the somewhat odd pattern of results for this one particular condition is most likely attributable to this random variability.

[5] One interesting finding to note is that at $N = 400$ and $N = 800$ of Specification 3 of Model 3, the relative bias was actually a *negative* value (both approximately -18%). This finding was not predicted, but also was not consistent across model or specification. It is thus not immediately clear what this limited evidence of underestimation of variance implies, if anything at all.

Table 2
Parametric and Nonparametric Tests of Variance Differences in Test Statistic *T*

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Variance of *T* | Variance of SEM Simulated *T* | Chi-Square Probability | Variance of CHI Simulated *T* | Siegel-Tukey Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 50 | 22 | 0.0000 | 44.0000 | 54.2400 | 0.0003 | 45.3600 | 0.1603 | 23.2730 |
| 1 | 1 | 75 | 22 | 0.0000 | 44.0000 | 45.9400 | 0.2398 | 43.4500 | 0.5219 | 4.4060 |
| 1 | 1 | 100 | 22 | 0.0000 | 44.0000 | 52.1000 | 0.0028 | 43.2300 | 0.0278 | 18.4080 |
| 1 | 1 | 200 | 22 | 0.0000 | 44.0000 | 45.1500 | 0.3328 | 43.6800 | 0.9937 | 2.6250 |
| 1 | 1 | 400 | 22 | 0.0000 | 44.0000 | 42.5200 | 0.6969 | 45.3400 | 0.4548 | -3.3590 |
| 1 | 1 | 800 | 22 | 0.0000 | 44.0000 | 40.3800 | 0.9056 | 44.0700 | 0.6296 | -8.2190 |
| 1 | 1 | 1000 | 22 | 0.0000 | 44.0000 | 44.4700 | 0.4254 | 45.1700 | 0.1883 | 1.0610 |
| 1 | 2 | 50 | 23 | 0.8100 | 49.2500 | 61.0200 | 0.0002 | 51.0400 | 0.1190 | 23.8990 |
| 1 | 2 | 75 | 23 | 1.2300 | 50.9000 | 52.8600 | 0.2671 | 50.3100 | 0.8613 | 3.8510 |
| 1 | 2 | 100 | 23 | 1.6400 | 52.5600 | 60.3100 | 0.0124 | 52.9300 | 0.1180 | 14.7500 |
| 1 | 2 | 200 | 23 | 3.3000 | 59.1800 | 58.9700 | 0.5144 | 58.7300 | 0.9289 | -0.3610 |
| 1 | 2 | 400 | 23 | 6.6100 | 72.4300 | 74.8300 | 0.2951 | 72.8000 | 0.2585 | 3.3130 |
| 1 | 2 | 800 | 23 | 13.2300 | 98.9300 | 101.2600 | 0.3482 | 99.2900 | 0.8229 | 2.3550 |
| 1 | 2 | 1000 | 23 | 16.5400 | 112.1800 | 113.8800 | 0.3980 | 109.1400 | 0.0148 | 1.5110 |
| 1 | 3 | 50 | 24 | 1.8400 | 55.3700 | 64.1200 | 0.0083 | 55.7800 | 0.0108 | 15.7870 |
| 1 | 3 | 75 | 24 | 2.7800 | 59.1400 | 59.6700 | 0.4358 | 58.8000 | 0.8473 | 0.8930 |
| 1 | 3 | 100 | 24 | 3.7200 | 62.9000 | 69.3700 | 0.0556 | 61.5700 | 0.0063 | 10.2820 |
| 1 | 3 | 200 | 24 | 7.4900 | 77.9500 | 74.9400 | 0.7248 | 77.9300 | 0.2978 | -3.8630 |
| 1 | 3 | 400 | 24 | 15.0100 | 108.0500 | 111.4400 | 0.3047 | 109.3600 | 0.4540 | 3.1340 |
| 1 | 3 | 800 | 24 | 30.0600 | 168.2500 | 169.2400 | 0.4547 | 169.6300 | 0.8714 | 0.5890 |
| 1 | 3 | 1000 | 24 | 37.5900 | 198.3500 | 206.4500 | 0.2556 | 198.3500 | 0.2829 | 4.0810 |
| 1 | 4 | 50 | 25 | 4.6300 | 68.5000 | 76.7000 | 0.0329 | 66.9800 | 0.0391 | 11.9620 |

Table 2 (*cont.*)

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Variance of *T* | Variance of SEM Simulated *T* | Chi-Square Probability | Variance of CHI Simulated *T* | Siegel-Tukey Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 75 | 25 | 6.9900 | 77.9500 | 79.6300 | 0.3595 | 77.0900 | 0.4901 | 2.1590 |
| 1 | 4 | 100 | 25 | 9.3500 | 87.3900 | 97.9600 | 0.0315 | 87.6500 | 0.1980 | 12.0960 |
| 1 | 4 | 200 | 25 | 18.7900 | 125.1500 | 127.3600 | 0.3830 | 125.0900 | 0.4716 | 1.7600 |
| 1 | 4 | 400 | 25 | 37.6700 | 200.6800 | 217.7000 | 0.0927 | 200.0400 | 0.3390 | 8.4790 |
| 1 | 4 | 800 | 25 | 75.4400 | 351.7400 | 359.1700 | 0.3622 | 358.9800 | 0.7278 | 2.1130 |
| 1 | 4 | 1000 | 25 | 94.3200 | 427.2700 | 455.1900 | 0.1512 | 417.5100 | 0.8286 | 6.5340 |
| 1 | 5 | 50 | 36 | 174.8600 | 771.4300 | 1253.0200 | 0.0000 | 770.8100 | 0.0000 | 62.4280 |
| 1 | 5 | 75 | 36 | 264.0700 | 1128.2800 | 1952.9000 | 0.0000 | 1119.1100 | 0.0000 | 73.0870 |
| 1 | 5 | 100 | 36 | 353.2800 | 1485.1300 | 2481.5300 | 0.0000 | 1514.3700 | 0.0000 | 67.0920 |
| 1 | 5 | 200 | 36 | 710.1300 | 2912.5300 | 4912.1500 | 0.0000 | 2852.9800 | 0.0000 | 68.6560 |
| 1 | 5 | 400 | 36 | 1423.8300 | 5767.3400 | 9303.8600 | 0.0000 | 5718.8400 | 0.0000 | 61.3200 |
| 1 | 5 | 800 | 36 | 2851.2400 | 11476.9500 | 20973.2300 | 0.0000 | 11448.6300 | 0.0000 | 82.7420 |
| 1 | 5 | 1000 | 36 | 3564.9400 | 14331.7600 | 27547.1200 | 0.0000 | 14342.5700 | 0.0000 | 92.2100 |
| 2 | 1 | 50 | 85 | 0.0000 | 170.0000 | 204.5800 | 0.0012 | 161.7900 | 0.0000 | 20.3400 |
| 2 | 1 | 75 | 85 | 0.0000 | 170.0000 | 201.1600 | 0.0029 | 172.9500 | 0.0015 | 18.3290 |
| 2 | 1 | 100 | 85 | 0.0000 | 170.0000 | 199.5700 | 0.0044 | 167.1200 | 0.0001 | 17.3920 |
| 2 | 1 | 200 | 85 | 0.0000 | 170.0000 | 190.5200 | 0.0317 | 170.7700 | 0.9153 | 12.0720 |
| 2 | 1 | 400 | 85 | 0.0000 | 170.0000 | 166.3800 | 0.6246 | 168.1100 | 0.3497 | -2.1280 |
| 2 | 1 | 800 | 85 | 0.0000 | 170.0000 | 159.1200 | 0.8442 | 169.4500 | 0.1888 | -6.3980 |
| 2 | 1 | 1000 | 85 | 0.0000 | 170.0000 | 161.9600 | 0.7696 | 166.8400 | 0.9936 | -4.7270 |
| 2 | 2 | 50 | 86 | 1.9400 | 179.7700 | 212.6000 | 0.0030 | 178.3700 | 0.0000 | 18.2640 |
| 2 | 2 | 75 | 86 | 2.9300 | 183.7300 | 211.0800 | 0.0118 | 179.6400 | 0.0003 | 14.8900 |

Table 2 (*cont.*)

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Variance of $T$ | Variance of SEM Simulated $T$ | Chi-Square Probability | Variance of CHI Simulated $T$ | Siegel-Tukey Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 100 | 86 | 3.9200 | 187.6900 | 220.8100 | 0.0039 | 183.1400 | 0.0002 | 17.6480 |
| 2 | 2 | 200 | 86 | 7.8800 | 203.5400 | 223.0700 | 0.0680 | 203.3100 | 0.7279 | 9.5970 |
| 2 | 2 | 400 | 86 | 15.8100 | 235.2300 | 235.5400 | 0.4832 | 229.5900 | 0.8319 | 0.1330 |
| 2 | 2 | 800 | 86 | 31.6500 | 298.6200 | 323.2900 | 0.0983 | 302.8700 | 0.8507 | 8.2610 |
| 2 | 2 | 1000 | 86 | 39.5800 | 330.3100 | 350.9000 | 0.1622 | 333.0800 | 0.0515 | 6.2310 |
| 2 | 3 | 50 | 87 | 4.0500 | 190.2200 | 224.0400 | 0.0037 | 189.0500 | 0.0000 | 17.7820 |
| 2 | 3 | 75 | 87 | 6.1200 | 198.4900 | 226.7700 | 0.0149 | 195.2300 | 0.0019 | 14.2450 |
| 2 | 3 | 100 | 87 | 8.1900 | 206.7700 | 248.9600 | 0.0012 | 209.0500 | 0.0840 | 20.4060 |
| 2 | 3 | 200 | 87 | 16.4700 | 239.8700 | 255.0900 | 0.1579 | 245.0600 | 0.1430 | 6.3480 |
| 2 | 3 | 400 | 87 | 33.0200 | 306.0700 | 330.3600 | 0.1069 | 314.6600 | 0.6377 | 7.9370 |
| 2 | 3 | 800 | 87 | 66.1200 | 438.4700 | 443.6700 | 0.4177 | 440.5300 | 0.7258 | 1.1870 |
| 2 | 3 | 1000 | 87 | 82.6700 | 504.6600 | 498.2200 | 0.5720 | 505.6500 | 0.3980 | -1.2770 |
| 2 | 4 | 50 | 88 | 6.9000 | 203.6000 | 232.9600 | 0.0140 | 208.5500 | 0.0000 | 14.4230 |
| 2 | 4 | 75 | 88 | 10.4200 | 217.6800 | 245.2100 | 0.0262 | 217.7000 | 0.0010 | 12.6500 |
| 2 | 4 | 100 | 88 | 13.9400 | 231.7600 | 275.2900 | 0.0024 | 233.3300 | 0.0847 | 18.7840 |
| 2 | 4 | 200 | 88 | 28.0200 | 288.0800 | 289.6000 | 0.4586 | 291.6300 | 0.0735 | 0.5260 |
| 2 | 4 | 400 | 88 | 56.1800 | 400.7300 | 454.7600 | 0.0196 | 410.3200 | 0.1268 | 13.4850 |
| 2 | 4 | 800 | 88 | 112.5000 | 626.0200 | 623.3800 | 0.5182 | 639.2300 | 0.8790 | -0.4210 |
| 2 | 4 | 1000 | 88 | 140.6700 | 738.6600 | 723.5900 | 0.6192 | 710.9000 | 0.3342 | -2.0400 |
| 2 | 5 | 50 | 105 | 328.0000 | 1521.9900 | 3483.9800 | 0.0000 | 1537.4900 | 0.0000 | 128.9090 |
| 2 | 5 | 75 | 105 | 495.3400 | 2191.3800 | 4915.4700 | 0.0000 | 2230.1000 | 0.0000 | 124.3100 |
| 2 | 5 | 100 | 105 | 662.6900 | 2860.7600 | 6252.5800 | 0.0000 | 2944.9900 | 0.0000 | 118.5640 |

Table 2 (*cont.*)

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Variance of *T* | Variance of SEM Simulated *T* | Chi-Square Probability | Variance of CHI Simulated *T* | Siegel-Tukey Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 5 | 200 | 105 | 1332.0700 | 5538.3000 | 12426.7500 | 0.0000 | 5465.9000 | 0.0000 | 124.3790 |
| 2 | 5 | 400 | 105 | 2670.8400 | 10893.3700 | 26462.4900 | 0.0000 | 10862.4600 | 0.0000 | 142.9230 |
| 2 | 5 | 800 | 105 | 5348.3800 | 21603.5200 | 57124.1300 | 0.0000 | 22203.6800 | 0.0000 | 164.4200 |
| 2 | 5 | 1000 | 105 | 6687.1500 | 26958.5900 | 67060.5700 | 0.0000 | 27060.6200 | 0.0000 | 148.7540 |
| 3 | 1 | 50 | 50 | 0.0000 | 100.0000 | 139.8500 | 0.0000 | 98.1200 | 0.0000 | 39.8530 |
| 3 | 1 | 75 | 50 | 0.0000 | 100.0000 | 121.4300 | 0.0007 | 99.7800 | 0.0133 | 21.4260 |
| 3 | 1 | 100 | 50 | 0.0000 | 100.0000 | 102.8200 | 0.3217 | 96.4100 | 0.4310 | 2.8240 |
| 3 | 1 | 200 | 50 | 0.0000 | 100.0000 | 109.2400 | 0.0752 | 97.2600 | 0.7120 | 9.2450 |
| 3 | 1 | 400 | 50 | 0.0000 | 100.0000 | 95.4900 | 0.7586 | 104.0700 | 0.7198 | -4.5070 |
| 3 | 1 | 800 | 50 | 0.0000 | 100.0000 | 90.9400 | 0.9272 | 101.3700 | 0.2008 | -9.0580 |
| 3 | 1 | 1000 | 50 | 0.0000 | 100.0000 | 100.3800 | 0.4678 | 100.0000 | 0.5712 | 0.3790 |
| 3 | 2 | 50 | 53 | 6.1900 | 130.7500 | 179.4400 | 0.0000 | 133.4800 | 0.0000 | 37.2470 |
| 3 | 2 | 75 | 53 | 9.3400 | 143.3700 | 163.4800 | 0.0162 | 143.5400 | 0.3343 | 14.0230 |
| 3 | 2 | 100 | 53 | 12.5000 | 156.0000 | 156.8200 | 0.4586 | 156.1400 | 0.8556 | 0.5260 |
| 3 | 2 | 200 | 53 | 25.1200 | 206.5000 | 219.3500 | 0.1625 | 213.0500 | 0.8804 | 6.2240 |
| 3 | 2 | 400 | 53 | 50.3700 | 307.5000 | 309.8000 | 0.4446 | 299.0600 | 0.2353 | 0.7500 |
| 3 | 2 | 800 | 53 | 100.8700 | 509.5000 | 515.1700 | 0.4222 | 504.1200 | 0.2938 | 1.1140 |
| 3 | 2 | 1000 | 53 | 126.1200 | 610.5000 | 684.9300 | 0.0305 | 607.1400 | 0.0999 | 12.1920 |
| 3 | 3 | 50 | 54 | 18.8400 | 183.3800 | 207.7300 | 0.0211 | 188.3700 | 0.0001 | 13.2790 |
| 3 | 3 | 75 | 54 | 28.4600 | 221.8300 | 239.3800 | 0.1077 | 221.1400 | 0.9915 | 7.9090 |
| 3 | 3 | 100 | 54 | 38.0700 | 260.2900 | 268.9700 | 0.2940 | 262.0100 | 0.7011 | 3.3330 |
| 3 | 3 | 200 | 54 | 76.5300 | 414.1200 | 397.9300 | 0.7273 | 408.2100 | 0.7543 | -3.9090 |

Table 2 (*cont.*)

| Model Type | Specification | Sample Size | Degrees of Freedom | Population Lambda | Population Variance of $T$ | Variance of SEM Simulated $T$ | Chi-Square Probability | Variance of CHI Simulated $T$ | Siegel-Tukey Probability | Relative Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 3 | 400 | 54 | 153.4400 | 721.7800 | 595.8200 | 0.9982 | 703.8100 | 0.0577 | -17.4510 |
| 3 | 3 | 800 | 54 | 307.2700 | 1337.1000 | 1090.2500 | 0.9990 | 1347.5600 | 0.0619 | -18.4620 |
| 3 | 3 | 1000 | 54 | 384.1900 | 1644.7600 | 1480.0200 | 0.9471 | 1703.4400 | 0.1788 | -10.0160 |
| 3 | 4 | 50 | 57 | 26.3100 | 219.2300 | 268.9900 | 0.0004 | 215.2800 | 0.0000 | 22.6990 |
| 3 | 4 | 75 | 57 | 39.7300 | 272.9200 | 310.8100 | 0.0170 | 277.8400 | 0.5676 | 13.8840 |
| 3 | 4 | 100 | 57 | 53.1500 | 326.6000 | 352.6400 | 0.1060 | 325.7000 | 0.6906 | 7.9720 |
| 3 | 4 | 200 | 57 | 106.8400 | 541.3600 | 535.1400 | 0.5639 | 541.8700 | 0.8566 | -1.1480 |
| 3 | 4 | 400 | 57 | 214.2200 | 970.8600 | 897.1700 | 0.8867 | 958.8900 | 0.3596 | -7.5900 |
| 3 | 4 | 800 | 57 | 428.9700 | 1829.8700 | 1641.0400 | 0.9524 | 1793.6000 | 0.3567 | -10.3190 |
| 3 | 4 | 1000 | 57 | 536.3400 | 2259.3800 | 2256.9700 | 0.4983 | 2234.9800 | 0.1605 | -0.1070 |
| 3 | 5 | 50 | 78 | 268.8100 | 1231.2600 | 1685.8300 | 0.0000 | 1240.6800 | 0.0000 | 36.9190 |
| 3 | 5 | 75 | 78 | 405.9600 | 1779.8600 | 2466.0300 | 0.0000 | 1821.3800 | 0.0000 | 38.5520 |
| 3 | 5 | 100 | 78 | 543.1100 | 2328.4600 | 3386.1900 | 0.0000 | 2323.3000 | 0.0000 | 45.4270 |
| 3 | 5 | 200 | 78 | 1091.7100 | 4522.8600 | 6992.3100 | 0.0000 | 4482.9000 | 0.0000 | 54.5990 |
| 3 | 5 | 400 | 78 | 2188.9100 | 8911.6600 | 12477.9900 | 0.0000 | 8664.6000 | 0.0000 | 40.0190 |
| 3 | 5 | 800 | 78 | 4383.3100 | 17689.2600 | 23401.8800 | 0.0000 | 17873.4400 | 0.0008 | 32.2940 |
| 3 | 5 | 1000 | 78 | 5480.5100 | 22078.0600 | 32720.5600 | 0.0000 | 21361.9100 | 0.0000 | 48.2040 |

1 through 4 for all three model types (e.g., the one proper and three improper specifications), the variance of the SEM test statistics consistently overestimated the corresponding variance of the expected underlying central and noncentral chi-square population distribution at smaller sample sizes. Second, for Specification 5 (the uncorrelated variables baseline model), the expected population variance was significantly overestimated for every model type at every sample size under study. The smallest bias found was 32%, but for most conditions bias ranged between 50% and 150%. Consistent with statistical theory, in terms of dispersion the SEM test statistics are not following the expected noncentral chi-square distributions at smaller sample sizes for the properly specified or moderately misspecified conditions, nor at any sample size for the severely misspecified uncorrelated variables model.

### *Potential Implications of Findings*

Our simulation results demonstrate that the likelihood ratio test statistic *T* does follow the expected noncentral chi-square distribution under some experimental conditions, but does not follow this distribution under others. As we discussed in the introduction, the failure of the test statistic to follow the expected underlying distribution may threaten the validity of a variety of measures of fit and methods of power estimation that rely upon the sample test statistic *T*. A comprehensive examination of the various ways in which these measures and methods may be adversely affected is beyond the scope of the current article, and we are examining these issues in greater detail elsewhere. However, we will briefly examine the implications of our simulation results for a single measure that relies directly on the condition that the test statistic follows a noncentral chi-square distribution, namely the computation of confidence intervals for the RMSEA.

*RMSEA Confidence Intervals*. The RMSEA was originally proposed by Steiger and Lind (1980) and was further developed by Browne and Cudeck (1993). The point estimate of the RMSEA is given as

(5) $$RMSEA = \sqrt{\frac{T - df}{df\,(N-1)}}$$

where *T* and *df* represent the likelihood ratio test statistic and degrees-of-freedom from the hypothesized model, respectively, and *N* represents sample size. The numerator thus represents the sample estimate of the noncentrality parameter λ and it is an estimate of the degree of model misspecification. A unique characteristic of the RMSEA is that, under the assumptions that lead the test statistic to follow a noncentral chi-square distribution, the sampling distribution is known. This allows for the computation of confidence intervals (CIs) around the point estimate. The CIs for the RMSEA are computed using appropriate upper and lower percentile limits of a noncentral chi-square distribution for given degrees-of-freedom and noncentrality parameter $\lambda$ (see Equation 14, Browne & Cudeck, 1993). The CIs are asymmetric around the point estimate and range from zero to positive infinity.

A key condition for the computation of these CIs is that the test statistic *T* from the tested model follows a noncentral chi-square distribution. Our simulation results suggest that *T* does indeed follow a noncentral chi-square distribution under some experimental conditions, but not under others. To briefly examine the potential impact of these findings on the computation of the CIs for the RMSEA, we compared the percent of sample CIs from the SEM simulations that covered the known population RMSEA value for two conditions under which we found *T* to follow a noncentral chi-square distribution and for two conditions under which it did not. Under conditions in which the test statistic follows the population noncentral chi-square distribution, we expect that approximately 90% of the 90% CIs would cover the known population value of the RMSEA.

Recall that the simulation results indicated that the likelihood ratio test statistic *T* closely followed the noncentral chi-square distribution in terms of both central tendency and dispersion for the moderately misspecified Model 3 (the full SEM), Specification 2 (omitting three cross loadings) at $N = 400$ and $N = 800$. For Specification 2 of Model 3, 91% and 90% of the sample 90% CIs covered the population RMSEA value for $N = 400$ and $N = 800$, respectively. Thus, under conditions in which *T* does follow the expected underlying distribution, the 90% CIs appear to be covering the population RMSEA at the expected rate. In contrast, recall that the simulated test statistics failed to follow a noncentral chi-square distribution for the same Model 3, Specification 2 at $N = 50$ and $N = 75$ in terms of both central tendency and dispersion. Here, 79% and 86% of the sample 90% CIs covered the population RMSEA value for $N = 50$ and $N = 75$, respectively. Thus, under experimental conditions in which *T* does not follow the expected noncentral chi-square distribution, the computed CIs based on this underlying distribution are not covering the population parameter at the expected rate.

*Summary*. This brief exploration of the RMSEA CIs suggests that the departure of *T* from the expected noncentral chi-square distribution may indeed exert a negative influence on other measures of fit based on *T*, at least under the conditions studied here. The RMSEA confidence intervals are only one of many measures of fit and methods of power estimation that might be influenced by the departures of $T$ from the expected underlying distribution. Future research is necessary to fully understand the implications of these findings across a much broader range of outcomes.

## *Discussion*

A central goal of our article was to empirically evaluate the degree to which the SEM likelihood ratio test statistic $T$ follows a central chi-square distribution under proper model specification, and a noncentral chi-square distribution under improper model specification. This is a critically important issue to better understand given the reliance on the test statistic following this known distribution across many areas of SEM applications and research, particularly in terms of fit indices and statistical power. Drawing on statistical theory and prior research, we proposed three research hypotheses that we empirically evaluated using data generated from Monte Carlo simulations. Experimental conditions included 15 different models varying both in complexity and in degree of misspecification, as well as a range of sample sizes falling between 50 and 1000. Though we feel that we exercised considerable care in the selection of our experimental conditions, we of course need to keep in mind the usual limitations that must accompany any Monte Carlo simulation design. That is, we cannot be certain about the degree to which we can extrapolate from our conditions to other modeling conditions; however, given the close correspondence of our findings to what was predicted from statistical theory, we feel confident that these findings do generalize across many research settings commonly encountered in practice. Keeping these caveats in mind, we find several interesting results.

### *Proposed Research Hypotheses*

*Hypothesis 1*. Our first hypothesis predicted that for properly specified models, the SEM test statistic $T$ would follow a central chi-square distribution with mean $df$ and variance $2df$, but only at moderate to large sample sizes. Consistent with both statistical theory (e.g., Browne, 1984) and prior research findings (e.g., Curran et al., 1996; Hu et al., 1992), our results supported this hypothesis. In terms of bias in the central tendency of the distribution, the mean of the sample estimates of $T$ consistently overestimated the mean of the expected population distribution at the smaller sample sizes. This bias in the mean became negligible at sample sizes of $N = 200$ and higher. A similar pattern of bias was found in terms of the dispersion of the distribution such that the variance of the sample estimates of $T$ significantly overestimated the variance of the expected population distribution at the smaller sample sizes for all three properly specified models. Although these results replicate several previous studies of similar research questions, the vast majority of these studies only examined the distribution of $T$ in terms of central tendency. We extend these findings by demonstrating important departures in the distribution of $T$ in terms of variance as well. This finding has important potential implications for the computation of confidence intervals, a point that we will discuss further below.

*Hypothesis 2*. Our second hypothesis was that under small to moderate model misspecification, the test statistic $T$ would follow a *noncentral* chi-square distribution with noncentrality parameter $\lambda$, mean $df + \lambda$ and variance $2df + 4\lambda$, but this was expected to only hold at moderate to large samples. Again consistent with both statistical theory (e.g., Steiger et al., 1985) and prior research (e.g., Curran et al., 1996; Fan et al., 1999), our results clearly demonstrated that for models that were misspecified, but not severely so, the distribution of $T$ did indeed follow the expected underlying noncentral chi-square distribution. However, this only held for moderate to large samples, and $T$ did not follow a noncentral chi-square distribution at smaller sample sizes. As was found for the properly specified conditions, both the mean and variance of the empirical distribution of $T$ showed significant bias relative to the expected population distribution, but the magnitude of bias was larger for the variance compared to the mean of $T$.

*Hypothesis 3*. Our third and final hypothesis was that under severe model misspecification, especially the uncorrelated variable model, the test statistic $T$ would follow neither the central nor the noncentral chi-square distribution, and we expected this to occur across all sample sizes. Consistent with both statistical theory (e.g., Steiger et al., 1985) and some limited prior research (e.g., Rigdon, 1998), our findings indicated that the empirical distribution of the test statistic $T$ did not follow the expected noncentral chi-square distribution for any model at any sample size. However, there was an intriguing aspect about this pattern of bias. When comparing the *mean* of the empirical distribution of $T$ to the expected population counterpart, there was no evidence of bias found based on the parametric, nonparametric, or effect size estimates. Again, prior simulation studies have typically only considered departures of $T$ from the expected distribution in terms of central tendency. When comparing the *variance* of the empirical distribution of $T$ to the expected population counterpart, there was significant bias evident across all models and all sample sizes. Indeed, relative bias in the variance of $T$ ranged from 32% to 164% with a median bias of 69% across all sample sizes and model types. Thus, if bias were only considered in terms of central tendency, it would be concluded that $T$ did not depart from the expected underlying distribution. However, when measuring bias in terms of dispersion, there were pervasive and significant departures of *T* from the expected underlying distribution across all modeling conditions.

*Implications*

We briefly explored the potential implications of these findings on a single measure of fit and demonstrated that the failure of *T* to follow the expected underlying distribution may indeed have negative consequences for other applications in SEM. Other areas in SEM that might be adversely affected include the computation of many fit indices and corresponding confidence intervals, as well as several power estimation procedures. For example, many relative fit indices incorporate the test statistic from the uncorrelated variables baseline model in the computation of the index, and are based on the condition that *T* follows a noncentral chi-square distribution (e.g., RNI, CFI).[6] However, our results provide strong evidence that this condition is not valid under any condition studied here. The variance of the *T* statistic for the null baseline model departs from that of the corresponding expected noncentral chi-square distribution across every sample size considered, although the mean of *T* showed only negligible bias. From one perspective, the lack of strong bias in the mean suggests that the point estimates of these relative fit indices might not greatly suffer as long as the sample is not small. On the other hand, the relative fit indices are nonlinear functions of the test statistics for the uncorrelated baseline and the hypothesized structure and this nonlinear structure complicates the assessment of the potential impact of the bias. Further, one goal has been to work toward the development of confidence intervals around the point estimates of these relative fit indices (e.g., Bentler, 1990). So, although the lack of bias in the mean of *T* may allow for appropriate point estimation, the substantial bias in the variance of *T* may have a much greater impact on the ability to compute corresponding confidence intervals. Further research is needed to more fully address these implications.

In contrast to the relative fit indices, the computation of the RMSEA does *not* involve a baseline chi-square test statistic. Since the RMSEA only assumes a noncentral chi-square distribution of *T* for the hypothesized model, our results imply that the greatest possible bias will occur in the smaller sample sizes where the test statistic does not appear to follow the noncentral chi-square distribution. An important finding here is that at sample sizes of

[6] An exception is the IFI where Bollen (1989, p. 305) suggests that the test statistic for the baseline model will not always follow a noncentral chi-square distribution and thus uses the baseline chi-square rather than the noncentrality estimate in the IFI calculation.

around $N$ = 100 and above, the mean of the $T$ statistic was not biased even under the most severe misspecification of a model that might be considered theoretically tenable in practice. However, a significant advantage of the RMSEA is the direct calculation of corresponding confidence intervals (e.g., Browne & Cudeck, 1993; Steiger & Lind, 1980). Our results indicate that the variance of $T$ closely corresponds to that of the underlying noncentral chi-square distribution, but only at sample sizes above around $N$ = 200. Thus, we might expect that both the point estimates and the confidence intervals of the RMSEA are well validated for use in practical research applications given moderate to large sample sizes, but may be biased at smaller sample sizes. Our ongoing work is directly exploring these very issues with regard to point estimation, Type I error, and power of the RMSEA.

Finally, our results indicate that power estimation procedures that depend on the condition that $T$ follows a noncentral chi-square distribution require special care when $N$ is small. It was beyond the scope of this article to examine the power of the likelihood ratio test statistic to reject an incorrect model, but our findings imply that the accuracy of methods such as those proposed by Satorra and Saris (1985), MacCallum et al. (1996), and Muthén and Curran (1997) may degrade as a function of decreasing sample size and increasing model misspecification. Further empirical work is needed to better understand the conditions under which these power estimates may become inaccurate.

It is important to stress that, although there is strong evidence that there are key experimental conditions under which $T$ does not follow a noncentral chi-square distribution, given the scope of this article we have not explicitly considered the robustness of $T$ not following this underlying distribution on the baseline fit indices, stand-alone measures such as the RMSEA, or power estimation procedures. It is possible that even though the test statistic significantly departs from the noncentral chi-square, it may very well be a good enough approximation for practical utilization. Our study provides important insights into this potential problem in terms of the distribution of $T$, but caution dictates that additional research is needed that focuses explicitly on each of these particular applications that utilize $T$.

### *Limitations and Future Directions*

As we raised earlier, an inherent limitation to any computer simulation study is that it is possible that the resultant findings can not be generalized beyond the experimental conditions under study. We endorse this as a potential limitation, but we also took great care in designing our experimental conditions to reflect a wide variety of sample sizes and model types commonly encountered in applied behavioral research. Given this, findings may differ with variations in factors such as model complexity, model parameterization, and degree of misspecification, and future research will do well to further explore these issues. A related limitation of our study is that we examined only data generated from a multivariate normal distribution. Prior research has indicated that it is important to also consider non-normally distributed data (e.g., Muthén & Kaplan, 1985, 1992), but an examination of this was beyond the scope of the current manuscript. Given that non-normal distributions are a significant problem in social science research (e.g., Micceri, 1989), much can be learned about the distribution of *T* under combinations of sample size, model specification, and multivariate distribution. These limitations should warrant some caution in over generalizing from our results, but we feel our findings provide an important first glimpse into the empirical distribution of *T* and may serve as a starting point for future research in this important area of structural equation modeling.

## References


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.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31*, 419-456.

Bentler, P. M. (1983). Some contributions to efficient statistics for 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 program manual, Version 5.0* [Computer software]. Los Angeles: BMDP Statistical Software.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wolds (Eds.), *Systems under indirect observation: Causality, structure, prediction (Part 1)*. Amsterdam: North-Holland.

Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 127-141.

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

Chen, F., Bollen, K., Paxton, P., Curran, P. J., & Kirby, J. (2001). Improper solutions in structural equation models: Causes consequences, and strategies. *Sociological Methods and Research, 29,* 468-508.

Cudeck, R. & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 141-167.

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

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.

Freund, J. E. (1992). *Mathematical statistics, fifth edition*. Englewood Cliffs, NJ: Prentice Hall.

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

Jöreskog, K. G. (1971a). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Jöreskog, K. G. (1971b). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G. & Sörbom, D. (1978). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Kanji, G. K. (1993). *100 statistical tests*. Thousand Oaks, CA: Sage.

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.

Maiti, S. S. & Mukherjee, B. N. (1990). A note on distributional properties of the Joreskog-Sorbom fit indices. *Psychometrika, 55*, 721-726.

Matsueda, R. L. & Bielby, W. T. (1986). Statistical power in covariance structure models. In N. Tuma (Ed.), *Sociological methodology 1986* (pp. 120-158). Washington, DC: American Sociological Association

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

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

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

Muthén, B. O. & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*, 371–402.

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

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

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*, 287-312.

Rigdon, E. E. (1998). The equal correlation baseline model for comparative fit assessment in structural equation modeling. *Structural Equation Modeling, 5*, 63-77.

SAS (1999). *SAS/STAT user's guide: Version 7*. Cary, NC: Author.

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

Satorra, A. & Saris, W. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 51*, 83-90.

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

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

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50*, 253-263.

*Accepted December, 2000.*