Chi-square Difference Tests for Comparing Nested Models:

An Evaluation with Non-normal Data

Goran Pavlov[1,2]

Dexin Shi[1]

Alberto Maydeu-Olivares[1,2]

[1] University of South Carolina

[2] University of Barcelona

Authors note: Correspondence concerning this article should be addressed to Goran Pavlov: Department of Psychology, University of South Carolina. Barnwell College. 1512 Pendleton St. Columbia, SC 29208; E-mail: pavlovg@mailbox.sc.edu. This research was supported by the National Science Foundation under Grant No. SES-1659936.

Abstract

The relative fit of two nested models can be evaluated using a chi-square difference statistic. We evaluate the performance of five robust chi-square difference statistics in the context of confirmatory factor analysis with non-normal continuous outcomes. The mean and variance corrected difference statistics performed adequately across all conditions investigated. In contrast, the mean corrected difference statistics required larger samples for the *p*-values to be accurate. Sample size requirements for the mean corrected difference statistics increase as the degrees of freedom for difference testing increase. We recommend that the mean and variance corrected difference testing be used whenever possible. When performing mean corrected difference testing, we recommend that the expected information matrix is used (i.e., choice MLM), as the use of the observed information matrix (i.e., choice MLR) requires larger samples for *p*-values to be accurate. Supplementary materials for applied researchers to implement difference testing in their own research are provided.

Keywords: structural equation modeling, nested models, chi-square difference test, non-normal data.

Chi-square Difference Tests for Comparing Nested Models: An Evaluation with Non-normal Data

Structural equation modeling (SEM) is a general statistical framework appropriate for modeling multivariate datasets. Over the past few decades, SEM has been steadily gaining in popularity among applied researchers across a broad range of scientific disciplines. One of the essential and frequently used features available within the SEM framework is the statistical evaluation of how well hypothesized models fit the observed data.

Maximum likelihood (ML) is the most widely used estimation method for modeling continuous data within the SEM framework (Maydeu-Olivares, 2017). When the model is correctly specified and data follow a multivariate normal distribution, the minimum of the ML fit function can be used to construct a chi-square distributed test statistic, thus enabling a statistical evaluation of the fit of the model to the data at hand. The assumption of multivariate normality, however, need not be tenable in empirical research (Cain, Zhang, & Yuan, 2017; Micceri, 1989). If data are not normal, relying on the normal-theory ML statistic to evaluate model fit may result in erroneous statistical conclusions (Hu, Bentler, & Kano, 1992; Satorra, 1990; Satorra & Bentler, 1994). To address this problem, various corrections to the chi-square test statistic have been proposed. Specifically, the chi-square statistic can be corrected so that in large samples it agrees in mean with a chi-square distribution (Asparouhov & Muthén, 2005; Satorra & Bentler, 1994; Yuan & Bentler, 2000), or it can be corrected so that it agrees in both mean and variance (Asparouhov & Muthén, 2010; Satorra & Bentler, 1994). In large samples, the mean and variance corrected chi-square statistics should be superior over the less computationally expensive mean corrected chi-squares (Asparouhov & Muthén, 2013). Maydeu-Olivares (2017) summarizes the various mean and mean and variance corrected chi-square statistics proposed in

the literature. In a simulation study, he also shows that mean and variance corrections provide more accurate *p*-values than mean corrections when assessing the absolute (i.e., model-data) fit of the model. In this article, we refer to these corrected chi-square statistics as robust chi-square statistics.

Chi-square tests can also be used to compare the fit of two models that are nested. There are many applications in which this is of interest (e.g., Elkins et al., 2018; Lai et al., 2015; Pappu & Quester, 2016; Schivinski & Dabrowski, 2016; Wingate & Bourdage, 2019). In particular, testing for differences in fit is routinely performed in the measurement invariance literature (e.g., Guhn, Ark, Emerson, Schonert-Reichl, & Gadermann, 2018; Hawes et al., 2018; Huhtala, Kangas, Kaptein, & Feldt, 2018; Jenkins, Fredrick, & Nickerson, 2018; Krieg, Xu, & Cicero, 2018).

Consider two models, Model 0 and Model 1, with degrees of freedom df_0 and df_1, respectively, where $df_0 > df_1$. Model 0 is nested within Model 1 if the mean and covariance structures implied by Model 0 can be reproduced exactly by fitting Model 1 (Bentler & Satorra, 2010). Using ML, and if the normality assumption holds, the difference in model fit can be conveniently tested by computing the difference between chi-square statistics of the two nested models under consideration. When the larger model (Model 1) is correctly specified, the difference statistic asymptotically follows a chi-square distribution. If the chi-square difference statistic cannot be rejected, the more parsimonious model (Model 0), should be preferred over the less restricted one (Model 1).

If the normality assumption does not hold, the difference between the two robust fit statistics will not be chi-square distributed, thus compromising the accuracy of statistical conclusions (Satorra, 2000). To facilitate appropriate statistical testing for differences in fit under

non-normality, several corrections to the chi-square difference statistic have been proposed (e.g., Asparouhov & Muthén, 2006; Asparouhov & Muthén, 2010; Satorra, 2000; Satorra & Bentler, 2001; Satorra & Bentler, 2010). To date, the two most commonly utilized options among applied researchers have been the two versions of Satorra-Bentler mean-adjusted chi-square difference statistic (Satorra & Bentler, 2001, and Satorra & Bentler, 2010). Surprisingly, notwithstanding the frequent and ongoing application of these two corrected statistics, only two studies have thoroughly assessed their performance under non-normality: Chuang, Savalei, and Falk (2015), and Brace and Savalei (2017). The results of both studies reinforced concerns regarding the application of uncorrected difference statistics to non-normal data and provided evidence of the robustness to non-normality of the Satorra and Bentler (2001) and (2010) corrections under a variety of plausible research scenarios, gently favoring the more recent one.

However, these recent studies did not include an investigation of the mean and variance adjusted difference statistics (Asparouhov & Muthén, 2006, 2010), which may perform better than the mean corrected difference statistics currently in use in applications. Accordingly, the current investigation is aimed at addressing this gap in the literature to date. The remaining of this paper is organized as follows. First, we describe the mean, and mean and variance corrections to chi-square statistics for comparing nested models. Next, we summarize previous studies on the behavior of mean corrected difference statistics when data is non-normal and emphasize the rationale for investigating the performance of the mean and variance corrected test statistic. Afterwards, we present the results of a simulation study comparing the performance of Asparouhov and Muthén's (2006, 2010) mean and variance adjusted difference chi-square to Satorra and Bentler's (i.e., 2001, 2010) mean adjusted difference statistics with respect to both empirical Type I error rates and power. Finally, we discuss the results and provide some

recommendations for substantive researchers. In the supplementary materials to this article, we provide a worked out example in order to facilitate the application of the discussed methods.

Mean, and Mean and Variance Corrections to Chi-square Statistics for Comparing Nested Models

In this article we focus on structural equation models for continuous outcomes estimated by ML as this is the most commonly used setup in applications. Under a multivariate normality assumption, and when no constraints are imposed on the means, the ML fit function is given by:

$$F_{ML}(\mathbf{S},\boldsymbol{\Sigma}(\boldsymbol{\theta})) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \log|\mathbf{S}| + tr(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})) - p, \quad (1)$$

where $\mathbf{S}$ is the sample covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the model implied covariance matrix, $\boldsymbol{\theta}$ is the vector of model parameters with length q, and p is the number of observed variables. Within this setup, the most widely used test statistic used to assess fit of the hypothesized model is the likelihood ratio test statistic,

$$T = (N-1)\hat{F}_{ML}, \quad (2)$$

where N denotes sample size, and $\hat{F}_{ML}$ is obtained by minimizing the ML fit function with respect to $\boldsymbol{\theta}$. If the multivariate normality assumption holds and the model is correctly specified, T asymptotically follows a chi-square distribution with degrees of freedom (*df*) equal to $p(p+1)/2 - q$, hence allowing for statistical evaluation of model fit. In the applied literature, T is commonly referred to as the chi-square test statistic. However, if data are not normally distributed, T will not be χ^2 distributed. In this case, the chi-square statistic can be adjusted so that it matches asymptotically a χ^2 distribution either in its mean (e.g., Satorra & Bentler, 1994; Yuan & Bentler, 2000; Asparouhov & Muthén, 2005), or in its mean and its variance (Satorra & Bentler, 1994; Asparouhov & Muthén, 2010). The mean adjusted test statistics can be written as

$\bar{T} = \frac{T}{c}$, where c is the scaling correction; the mean and variance adjusted statistic can be written as $\bar{\bar{T}} = aT + b$ (Asparouhov & Muthén, 2010). Two variants of the mean adjusted statistic have been proposed. They differ on how c is computed and on their suitability in the presence of missing data. The first one was originally proposed by Satorra and Bentler (1994) and can only be used with complete data. In widely used Mplus (Muthén & Muthén, 2017) and lavaan (Rosseel, 2012) SEM packages, it is obtained when choice MLM is selected. The second one was originally proposed by Yuan and Bentler (2000) and later modified by Asparouhov and Muthén (2005). It is suitable for both complete and incomplete data, and obtained in Mplus and lavaan using choice MLR. The main difference between the MLR and MLM version of the test is how the information matrix used in computing the scaling correction is estimated. In MLM, the expected information matrix is used, whereas in MLR, the observed information matrix is used. The latter should provide more accurate results (Efron & Hinkley, 1978; Maydeu-Olivares, 2017; Savalei, 2010). A detailed technical account of the differences between choices MLM and MLR can be found in Maydeu-Olivares (2017).

In applications, it is often of interest to compare the fit of competing models. When the comparison between two models involves one model nested within another, a test can be performed to determine whether the difference in fit is statistically significant. We use M_0 and df_0 to denote the more restricted model to be compared and its degrees of freedom. We denote by M_1 with df_1 the less restricted model. M_0 will be nested within M_1, for instance, if M_0 is the result of placing constraints on some of the model parameters of M_1. Under normality assumptions, and for ML estimation, the difference in fit between two nested models can be tested simply by subtracting the two chi-square fit statistics:

$$D = T_0 - T_1, \tag{3}$$

where T_0 and T_1 are chi-square statistics for models M_0 and M_1, respectively. Under these conditions, and when both models are correctly specified, D asymptotically follows a chi-square distribution with degrees of freedom $df = df_0 - df_1$ (Steiger, Shapiro, & Browne, 1985).

When data are not normal, D does not result in a χ^2 distributed statistic (Satorra, 2000). To account for that, Satorra (2000) developed a scale corrected χ^2 difference test robust to non-normality. However, his computationally taxing implementation was quickly followed by an alternative correction to D that can be conveniently computed from a standard SEM software output (Satorra & Bentler, 2001). The scale corrected difference test statistic is given by

$$\bar{D}_{01} = \frac{D}{c_{01}}, \qquad c_{01} = \frac{df_0 c_0 - df_1 c_1}{df_0 - df_1}, \tag{4}$$

where c_0 and c_1 are the scaling corrections for testing the absolute fit of M_0 and M_1, respectively. We note that if T_0, $\bar{T}_0$ and T_1, $\bar{T}_1$ denote the uncorrected and mean-corrected chi-square statistics for the two models, respectively, then $c_0 = \frac{T_0}{\bar{T}_0}$ and $c_1 = \frac{T_1}{\bar{T}_1}$. We refer to this robust difference statistic as DSB1, and consider two variants of it. The first one employs the Satorra-Bentler mean-adjusted χ^2 (Satorra & Bentler, 1994) to obtain $\bar{T}_0$ and $\bar{T}_1$. Following Mplus/lavaan terminology, we refer to this option in the current study with DSB1_{MLM}. The second option considered uses Asparouhov and Muthén's (2005) mean-adjusted correction to obtain $\bar{T}_0$ and $\bar{T}_1$. We refer to this option here with DSB1_{MLR}. We note that what we refer to in this paper as DSB1_{MLM} corresponds to the difference statistic D_{R1} evaluated by Chuang and colleagues (2015), and to the D_{SB1} statistic evaluated by Brace and Savalei (2017).

A drawback of the DSB1 statistic proposed by Satorra and Bentler (2001) is that when sample size is small, the correction in (4) can take a negative value leading to a negative estimate of the test statistic. To avoid this shortcoming of the scaling correction in (4), Satorra and Bentler (2010) proposed another version of mean-adjusted scaling correction that can take only positive values. The "strictly positive" Satorra-Bentler corrected difference test statistic is identical to (4) except that c_1 in (4) is replaced by $c^* = \frac{T^*}{\overline{T}^*}$, where T^*, $\overline{T}^*$ are uncorrected and robust chi-square statistics associated with an additional model run (M^*) of the less restricted model M_1 using the parameter estimates of the more restricted model M_0 as starting values and with the number of iterations set to 0 (Bryant & Satorra, 2012). We refer to this robust difference statistic here as DSB10. The DSB10 statistic is asymptotically equivalent to DSB1, yet it is guaranteed to be positive (Satorra & Bentler, 2010). As with DSB1, we consider two options of DSB10. The first one employs the Satorra-Bentler χ^2 (Satorra & Bentler, 1994) to obtain $\overline{T}_0$ and $\overline{T}^*$. We refer to this option here as $\text{DSB10}_{\text{MLM}}$. The second option employs Asparouhov and Muthén's (2005) mean-adjusted correction to obtain $\overline{T}_0$ and $\overline{T}^*$, and we refer to this option here as $\text{DSB10}_{\text{MLR}}$. We note that what we refer to in this paper as $\text{DSB10}_{\text{MLM}}$ corresponds to the difference statistic D_{R2} evaluated by Chuang and colleagues (2015), and to the D_{SB10} statistic evaluated by Brace and Savalei (2017).

Of focal interest in the current study is, however, the second order (i.e., the mean and variance) adjusted difference statistics developed by Asparouhov and Muthén (2010), currently implemented in Mplus under the "MLMV" estimator using the "DIFFTEST" command. In contrast to the mean corrections, the second order adjustment takes the form $\overline{\overline{D}} = aD + b$, where a is the scaling correction and b is the shift parameter. In order to match the empirical mean and

variance of the difference statistic with those of a chi-square distribution, *a* and *b* need to meet $E(\bar{D}) = df$ and $Var(\bar{D}) = 2df$. The second order adjustment (Asparouhov & Muthén, 2010) is given by

$$\bar{\bar{D}} = \sqrt{\frac{df}{tr(\mathbf{M}^2)}} D + df - \sqrt{\frac{df\, tr(\mathbf{M})^2}{tr(\mathbf{M}^2)}}, \quad (5)$$

where $\mathbf{M}$ is given in formula (9) in Asparouhov and Muthén (2006). We refer to the difference statistic in (5) as D_{MLMV}. In Table 1, we summarize the choices of statistics available to substantive researchers to test differences in fit between nested models.

Insert Table 1 about here

Previous research and research hypotheses

Chuang and colleagues (2015) compared the Type I error rates between the two Satorra and Bentler's (Satorra & Bentler, 2001, 2010) mean corrected difference statistics, i.e., $DSB1_{MLM}$ and $DSB10_{MLM}$ (e.g., the expected information matrix was used in computing this statistic), also including the uncorrected statistic (D) suitable for normal data. Within a confirmatory factor analysis (CFA) framework, the types of constraints studied included constraining factor correlations to 0 or to 1, and constraining loadings to be equal. Both normal and non-normal data were considered. Two methods to generate non-normal data were used: the method proposed by Vale and Maurelli (1983), and a mixture of normal distributions (i.e., a contaminated multivariate normal distribution). In the first case, skewness was set to 2 and kurtosis to either 7 or 15; in the second case, skewness was set to 0 and kurtosis to 4.96. Models between $p = 8$ and 12 observed variables were considered, and the degrees of freedom available for difference testing ranged from 1 to 5. Sample sizes (N) ranged from 100 to 1,000

observations. The uncorrected statistic (D) performed well across conditions involving normally distributed data but was consistently overrejecting the true null when data were non-normal. Across the conditions involving non-normality, both mean corrected difference statistics outperformed the uncorrected test and overall performed reasonably well, with a slight tendency of $DSB1_{MLM}$ to underreject and $DSB10_{MLM}$ to overreject.

In a follow-up to the study by Chuang and colleagues (2015), Brace and Savalei (2017) investigated both Type I errors and power of the two Satorra and Bentler's mean corrected statistics in the context of evaluating measurement invariance in two-group CFA models. As in the previous study (Chuang et al., 2015), D, $DSB1_{MLM}$ and $DSB10_{MLM}$ were investigated using the same data generating procedures and skewness/kurtosis values. Total sample sizes (N) ranged from 220 to 1,760 observations, model size was either $p = 8$ or 16, and the degrees of freedom available for difference testing ranged from 6 to 16. Type I error results revealed that the mean corrected statistics overrejected the null hypothesis of overall model fit in the presence of non-normality in small samples. The overrejection was increasing with the increasing levels of non-normality and model size. Accurate Type I errors were obtained in most conditions in which the smallest sample size (recall that this is a two-group set up) was $N = 440$. In general, the mean corrected difference statistics behaved better than the statistics for overall model fit. As Brace and Savalei (2017, p. 477) put it, "rejection rates of scaled difference tests are related to the differences in the rejection rates of the corresponding scaled tests of overall model fit". Type I errors for $DSB10_{MLM}$ were accurate except for a few conditions involving the smallest sample sizes ($N = 220$). The behavior of $DSB1_{MLM}$ was noticeably worse in small samples.

We extend previous research by evaluating the performance of the mean and variance difference correction. One would expect that the mean and variance corrected test statistics

would perform better in large models than statistics that involve only a mean correction. In particular, Maydeu-Olivares (2017) showed that when $p = 16$, both types of robust statistics yielded adequate empirical Type I errors when assessing the overall model fit. However, when $p = 32$, the mean and variance corrected test statistic maintained nominal Type I error rates while the mean corrected statistics were overrejecting the model. The magnitude of overrejection was increasing as the sample size was decreasing. Accordingly, we expect similar behavior of the robust difference statistics, that is, more accurate Type I error rates in small samples and for large models when MLMV is used.

In addition, the current study goes beyond previous research by also evaluating the performance of the two Satorra-Bentler difference corrections coupled with the Asparouhov and Muthén's (2005) mean adjustment for absolute fit (i.e., $DSB1_{MLR}$ and $DSB10_{MLR}$). These combinations are of particular interest to substantive researchers because MLR is the only option currently available for modeling incomplete data. Previous research (Maydeu-Olivares, 2017) reports that when assessing the overall model fit, choices MLR and MLM provide similar results, except in smaller samples ($N \leq 500$) where MLM slightly outperforms MLR. Accordingly, we expect similar behavior of the difference statistics, namely, more accurate Type I error rates in small samples ($N \leq 500$) when MLM is used.

Simulation Study

A simulation study was conducted to assess the performance of five robust difference options: $DSB1_{MLM}$, $DSB1_{MLR}$, $DSB10_{MLM}$, $DSB10_{MLR}$, and D_{MLMV}. The uncorrected difference test, D, was also included in the study to serve as a baseline for comparison. The data were generated in the context of a two-wave longitudinal one factor model. Put differently, the population model is a one factor model measured at two time points. As a result, it has the form

of a two factor confirmatory factor analysis (CFA) model with correlated errors to account for dependencies across time. We display in Figure 1 one of the models used in our simulation.

The chi-square difference tests were conducted to examine the equivalence of factor loadings across the two occasions. It is important to note that such tests are routinely utilized, for example, when researchers test weak factorial invariance across time (Meredith, 1993; Shi, Song & Lewis, 2017). When generating data, both factor variances were set to one and the population value of the inter-factor correlation was set to 0.30. We set the population values of all factor loadings to 0.70, except for the factor loading value for the first indicator of the second factor. The value of this factor loading was varied as described below. The population values for residual correlations across the two time occasions was set to 0.15. Finally, the error variances were set such that the population variances of the observed variables were equal to one.

Insert Figure 1 about here

Study conditions

The simulation conditions were obtained by manipulating the following five factors: (a) level of non-normality, (b) sample size, (c) model size, (d) magnitude of (non)invariance, and (e) degrees of freedom of the difference test.

Level of non-normality. We used three levels of non-normality by manipulating the magnitude of skewness and (excess) kurtosis: Normal data (0,0), moderately non-normal (2,7), and severely non-normal (2,10). We chose these particular values of skewness and kurtosis to match the values used in studies by Chuang and colleagues (2015) and Brace and Savalei (2017). Until recently, the standard method for generating non-normal data was based on Vale and Maurelli (1983). However, Foldnes and Olsson (2016) have recently shown that the Vale-

Maurelli method gives an overly optimistic evaluations of the performance of estimators and fit statistics. Accordingly, in this paper non-normal data were generated using the procedure described by Foldnes and Olsson (2016).

Sample size. Four typical sample size variants were included in the study: extremely small (100), small (200), moderate (500) and large (1,000) sample size.

Model size. Model size refers to the total number of observed variables (*p;* Shi, Lee, & Terry, 2015, 2018). Two model sizes were considered: small model with five indicators per factor ($p = 10$), and large model with fifteen indicators per factor ($p = 30$). We chose $p = 30$ because Maydeu-Olivares (2017) showed that the behavior of mean corrected test statistics for assessing model-data fit deteriorate in models of this (and larger) model size.

Magnitude of noninvariance. Three levels of noninvariance were considered by manipulating the population values of the first indicator across factors: invariant, small, and large noninvariance. For the invariant conditions, all factor loadings were equivalent across two occasions (i.e., $\lambda = 0.70$). Therefore, rejecting the chi-square difference test implies that a Type I error is made. The condition with small noninvariance corresponds to setting the population loadings of the first indicator to 0.70 in one factor and to 0.50 in the second factor ($\Delta\lambda = 0.20$). In the large noninvariance condition these values were $\lambda = 0.70$ and $\lambda = 0.30$ ($\Delta\lambda = 0.40$), respectively. Under both small and large noninvariant conditions, the probability of rejecting the chi-square difference test informs us of the power rates of the test.

Degrees of freedom of the difference test (df). We manipulated the degrees of freedom of the test by varying the number of equality constraints imposed (i.e., the number of tested factor loadings). The invariance tests were conducted on the first factor loading and on all factor loadings across two occasions. That is, when $p = 10$ (i.e., five factor loadings loaded on each

factor), the difference tests had either *df* = 1 (small) or *df* = 5 (large); whereas when *p* = 30 (i.e., 15 factor loadings loaded on each factor), the difference tests had either *df* = 1 (small) or *df* = 15 (large).

In sum, the simulation study consisted of a fully crossed design including three distributional shapes (normal, moderately non-normal, and severely non-normal), three (non)invariance options (invariance, small noninvariance, and large noninvariance), four sample sizes (100, 200, 500, and 1,000), two model sizes (small and large), and two *df* options (small and large). One hundred and forty-four (144) conditions were created ($3 \times 3 \times 4 \times 2 \times 2$) in total. One thousand replications were generated for each condition using the function *nnig_sim* in the *miceadds* package in R (R Core Team, 2019; Robitzsch, 2019).

Estimation

The chi-square difference tests were conducted by comparing two nested models. The less restricted (baseline) model M_1 was a two-wave longitudinal CFA model with all parameters freely estimated (the factor variances were fixed to one for model identification purposes). The more restricted models M_0 had either one (the first one) or all factor loadings constrained to be equal across occasions. For each dataset, we fitted the nested models and conducted chi-square difference tests using ML and the robust ML (i.e., MLM, MLR and MLMV) estimation methods. As previously described, for both MLM and MLR, two variants of the mean corrected difference tests were computed (i.e., DSB1 and DSB10). In total, the performance of six maximum likelihood (ML) based chi-square difference tests (D, $DSB1_{MLM}$, $DSB1_{MLR}$, $DSB10_{MLM}$, $DSB10_{MLR}$, and D_{MLMV}) was compared across the simulated conditions.

In order to evaluate the performance of different robust chi-square difference tests, empirical rejection rates for nominal alpha levels of 5% were computed across all replications

within each simulation condition. To reiterate, under the invariant conditions (i.e., the null hypotheses are correct), the empirical rejection rates are Type I error rates. When the tested factor loadings are noninvariant in the population (i.e., the null hypotheses are wrong) the proportions of rejections across all replications are to be interpreted as the power of the chi-square difference test. All estimations were performed using lavaan 0.6-5 (Rosseel, 2012) except for MLMV, for which Mplus 8 (Muthén & Muthén, 2017) was used.

Results

For all of the study conditions all replications successfully converged. Accordingly, results for each condition under investigation were based on all 1,000 replications.

Type I error rates

For the Type I error rate analysis, we used results involving the invariant population model. The less restricted model M_1 and additionally restricted models M_0 were correctly specified in all conditions. In Table 2 and Table 3 we provide empirical Type I error rates of the difference tests at the 5% level of significance for small ($p = 10$) and large models ($p = 30$) respectively. Following Bradley (1978), and taking into account rounding error, we considered Type I error rates in [.02, .08] to be adequate. Conditions with Type I error rates outside this range are highlighted in Tables 2 and 3.

Under normality, all examined difference tests performed well across conditions involving M_0 with a single constraint ($df = 1$; Tables 2 and 3), regardless of model size and sample size. In conditions with small models ($p = 10$) and M_0 with multiple constraints ($df = 5$; see Table 2), the Type I error rates were also appropriate for all examined statistics. Finally, conditions involving large models ($p = 30$) and M_0 with multiple constraints ($df = 15$; Table 3) were more challenging for the studied difference statistics to maintain Type I accuracy. In these

conditions, the difference statistics involving MLR and MLMV (i.e., $DSB1_{MLR}$, $DSB10_{MLR}$, and D_{MLMV}) tended to slightly underreject.

In conditions with non-normal data, the uncorrected difference test (D) did not maintain its accuracy and, as expected, was overrejecting the true null, regardless of model size, sample size, and degrees of freedom. No large differences in rejection rates were observed across conditions involving different model sizes, severity of non-normality, sample sizes, and degrees of freedom (see Tables 2 and 3).

Conversely, in all conditions with non-normal data, the robust difference statistics were outperforming the uncorrected option. However, their behavior was differently affected by non-normality. Both versions of the Satorra-Bentler mean corrected difference statistics (Satorra & Bentler, 2001, 2010) were overrejecting the true null in several conditions with non-normal data. Conversely, the mean and variance corrected difference statistic (D_{MLMV}; Asparouhov & Muthén, 2010) was performing consistently and it was the only option that yielded adequate Type I error rates across all non-normal conditions (see Tables 2 and 3). Overall, as hypothesized, the mean and variance corrected statistic, D_{MLMV}, outperformed the two Satorra and Bentler's (2001, 2010) mean corrected difference statistics.

As it can be observed in Tables 2 and 3, with respect to Type I error rates, the main effect of Satorra-Bentler (2001) vs. (2010) option was small. A more substantial effect was found for the MLM vs. MLR option. Specifically, larger sample sizes were needed for MLR (i.e., $SB1_{MLR}$ and $SB10_{MLR}$) than for MLM options (i.e., $SB1_{MLM}$ and $SB10_{MLM}$) to reach adequate Type I error rates. The model size effect was not observed. As can be seen in Tables 2 and 3, holding all other factors constant and simply increasing the number of variables had no effect on the performance of the two mean corrected difference statistics. However, the number of degrees of freedom

available for difference testing did have an impact on the performance of the robust difference statistics. Holding all other factors constant, the larger the number of degrees of freedom, the poorer was the performance of the mean corrected statistics. Within the limited conditions of this study, the mean and variance difference statistic (D_{MLMV}) seemed robust to this effect.

Finally, a small interaction effect between the version of the difference statistic, i.e., Satorra-Bentler (2001) vs. (2010), and the choice of formula used to obtain the standard errors for the model parameters (i.e., MLM vs. MLR) was observed. As it can be seen in Tables 2 and 3, when there was a difference in Type I error rates between the two Satorra-Bentler difference corrections, a slightly more accurate results were observed for the original version when both were coupled with the MLM option (i.e., $SB1_{MLM}$), whereas a slightly more accurate results were obtained using the "strictly positive" version when both were coupled with the MLR option (i.e., $SB10_{MLR}$).

Insert tables 2 and 3 about here

Power

Power analysis was based on two population models with one noninvariant factor loading. The less restricted model M_1 was correctly specified in all conditions. Conversely, both more restricted models M_0 were misspecified, simulating a small misspecification when the difference of the constrained factor loading across occasions was $\Delta\lambda = 0.20$, and a large misspecification when the difference was $\Delta\lambda = 0.40$. The power of the difference test thus reflects the sensitivity of the test to identify this misspecification in M_0.

Power results are provided in Tables 4 and 5 for small (p = 10) and large model (p = 30) respectively. In the tables, conditions with incorrect Type I error rates identified earlier are highlighted. We evaluate only power results in conditions with adequate Type I error rates, that is, in those conditions not highlighted in the tables. As expected, power of the difference statistics was increasing with the increasing sample size and severity of misspecification and was decreasing with the increasing degrees of freedom for the difference test. Overall, we did not observe substantial differences in power among difference statistics in conditions with adequate Type I error rates (see Tables 4 and 5).

Insert tables 4 and 5 about here

Discussion

Applied researchers are often interested in assessing if a plausible and more parsimonious model fits the data as well as the initial model under consideration. If the two models of interest are nested and if data are normally distributed, evaluating the difference in model fit can be conveniently performed, because the difference in absolute fit of the two models will result in a statistic that follows a chi-square distribution. However, if data are not normal, a difference statistic obtained by subtracting the two robust absolute fit statistics will not necessarily be chi-square distributed, requiring a unique adjustment (Satorra, 2000; Satorra & Bentler, 2001). In order to facilitate appropriate selection of difference statistics in substantive research, we evaluated the performance of several difference options appropriate for non-normal continuous outcomes.

Of focal interest in the current investigation was the performance of a seldom utilized yet potentially advantageous second order adjustment, that is, the mean and variance corrected difference statistic proposed by Asparouhov and Muthén (D_{MLMV}; 2010). In order to provide a more thorough evaluation of this robust difference statistic, we pitted its behavior against the two more popular mean corrected statistics, DSB1 and DSB10, proposed by Satorra and Bentler (2001, 2010). The Satorra-Bentler difference statistics can be used in concert with the Satorra and Bentler's (1994) model-data fit statistic appropriate for complete data (MLM), or the Asparouhov and Muthén's (2005) model-data fit statistic appropriate for both complete and incomplete data (MLR). Accordingly, the options under investigation were $DSB1_{MLM}$, $DSB1_{MLR}$, $DSB10_{MLM}$, $DSB10_{MLR}$, and D_{MLMV}. We also included in the comparison the uncorrected difference statistic (D) as a baseline. We evaluated the chosen options with respect to both Type I error rate accuracy and power of the test.

As expected, our investigation reconfirms that the uncorrected difference statistic can only be used with normally distributed data. When data is non-normal, it overrejects the true null, informing the researcher than the two models are different (and therefore the more complex model should be selected), when in fact the fit of both models is comparable. In the current investigation, the two Satorra-Bentler mean corrected difference statistics (DSB1 and DSB10) tended to overreject when sample size was small ($N < 200$). Their performance worsened as sample size decreased, kurtosis increased, and the degrees of freedom available for testing increased. Conversely and as hypothesized, the mean and variance corrected difference statistic (D_{MLMV}; Asparouhov & Muthén, 2010) outperformed the mean corrected options, and also provided the adequate Type I error rates across all non-normal conditions investigated. In terms of power, and holding Type I errors constant, no substantial differences were found among the

difference statistics considered (the uncorrected, mean corrected, and mean and variance corrected). Overall, a clear winner among the difference statistics considered in the current investigation is the mean and variance corrected difference statistic.

Among the mean corrected difference statistics studied, choices with MLM outperformed choices with MLR, especially in small samples. In contrast to previous studies, we did not find the Satorra and Bentler's (2010) procedure of combining the mean corrected statistics to obtain the difference statistic advantageous over the original Satorra and Bentler's (2001) proposal. This simply means that in our simulation setup, the original procedure did not fail (recall that the "strictly positive" procedure is essentially a way to obtain the difference statistic when the original procedure yields an improper value).

Limitations and directions for future research

As in any other simulation study, our conclusions are limited by the conditions included in the current investigation. We simulated conditions involving measurement invariance over time and found that the computationally more demanding mean and variance difference test statistic outperforms statistics that only involve a mean correction. However, nested tests are also widely used to assess measurement invariance across populations (e.g., males vs. females). Therefore, future research should be aimed at replicating our findings in this setup.

Moreover, we found that the performance of the mean corrected difference statistics worsened as the number of degrees of freedom for the difference test increased. In contrast, the mean and variance statistic maintained nominal Type I error rates in all conditions investigated. Nevertheless, it is reasonable to suspect that as degrees of freedom increase, *p*-values obtained using the mean and variance corrected difference statistic would eventually break down as well.

Accordingly, it would be of interest for future research to consider large models involving larger numbers of degrees of freedom for difference testing than those used in the current study.

It seems of interest to note that the mean and variance difference statistics are also available when estimating ordinal factor analysis using polychoric correlations. In this case, Mplus implements these statistics for the unweighted and diagonally weighted least squares estimators (choices ULSMV and WLSMV in Mplus terminology; see Asparouhov & Muthén, 2010). Additional research is needed to investigate the performance of the mean and variance difference statistics in setups involving ordinal data.

In closing, we must reiterate that statistical theory for chi-square difference testing relies on the assumption that the larger model being compared is correctly specified (Haberman, 1977; Yuan & Bentler, 2004), but it may not be able to assess this assumption because of the model size effect (Moshagen, 2012). Nevertheless, *p*-values for difference testing may be accurate even when *p*-values for overall model testing are not (e.g., see Brace & Savalei, 2017; Maydeu-Olivares & Cai, 2006). Accordingly, chi-square difference testing should be performed with care (Yuan & Bentler, 2004).

Recommendations

Based on the evidence of the current evaluation, we recommend that the mean and variance difference correction be used whenever possible, both for continuous outcomes and (pending further evaluation) for ordinal outcomes as well. For continuous outcomes, the mean and variance corrected difference test proposed by Asparouhov and Muthén (2010) can be conveniently performed in Mplus by selecting as estimator MLMV in concert with the DIFFTEST option. For binary and ordinal outcomes, this option is available for estimation choices ULSMV and WLSMV. Researchers that do not have access to this software may use the

mean corrected difference tests provided their sample is large enough (i.e., $N \geq 500$). If opting for the mean corrected statistics, we recommend that statistics using the expected information matrix (MLM in Mplus terminology) are preferred over statistics using the observed information matrix (MLR in Mplus terminology), as the latter require larger samples to perform adequately. The original Satorra-Bentler mean difference correction (2001) may be preferred over the "strictly positive" option (Satorra & Bentler, 2010), unless it yields an improper value. We provide as supplementary material a worked-out example and Mplus code for all the evaluated robust difference tests so that substantive researchers can conveniently use them in their own research.

References

Asparouhov, T., & Muthén, B. (2005). Multivariate statistical modeling with survey data. In *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference,* 1-30. Retrieved from http://statmodel2.com/download/AsparouhovMuthen_MultivariateModeling3.pdf

Asparouhov, T., & Muthén, B. (2006). Robust chi square difference testing with mean and variance adjusted test statistics (Mplus Web Notes No. 10). Retrieved from http://www.statmodel.com/download/webnotes/webnote10.pdf

Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction (Technical appendix). Retrieved from https://www.statmodel.com/download/WLSMV_new_chi21.pdf

Asparouhov, T., & Muthén, B. (2013). Computing the strictly positive Satorra-Bentler chi-square test in Mplus (Mplus Web Notes No. 12). Retrieved from https://www.statmodel.com/examples/webnotes/SB5.pdf

Brace, J. C., & Savalei, V. (2017). Type I error rates and power of several versions of scaled chi-square difference tests in investigations of measurement invariance. *Psychological Methods*, *22*(3), 467-485. https://doi.org/10.1037/met0000097

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods*, *15*(2), 111–123. http://doi.org/10.1037/a0019625

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 372-398. https://doi.org/10.1080/10705511.2012.687671

Cain, M. K., Zhang, Z., & Yuan, K. -H. (2017). Univariate and multivariate skewness and kurtosis for measuring non-normality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716-1735. https://doi.org/10.3758/s13428-016-0814-1

Chuang, J., Savalei, V., & Falk, C. F. (2015). Investigation of Type I error rates of three versions of robust chi-square difference tests. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 517-530. https://doi.org/10.1080/10705511.2014.938713

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*(3), 457–483. https://doi.org/10.1093/biomet/65.3.457

Elkins, I. J., Saunders, G. R. B., Malone, S. M., Wilson, S., McGue, M., & Iacono, W. G. (2018). Mediating pathways from childhood ADHD to adolescent tobacco and marijuana problems: roles of peer impairment, internalizing, adolescent ADHD symptoms, and gender. *Journal of Child Psychology and Psychiatry*, *59*(10), 1083-1093. https://doi.org/10.1111/jcpp.12977

Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate Behavioral Research*, *51*(2–3), 207–219. https://doi.org/10.1080/00273171.2015.1133274

Guhn, M., Ark, T. K., Emerson, S. D., Schonert-Reichl, K. A., & Gadermann, A. M. (2018). The satisfaction with life scale adapted for children: Measurement invariance across gender

and over time. *Psychological Assessment, 30*(9), 1261-1266. https://doi.org/10.1037/pas0000598

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *The Annals of Statistics*, *5*(6), 1148–1169. https://doi.org/10.1214/aos/1176344001

Hawes, S. W., Byrd, A. L., Kelley, S. E., Gonzalez, R., Edens, J. F., & Pardini, D. A. (2018). Psychopathic features across development: Assessing longitudinal invariance among Caucasian and African American youths. *Journal of Research in Personality*, *73*, 180-188. https://doi.org/10.1016/j.jrp.2018.02.003

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling*, *14*(3), 361–390. https://doi.org/10.1080/10705510701301602

Hu, L. -t., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351-362. https://doi.org/10.1037/0033-2909.112.2.351

Huhtala, M., Kangas, M., Kaptein, M., & Feldt, T. (2018). The shortened Corporate Ethical Virtues scale: Measurement invariance and mean differences across two occupational groups. *Business Ethics: A European Review*, *27*(3), 238–247. https://doi.org/10.1111/beer.12184

Jenkins, L. N., Fredrick, S. S., & Nickerson, A. (2018). The assessment of bystander intervention in bullying: Examining measurement invariance across gender. *Journal of School Psychology*, *69*, 73-83. https://doi.org/10.1016/j.jsp.2018.05.008

Krieg, A., Xu, Y., & Cicero, D. C. (2018). Comparing social anxiety between Asian Americans and European Americans: An examination of measurement invariance. *Assessment*, *25*(5), 564–577. https://doi.org/10.1177/1073191116656438

Lai, C. M., Mak, K. K., Watanabe, H., Jeong, J., Kim, D., Bahar, N., … Cheng, C. (2015). The mediating role of Internet addiction in depression, social anxiety, and psychosocial well-being among adolescents in six Asian countries: A structural equation modelling approach. *Public Health*, *129*(9), 1224-1236. https://doi.org/10.1016/j.puhe.2015.07.031

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 383-394. https://doi.org/10.1080/10705511.2016.1269606

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*(1), 55–64. https://doi.org/10.1207/s15327906mbr4101_4

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika*, *58*, 525-543. https://doi.org/10.1007/BF02294825

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156. http://dx.doi.org/10.1037/0033-2909.105.1.156

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*(1), 86–98. https://doi.org/10.1080/10705511.2012.634724

Muthén, L. K., & Muthén, B. (2017). MPLUS 8 [Computer program]. Los Angeles, CA: Muthén & Muthén.

Pappu, R., & Quester, P. G. (2016). How does brand innovativeness affect brand loyalty? *European Journal of Marketing*, *50*(1-2), 2-28. https://doi.org/10.1108/EJM-01-2014-0020

R Core Team. (2019). A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www. R-project. org.

Robitzsch, A. (2019). R package miceadds: Some additional multiple imputation functions. Retrieved from http://cran.r-project.org

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity*, *24*(4), 367-386. https://doi.org/10.1007/BF00152011

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. Advanced studies in theoretical and applied econometrics, 36*. (pp. 233-247). Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-4603-0_17

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.) *Latent variable*

analysis: Applications for developmental research (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507-514. https://doi.org/10.1007/BF02296192

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243-248. https://doi.org/10.1007/s11336-009-9135-y

Schivinski, B., & Dabrowski, D. (2016). The effect of social media communication on consumer perceptions of brands. *Journal of Marketing Communications*, *22*(2), 189-214. https://doi.org/10.1080/13527266.2013.871323

Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods*, *15*(4), 352–367. https://doi.org/10.1037/a0020143

Shi, D., Lee, T., & Terry, R. A. (2015). Revisiting the model size effect in structural equation modeling (SEM). *Multivariate Behavioral Research*, *50*(1), 142. https://doi.org/10.1080/00273171.2014.989012

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 21-40. https://doi.org/10.1080/10705511.2017.1369088

Shi, D., Song, H., & Lewis, M. D. (2019). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, *26*(7), 1217-1233 https://doi.org/10.1177/1073191117711020

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*(3), 253-263. https://doi.org/10.1007/BF02294104

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465–471. https://doi.org/10.1007/BF02293687

Wingate, T. G., & Bourdage, J. S. (2019). Liar at first sight? Early impressions and interviewer judgments, attributions, and false perceptions of faking. *Journal of Personnel Psychology*, *18*(4), 177. https://doi.org/10.1027/1866-5888/a000232

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*(1), 165-200. https://doi.org/10.1111/0081-1750.00078

Table 1

Choices of Chi-square Statistics for Comparing the Fit of Nested Models for Continuous Outcomes

Difference Statistic	**For models estimated using choice:**	**Suitable for:**	**Available for models with missing data?**	**Computable from the two models output?**	**Reference**
D	ML	normal outcomes	Yes	Yes	Steiger, Shapiro, and Browne (1985)
$DSB1_{MLM}$	MLM	non-normal outcomes	No	Yes	Satorra and Bentler (2001)
$DSB10_{MLM}$	MLM	non-normal outcomes	No	Yes[a]	Satorra and Bentler (2010)
$DSB1_{MLR}$	MLR	non-normal outcomes	Yes	Yes	Satorra and Bentler (2001)
$DSB10_{MLR}$	MLR	non-normal outcomes	Yes	Yes[a]	Satorra and Bentler (2010)
D_{MLMV}	MLMV	non-normal outcomes	No	No[b]	Asparouhov and Muthén (2006)

Notes: [a] It requires an additional run of the less restricted model using the parameter estimates of the more restricted model as starting values and with the number of iterations set to 0; [b] software is needed to compute it, at the time of this writing it is only available in Mplus, which directly outputs the difference statistic, df, and p-value (see supplementary materials to this article).

Table 2

Correctly Specified Small Model (p = 10). Empirical Type I Error Rates at the 5% Significance Level

Distribution			*df = 1*						*df = 5*					
Kurt	Skew	N	D	$DSB1_{MLM}$	$DSB1_{MLR}$	$DSB10_{MLM}$	$DSB10_{MLR}$	D_{MLMV}	D	$DSB1_{MLM}$	$DSB1_{MLR}$	$DSB10_{MLM}$	$DSB10_{MLR}$	D_{MLMV}
0.0	0.0	100	0.02	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.02	0.03
0.0	0.0	200	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02
0.0	0.0	500	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
0.0	0.0	1,000	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02
7.0	2.0	100	0.28	0.06	0.12	0.07	0.08	0.07	0.27	0.08	0.14	0.10	0.12	0.06
7.0	2.0	200	0.24	0.06	0.09	0.06	0.07	0.06	0.26	0.05	0.09	0.07	0.09	0.04
7.0	2.0	500	0.27	0.05	0.07	0.06	0.06	0.06	0.27	0.06	0.08	0.07	0.08	0.05
7.0	2.0	1,000	0.25	0.06	0.06	0.06	0.06	0.06	0.26	0.07	0.08	0.07	0.07	0.05
10.0	2.0	100	0.25	0.06	0.13	0.07	0.10	0.07	0.27	0.08	0.17	0.12	0.14	0.06
10.0	2.0	200	0.27	0.05	0.09	0.06	0.08	0.06	0.28	0.05	0.11	0.07	0.09	0.04
10.0	2.0	500	0.29	0.05	0.08	0.05	0.07	0.05	0.32	0.06	0.11	0.08	0.10	0.04
10.0	2.0	1,000	0.31	0.04	0.06	0.05	0.05	0.04	0.31	0.06	0.08	0.06	0.07	0.04

Notes: highlighted values fall outside [.02, .08]; *p* = number of indicators; Kurt = Kurtosis; Skew = Skewness; N = sample size; *df* = degrees of freedom; D = uncorrected ML $\Delta\chi^2$; $DSB1_{MLM}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Satorra-Bentler χ^2 (1994); $DSB1_{MLR}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Asparouhov-Muthén χ^2 (2005); $DSB10_{MLM}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Satorra-Bentler χ^2 (1994); $DSB10_{MLR}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Asparouhov-Muthén χ^2 (2005); D_{MLMV} = Asparouhov-Muthén $\Delta\chi^2$ (2010).

Table 3

Correctly Specified Large Model (p = 30). Empirical Type I Error Rates at the 5% Significance Level

Distribution			*df* = 1						*df* = 15					
Kurt	Skew	N	D	$DSB1_{MLM}$	$DSB1_{MLR}$	$DSB10_{MLM}$	$DSB10_{MLR}$	D_{MLMV}	D	$DSB1_{MLM}$	$DSB1_{MLR}$	$DSB10_{MLM}$	$DSB10_{MLR}$	D_{MLMV}
0.0	0.0	100	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01
0.0	0.0	200	0.02	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01
0.0	0.0	500	0.02	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.01	0.02	0.01	0.01
0.0	0.0	1,000	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01
7.0	2.0	100	0.28	0.09	0.12	0.08	0.09	0.08	0.26	0.13	0.22	0.17	0.18	0.08
7.0	2.0	200	0.25	0.06	0.09	0.06	0.08	0.07	0.26	0.09	0.15	0.11	0.13	0.05
7.0	2.0	500	0.30	0.05	0.06	0.06	0.06	0.06	0.26	0.06	0.08	0.06	0.07	0.05
7.0	2.0	1,000	0.30	0.06	0.07	0.06	0.06	0.06	0.29	0.07	0.08	0.07	0.08	0.04
10.0	2.0	100	0.26	0.08	0.13	0.08	0.09	0.08	0.28	0.14	0.22	0.17	0.18	0.07
10.0	2.0	200	0.29	0.07	0.10	0.07	0.08	0.07	0.30	0.10	0.18	0.13	0.15	0.06
10.0	2.0	500	0.30	0.05	0.08	0.05	0.07	0.05	0.32	0.08	0.12	0.10	0.11	0.05
10.0	2.0	1,000	0.32	0.05	0.07	0.06	0.06	0.06	0.35	0.07	0.10	0.08	0.09	0.04

Notes: highlighted values fall outside [.02, .08]; *p* = number of indicators; Kurt = Kurtosis; Skew = Skewness; N = sample size; *df* = degrees of freedom; D = uncorrected ML $\Delta\chi^2$; $DSB1_{MLM}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Satorra-Bentler χ^2 (1994); $DSB1_{MLR}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Asparouhov-Muthén χ^2 (2005); $DSB10_{MLM}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Satorra-Bentler χ^2 (1994); $DSB10_{MLR}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Asparouhov-Muthén χ^2 (2005); D_{MLMV} = Asparouhov-Muthén $\Delta\chi^2$ (2010).

Table 4

Misspecified Small Model (p = 10). Empirical Rejection Rates (Power) at the 5% Significance Level

	Distribution			*df = 1*						*df = 5*					
Δλ	Kurt	Skew	N	D	$DSB1_{MLM}$	$DSB1_{MLR}$	$DSB10_{MLM}$	$DSB10_{MLR}$	D_{MLMV}	D	$DSB1_{MLM}$	$DSB1_{MLR}$	$DSB10_{MLM}$	$DSB10_{MLR}$	D_{MLMV}
0.2	0.0	0.0	100	0.31	0.32	0.31	0.33	0.30	0.33	0.16	0.17	0.15	0.17	0.14	0.16
	0.0	0.0	200	0.58	0.58	0.57	0.58	0.57	0.58	0.32	0.34	0.33	0.33	0.31	0.32
	0.0	0.0	500	0.94	0.94	0.94	0.94	0.94	0.94	0.78	0.78	0.77	0.78	0.77	0.77
	0.0	0.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99
	7.0	2.0	100	0.47	0.25	0.30	0.27	0.27	0.29	0.46	0.21	0.28	0.25	0.25	0.17
	7.0	2.0	200	0.58	0.33	0.36	0.35	0.35	0.35	0.56	0.29	0.34	0.33	0.33	0.25
	7.0	2.0	500	0.83	0.59	0.59	0.61	0.59	0.61	0.81	0.58	0.59	0.60	0.59	0.54
	7.0	2.0	1,000	0.97	0.88	0.86	0.88	0.87	0.88	0.96	0.87	0.88	0.88	0.88	0.85
	10.0	2.0	100	0.46	0.23	0.30	0.25	0.27	0.25	0.46	0.20	0.30	0.25	0.25	0.18
	10.0	2.0	200	0.62	0.34	0.37	0.36	0.35	0.36	0.60	0.30	0.35	0.34	0.34	0.27
	10.0	2.0	500	0.83	0.59	0.59	0.60	0.60	0.60	0.82	0.56	0.58	0.58	0.58	0.52
	10.0	2.0	1,000	0.95	0.83	0.81	0.83	0.82	0.84	0.96	0.84	0.83	0.84	0.83	0.80
0.4	0.0	0.0	100	0.83	0.84	0.84	0.85	0.83	0.85	0.58	0.59	0.58	0.60	0.56	0.58
	0.0	0.0	200	0.99	0.99	0.99	0.99	0.99	0.99	0.92	0.92	0.92	0.92	0.91	0.92
	0.0	0.0	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.0	0.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	7.0	2.0	100	0.83	0.67	0.63	0.68	0.64	0.69	0.79	0.58	0.59	0.61	0.58	0.53
	7.0	2.0	200	0.96	0.88	0.84	0.89	0.86	0.89	0.95	0.83	0.82	0.85	0.83	0.80
	7.0	2.0	500	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99	0.99
	7.0	2.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10.0	2.0	100	0.82	0.68	0.66	0.70	0.67	0.70	0.79	0.59	0.62	0.64	0.59	0.54
	10.0	2.0	200	0.96	0.90	0.86	0.89	0.87	0.90	0.95	0.83	0.83	0.84	0.83	0.79
	10.0	2.0	500	1.00	0.99	0.98	0.99	0.98	0.99	1.00	0.99	0.98	0.99	0.98	0.98
	10.0	2.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: highlighted conditions have incorrect Type I errors; *p* = number of indicators; Δλ = noninvariance; Kurt = Kurtosis; Skew = Skewness; N = sample size; *df* = degrees of freedom. D = uncorrected ML $\Delta\chi^2$; $DSB1_{MLM}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Satorra-Bentler χ^2 (1994); $DSB1_{MLR}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Asparouhov-Muthén χ^2 (2005); $DSB10_{MLM}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Satorra-Bentler χ^2 (1994); $DSB10_{MLR}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Asparouhov-Muthén χ^2 (2005); D_{MLMV} = Asparouhov-Muthén $\Delta\chi^2$ (2010).

Table 5

Misspecified Small Model (p = 30). Empirical Rejection Rates (Power) at the 5% Significance Level

	Distribution			*df* = 1						*df* = 15					
Δλ	Kurt	Skew	N	D	$DSB1_{MLM}$	$DSB1_{MLR}$	$DSB10_{MLM}$	$DSB10_{MLR}$	D_{MLMV}	D	$DSB1_{MLM}$	$DSB1_{MLR}$	$DSB10_{MLM}$	$DSB10_{MLR}$	D_{MLMV}
0.2	0.0	0.0	100	0.35	0.37	0.37	0.36	0.35	0.36	0.06	0.07	0.07	0.07	0.06	0.05
	0.0	0.0	200	0.65	0.64	0.65	0.65	0.64	0.65	0.15	0.15	0.15	0.15	0.14	0.14
	0.0	0.0	500	0.98	0.97	0.98	0.97	0.97	0.97	0.61	0.61	0.61	0.61	0.61	0.59
	0.0	0.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	0.98	0.98
	7.0	2.0	100	0.47	0.26	0.29	0.28	0.27	0.28	0.45	0.26	0.37	0.31	0.33	0.17
	7.0	2.0	200	0.66	0.39	0.42	0.41	0.40	0.42	0.56	0.32	0.39	0.36	0.37	0.27
	7.0	2.0	500	0.89	0.65	0.64	0.66	0.64	0.66	0.81	0.62	0.64	0.64	0.64	0.58
	7.0	2.0	1,000	0.98	0.89	0.89	0.90	0.89	0.90	0.98	0.89	0.89	0.89	0.89	0.87
	10.0	2.0	100	0.50	0.29	0.34	0.32	0.30	0.31	0.46	0.30	0.38	0.35	0.34	0.21
	10.0	2.0	200	0.64	0.35	0.40	0.39	0.39	0.40	0.59	0.32	0.41	0.38	0.40	0.26
	10.0	2.0	500	0.85	0.61	0.62	0.63	0.62	0.64	0.82	0.58	0.61	0.60	0.61	0.50
	10.0	2.0	1,000	0.97	0.87	0.86	0.87	0.86	0.88	0.97	0.88	0.88	0.89	0.88	0.83
0.4	0.0	0.0	100	0.90	0.90	0.90	0.90	0.90	0.90	0.40	0.42	0.41	0.43	0.38	0.37
	0.0	0.0	200	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.84	0.83	0.84	0.82	0.82
	0.0	0.0	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.0	0.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	7.0	2.0	100	0.89	0.71	0.70	0.75	0.69	0.76	0.81	0.65	0.72	0.69	0.69	0.55
	7.0	2.0	200	0.97	0.92	0.89	0.92	0.89	0.92	0.93	0.86	0.87	0.87	0.86	0.81
	7.0	2.0	500	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	7.0	2.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10.0	2.0	100	0.86	0.71	0.70	0.75	0.70	0.76	0.78	0.64	0.69	0.68	0.66	0.54
	10.0	2.0	200	0.97	0.90	0.87	0.91	0.87	0.91	0.94	0.84	0.85	0.85	0.84	0.77
	10.0	2.0	500	1.00	0.99	0.98	0.99	0.98	0.99	1.00	1.00	0.99	1.00	0.99	0.99
	10.0	2.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: highlighted conditions have incorrect Type I errors; *p* = number of indicators; Δλ = noninvariance; Kurt = Kurtosis; Skew = Skewness; N = sample size; *df* = degrees of freedom. D = uncorrected ML $\Delta\chi^2$; $DSB1_{MLM}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Satorra-Bentler χ^2 (1994); $DSB1_{MLR}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Asparouhov-Muthén χ^2 (2005); $DSB10_{MLM}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Satorra-Bentler χ^2 (1994); $DSB10_{MLR}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Asparouhov-Muthén χ^2 (2005); D_{MLMV} = Asparouhov-Muthén $\Delta\chi^2$ (2010).

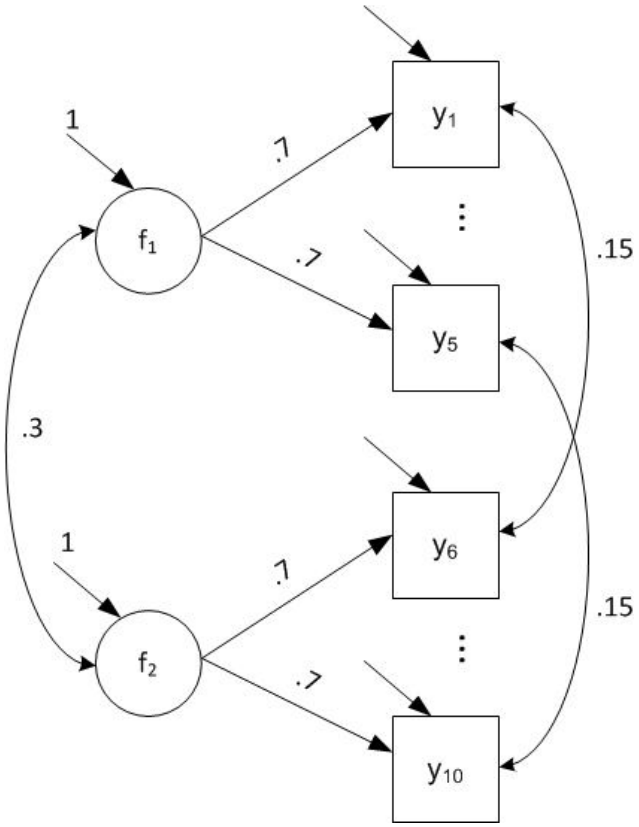

Figure 1. Small model used in the simulations.