Using the Standardized Root Mean Squared Residual (SRMR) to Assess Exact Fit in Structural Equation Models

Goran Pavlov[1,2] ⓘD, Alberto Maydeu-Olivares[1,2] and Dexin Shi[1] ⓘD

Abstract

We examine the accuracy of p values obtained using the asymptotic mean and variance (MV) correction to the distribution of the sample standardized root mean squared residual (SRMR) proposed by Maydeu-Olivares to assess the exact fit of SEM models. In a simulation study, we found that under normality, the MV-corrected SRMR statistic provides reasonably accurate Type I errors even in small samples and for large models, clearly outperforming the current standard, that is, the likelihood ratio (LR) test. When data shows excess kurtosis, MV-corrected SRMR p values are only accurate in small models ($p = 10$), or in medium-sized models ($p = 30$) if no skewness is present and sample sizes are at least 500. Overall, when data are not normal, the MV-corrected LR test seems to outperform the MV-corrected SRMR. We elaborate on these findings by showing that the asymptotic approximation to the mean of the SRMR sampling distribution is quite accurate, while the asymptotic approximation to the standard deviation is not.

Keywords

SRMR, exact fit, structural equation modeling

[1]University of South Carolina, Columbia, SC, USA
[2]University of Barcelona, Barcelona, Spain

Corresponding Author:
Goran Pavlov, Department of Psychology, Barnwell College, University of South Carolina, 1512 Pendleton Street, Columbia, SC 29208, USA.
Email: pavlovg@mailbox.sc.edu

Structural equation modeling (SEM) is a popular technique for modeling multivariate data because it provides a comprehensive framework for fitting theoretical models. Given that SEM is most often used for furthering theory development, a substantial body of literature to date has focused on the issue of how to assess model–data fit (i.e., goodness of fit) in SEM. There appear to be two general perspectives with regard to goodness of fit in SEM. One perspective revolves around the notion that one should not expect to find and thus not seek a model that may be considered as precisely true or correct in the population (e.g., MacCallum et al., 1992). From this perspective, applied researchers should aim at showing that a model provides a good approximation to real-world phenomena, as represented in an observed set of data. To do so, it is generally recommended that multiple approaches to assessment of fit be used (MacCallum, 1990). These may be purely descriptive, involving a comparison of the fitted model to another model, such as a saturated model, or to independence model (Bentler & Bonett, 1980). This perspective appears to be frequently employed, for instance, when fitting exploratory factor analysis models (Lim & Jahng, 2019). From this perspective, assessing whether the model fits the data exactly appears almost unnecessary.

The alternative perspective is concerned with the quality of inferences drawn using the fitted model. From this perspective, assessing the exact fit of a model is important because, provided that alternative equivalent models (Bentler & Satorra, 2010; MacCallum et al., 1993; Stelzl, 1986) can be ruled out theoretically and that the power of the test (Lee et al., 2012; Saris & Satorra, 1993) is sufficiently large, failing to reject the null hypothesis of exact fit enables drawing statistical inferences on the parameter estimates (Bollen & Pearl, 2013; Maydeu-Olivares et al., 2020). Of course, as sample size increases the power to reject the hypothesis of exact model fit increases (Jöreskog, 1967). Also, as model size increases it becomes increasingly difficult to find a well-fitting model, simply due to time constraints (Maydeu-Olivares, 2017a). From this perspective, assessing the exact fit of a model is a meaningful endeavor, always coupled with an assessment of the size of model misfit, with confidence intervals (Maydeu-Olivares, 2017a; Steiger, 1989).

Because sample goodness-of-fit indices are estimators of population quantities, both perspectives can be integrated by using confidence intervals (and if of interest, significance tests) for population effect sizes of misfit. Confidence intervals for the root mean squared error of approximation (RMSEA; Steiger & Lind, 1980; see also Browne & Cudeck, 1993) are well known and routinely used in applications. Steiger (1989) showed that it is possible to obtain confidence intervals for the population goodness-of-fit index (GFI; Jöreskog & Sörbom, 1988; see also MacCallum & Hong, 1997; Maiti & Mukherjee, 1990; Tanaka & Huba, 1985). The sampling distribution of the comparative fit index (CFI; Bentler, 1990) may also be approximated using asymptotic methods (Lai, 2019). Finally, confidence intervals for the standardized root mean squared residual (SRMR; Bentler, 1995) can be obtained using a normal distribution (Maydeu-Olivares, 2017a; Maydeu-Olivares et al., 2018; Ogasawara, 2001). Therefore, if the purpose of the analysis is simply to provide an approximate

representation of the phenomena under investigation, confidence intervals for any of these estimands should be obtained. It is important to use unbiased estimators of the estimands of interest as well as confidence intervals because at small to moderate sample sizes the sample goodness-of-fit indices commonly used in applications can be severely biased and may display a large sampling variability (Maydeu-Olivares et al., 2018; Shi et al., 2019; Steiger, 1990). On the other hand, if the purpose of the analysis is to draw causal inferences on the model parameters, then it makes more sense to test whether the population value of these effect sizes suggests a perfect fit.

The only effect size of model misfit that is currently used in applications is the RMSEA. Put differently, the RMSEA is the only goodness-of-fit index for which SEM software routinely provide a p value for a test of close fit. The null and alternative hypotheses can be written as

H_0^*: RMSEA $\leq$ RMSEA$_0$ versus H_1^*: RMSEA $>$ RMSEA$_0$ where RMSEA$_0$ is an arbitrary population value of the RMSEA. When data are normally distributed, a p value for a test of close fit can be obtained using

$$1 - F_{\chi^2}\left(\chi^2; df, N \times df \times \text{RMSEA}_0^2\right), \tag{1}$$

where N denotes sample size, $F_{\chi^2}(\chi^2; df, \lambda)$ denotes the noncentral chi-square distribution with df degrees of freedom and noncentrality parameter λ (Browne & Cudeck, 1993), and χ^2 denotes the chi-square statistic used to assess the exact fit of the model, usually the likelihood ratio test statistic (e.g., Jöreskog, 1969). We note that expression (1) can also be used to assess the exact fit of the model, that is, RMSEA$_0 = 0$. In this case, the noncentrality parameter $N \times df \times \text{RMSEA}_0^2$ becomes zero, and expression (1) reduces to the familiar equation to obtain a p value for the chi-square test using a central chi-square distribution.

When data are not normal, the most widely used test statistic is the likelihood ratio test statistic, either scaled by its asymptotic mean or adjusted by its asymptotic mean and variance as proposed by Satorra and Bentler (1994). When any of these chi-squares robust to nonnormality is used, (1) is replaced by

$$1 - F_{\chi^2}\left(\chi^2; df, N \times df \times \text{RMSEA}_0^2/c\right), \tag{2}$$

where χ^2 denotes the robust chi-square statistic used, and c denotes its scaling correction (Gao et al., 2020; Savalei, 2018). As in the normal case, expression (2) reduces to the usual chi-square testing in the special case of examining exact fit, for example, H_0^*: RMSEA $= 0$.

Recently, Maydeu-Olivares (2017a) introduced a framework for assessing the size of model misfit using the SRMR. Confidence intervals and, if of interest, tests of close fit can now be performed using the SRMR in addition to the RMSEA. Extant research (Maydeu-Olivares et al., 2018; Shi et al., 2020) has shown that more accurate confidence intervals and test of close fit are obtained using the SRMR than the RMSEA. The latter only provides accurate results in small models.

Maydeu-Olivares (2017a) also provided theory for utilizing the SRMR as a test of exact fit, both under normality assumptions and when data are not normal. In a simulation study, involving a confirmatory factor analysis (CFA) model and sample sizes (N) ranging from 100 to 3,000 observations, the author showed that the SRMR p values were accurate even when the smallest sample sizes were considered. Nevertheless, this simulation study relied on a CFA population model involving only eight variables ($p = 8$) and normally distributed data. In the literature to date, however, it has been repeatedly found that the performance of goodness-of-fit tests worsens as the model size (i.e., the number of variables being modeled) increases (Herzog et al., 2007; Maydeu-Olivares, 2017b; Moshagen, 2012; Shi, Lee, & Terry, 2018; Yuan et al., 2015) and with violations of the normality assumptions (e.g., Hu et al., 1992; Satorra, 1990).

In the current article, we address this gap in the literature and examine whether the SRMR test of exact fit yields accurate p values in a wider range of conditions, involving models of various sizes and both normal and nonnormal data. In addition, we pit the performance of the SRMR against the gold standard for the exact goodness-of-fit assessment, the likelihood ratio test (e.g., Jöreskog, 1969). In the SEM literature, this test statistic is commonly referred to as the chi-square test. In the comparison, we also include the robust, that is, the mean and variance adjusted, chi-square test statistic appropriate for nonnormal data (Asparouhov & Muthén, 2010; Satorra & Bentler, 1994). The remainder of this article is organized as follows. First, we summarize the existing statistical theory for the SRMR. Next, we describe the simulation study conducted to evaluate the accuracy of the asymptotic approximations to the finite sampling distribution of these test statistics. We then summarize the results and provide a discussion of our findings.

The Standardized Root Mean Squared Residual

The Sample SRMR

Let the standardized residual variances and covariances be

$$\hat{\varepsilon}_{ij} = \frac{s_{ij} - \hat{\sigma}_{ij}}{\sqrt{s_{ii}s_{jj}}}, \tag{3}$$

where s_{ij} denotes the sample covariance between variables i and j, with the model implied counterpart $\hat{\sigma}_{ij}$; when $i = j$, s_{ii} and $\hat{\sigma}_{ii}$ denote variances. Then, the sample SRMR (Bentler, 1995; Jöreskog & Sörbom, 1988) is the square root of the average of the squared standardized residual variances and covariances

$$\widehat{SRMSR} = \sqrt{\frac{1}{t} \sum_{i \leq j} \hat{\varepsilon}_{ij}^2} = \sqrt{\frac{1}{t} \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}, \tag{4}$$

where $t = p(p+1)/2$ denotes the number of nonredundant variances and covariances, and $\hat{\varepsilon}$ denotes the vector of t standardized residual covariances (3).

Equation (4) is the SRMR expression computed by the widely used software program LISREL (Jöreskog & Sörbom, 2017) and EQS (Bentler, 2004). It is suitable for assessing how well the assumed (theorized) model reproduces the observed associations among the variables in an interpretable manner. Roughly, it can be interpreted as the average of the absolute value of residual correlations.

On the other hand, the SRMR computed by default in Mplus software (Muthén & Muthén, 2017) is somewhat different:

$$\widehat{\text{SRMSR}}^* = \sqrt{\frac{1}{t+p} \left(\sum_{i \leq j} (\hat{\varepsilon}_{ij}^*)^2 + \sum_i (\hat{\varepsilon}_i^*)^2 \right)}. \tag{5}$$

$$\hat{\varepsilon}_{ij}^* = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}, \quad \hat{\varepsilon}_i^* = \frac{m_i}{\sqrt{s_{ii}}} - \frac{\hat{\mu}_i}{\sqrt{\hat{\sigma}_{ii}}}, \tag{6}$$

where m_i and $\hat{\mu}_i$ denote the sample and expected mean of variable i. It needs to be noted that in Equation (6), $\hat{\sigma}_{ii}$ is used when standardizing the expected covariance. In contrast, in Equation (3), the unrestricted estimate s_{ii} is used for standardization. This need not affect considerably the SRMR values because, in many applications, the estimated variances equal the sample variances, that is, $s_{ii} = \hat{\sigma}_{ii}$. However, the inclusion of the mean structure components $\hat{\varepsilon}_i^*$ in the Mplus version of the SRMR statistic may have a nonnegligible impact. Specifically, in many applications (e.g., in CFA models), the mean structure is saturated, that is, the mean residuals $\hat{\varepsilon}_i^*$ equal zero. In these applications, computing the SRMR in Equation (5) will result in a *lower value* than the value computed using the SRMR in Equation (4). Consequently, because all the SRMR cutoff values provided in the literature (e.g., Hu & Bentler, 1998, 1999; Shi, Maydeu-Olivares, & DiStefano, 2018) have been obtained relying on the LISREL/EQS definition of the SRMR, the utility of these cutoff values when applied to the Mplus SRMR becomes moot.[1] In this article, we focus on models with a saturated mean structure (i.e., no mean structure) and accordingly, on the sample SRMR in Equation (4).

Confidence Intervals for the Population SRMR

The sample SRMR provided in Equation (4) is an estimator of the population SRMR:

$$\text{SRMR} = \sqrt{\frac{1}{t}\sum_{i \leq j} \varepsilon_{ij}^2} = \sqrt{\frac{1}{t}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}, \quad \varepsilon_{ij} = \frac{\sigma_{ij} - \sigma_{ij}^0}{\sqrt{\sigma_{ii}\sigma_{jj}}}. \tag{7}$$

Here, σ_{ij} denotes the true and unknown population covariance between variables i and j (or variance if $i = j$) and σ_{ij}^0 denotes the population covariance (or variance) under the fitted model. The sample SRMR provided in Equation (4), however, is a
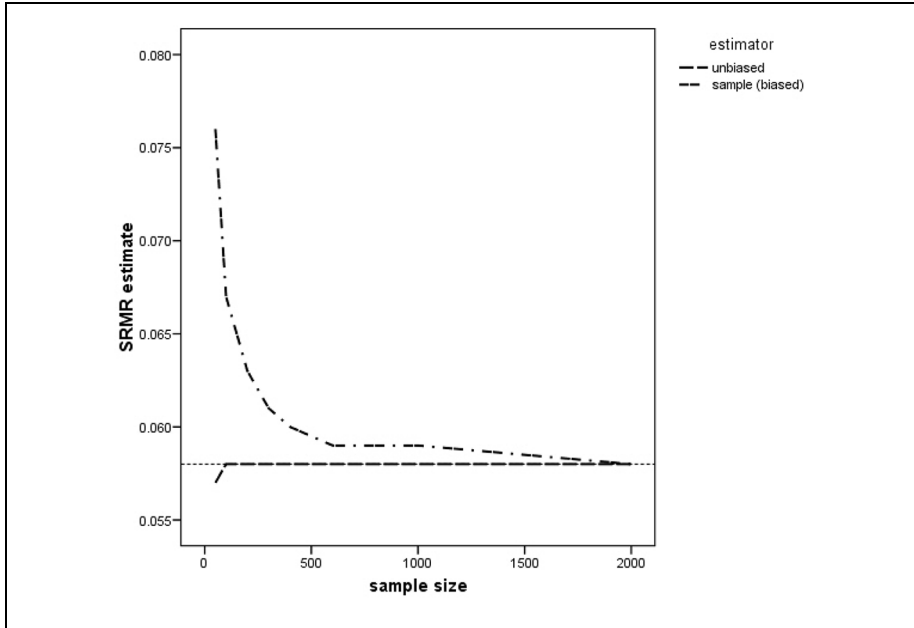
Figure 1. Average sample (i.e., biased) standardized root mean squared residual (SRMR) and unbiased SRMR estimates of the population SRMR of .058 across 1,000 replications as a function of sample size.

biased estimator of the population SRMR in finite samples. To illustrate the potential severity of the bias, we utilize simulation results reported recently by Shi, Maydeu-Olivares, and DiStefano (2018, Table 2). In Figure 1, we provide a plot of the average sample SRMR over 1,000 replications as sample size increases from 50 to 2,000 when the population SRMR = .058. As it can be clearly observed in the figure, the magnitude of the overestimation cannot be neglected for sample sizes smaller than 500 observations.

In Figure 1, we have also plotted the results of Shi, Maydeu-Olivares, and DiStefano (2018, Table 2) for the average unbiased estimator of the SRMR proposed by Maydeu-Olivares (2017a). As the figure reveals, the unbiased estimator of the SRMR is essentially unbiased for sample sizes over 100 observations. The unbiased estimator of the population SRMR proposed by Maydeu-Olivares (2017a) is

$$\widehat{\mathrm{SRMR}}_u = \hat{k}^{-1}\sqrt{\frac{\max\left(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} - \mathrm{tr}(\hat{\boldsymbol{\Xi}}), 0\right)}{t}}, \quad \hat{k} = 1 - \frac{\mathrm{tr}(\hat{\boldsymbol{\Xi}}^2) + 2\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\Xi}}\hat{\boldsymbol{\varepsilon}}}{4(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})^2}, \tag{8}$$

where $\boldsymbol{\Xi}$ denotes the asymptotic covariance matrix of the sample standardized residuals Equation (3), which can be computed either assuming that the observed

variables are normally distributed (NT) or under the asymptotically distribution free assumptions (ADF) put forth by Browne (1982).

Maydeu-Olivares (2017a) proposed using a normal distribution as reference for obtaining confidence intervals and tests of close fit for the population SRMR using the unbiased SRMR estimator. Using this reference distribution, a $(100 - \alpha)\%$ confidence interval for the population SRMR, can be obtained with

$$\Pr\left(\widehat{\mathrm{SRMR}}_u - z_{\alpha/2}SE(\widehat{\mathrm{SRMR}}_u) \leq \widehat{\mathrm{SRMR}} \leq \widehat{\mathrm{SRMR}}_u + z_{\alpha/2}SE(\widehat{\mathrm{SRMR}}_u)\right) = 1 - \alpha, \tag{9}$$

where $z_{\alpha/2}$ denotes the critical value under a standard normal distribution corresponding to a significance level α, and $SE(\widehat{\mathrm{SRMR}}_u)$ denotes the asymptotic standard error of the unbiased SRMR estimate

$$SE(\widehat{\mathrm{SRMR}}_u) = \sqrt{k^{-2}\frac{\mathrm{tr}(\hat{\boldsymbol{\Xi}}^2) + 2\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\Xi}}\hat{\boldsymbol{\varepsilon}}}{2t\ \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}}. \tag{10}$$

Finally, p values for a null hypothesis of close fit, H_0: SRMR $\leq$ SRMR$_0$ versus H_1: SRMR1 $>$ SRMR$_0$, where SRMR$_0$ denotes an arbitrary value of the population SRMR, can be obtained using

$$p = 1 - \Phi\left(\frac{\widehat{\mathrm{SRMR}}_u - \mathrm{SRMR}_0}{SE(\widehat{\mathrm{SRMR}}_u)}\right), \tag{11}$$

where $\Phi(\ )$ denotes a standard normal distribution function.

In needs to be noted that, in principle, these procedures could also be used to test whether a hypothesized SEM model fits exactly. In practice, when the population SRMR equals zero, often $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} - \mathrm{tr}(\hat{\boldsymbol{\Xi}}) < 0$, and the unbiased SRMR estimate is set to zero; see Equation (8). Put differently, when the model fits exactly, the sampling distribution of the $\widehat{\mathrm{SRMR}}u$ must be zero inflated and a normal distribution must provide a poor approximation. See Figure 1 of Shi et al. (2020) for an illustration of this result. Maydeu-Olivares (2017a) suggested that whether a model fits exactly could be tested approximating the sampling distribution of the biased SRMR using a normal distribution.

Testing for Exact Fit Using the SRMR

In SEM models without the mean structure, the null and alternative hypotheses of exact fit are generally written as: $H_0 : \Sigma = \Sigma_0$ versus $H_1 : \Sigma \neq \Sigma_0$, where Σ denotes the unknown population covariance matrix and Σ_0 denotes the population covariance matrix implied by the model. A number of test statistics have been proposed in the SEM literature to assess this null hypothesis of exact fit. In addition to the likelihood ratio test statistic described earlier, researchers may employ, for instance, the residual based chi-square statistic proposed by Browne (1974, 1982; Hayakawa, 2018), the *F*

test proposed by Yuan and Bentler (1999), or the chi-square test proposed by Yuan and Bentler (1997) to name a few.

Maydeu-Olivares (2017a) has proposed an additional test of the exact fit of the model based on the SRMR. The author showed that under the null hypothesis of exact model fit, the mean and standard error of the sample SRMR in (4) can be approximated in large samples using

$$\mu_{\widehat{SRMR}} = \sqrt{\frac{\text{tr}(\mathbf{\Xi})}{t}\left(1 - \frac{\text{tr}(\mathbf{\Xi}^2)}{4\text{tr}(\mathbf{\Xi})^2}\right)}, \tag{12}$$

$$\sigma_{\widehat{SRMR}} = \sqrt{\frac{\text{tr}(\mathbf{\Xi}^2)}{2t \ \text{tr}(\mathbf{\Xi})}}. \tag{13}$$

Then, the sample SRMR can be used to obtain p values for the null hypothesis of exact fit using

$$p = 1 - \Phi(z), \quad z = \frac{\widehat{SRMR} - \mu_{\widehat{SRMR}}}{\sigma_{\widehat{SRMR}}}. \tag{14}$$

To investigate the performance of the method above, Maydeu-Olivares (2017a) performed a simulation study involving a CFA model with eight observed variables ($p = 8$), sample sizes (N) ranging from 100 to 3,000, and normally distributed data. The results revealed that the proposed method provided accurate Type I error rates regardless of the sample size and significance level. Nevertheless, it has been repeatedly found in the literature that the performance of goodness-of-fit tests worsens as model size (i.e., the number of variables being modeled) increases (e.g., Herzog et al., 2007; Maydeu-Olivares, 2017b; Moshagen, 2012; Shi, Lee, & Terry 2018; Yuan et al., 2015). Because the initial evidence on the performance of SRMR was limited to a very small model, it seemed necessary to evaluate the performance of this test statistic also in large models. In addition, the SRMR proposal to assess the exact fit of SEM models was evaluated only in the case of normally distributed data (Maydeu-Olivares, 2017a). However, it has been well documented in the literature that the goodness-of-fit tests (e.g., the likelihood ratio test) fail when data are not normal (e.g., Hu et al., 1992; Satorra, 1990). Accordingly, it seemed warranted to evaluate the performance of the exact fit SRMR proposal also in the case of nonnormal data.

Method

We performed a simulation study to examine the performance of SRMR p values to assess the exact fit of SEM models as introduced by Maydeu-Olivares (2017a). The model used to generate the data was a CFA model because it is the most widely used SEM model in empirical research (DiStefano et al., 2018). The population and fitted models were a one-factor model. We used this simple model because the main aim of

the study was to investigate the performance of SRMR p values under nonnormality and large model size. The population values for all factor loadings were set to be .70, and all residual variances were set to .51.

Data Generation

Data were generated as follows. Using this population CFA model, we first generated continuous data from a multivariate normal distribution. The continuous data were then discretized into seven categories coded 0 to 6. Methodological studies have shown that when the number of response categories is large (i.e., seven), it is appropriate to treat the discretized data as continuous when fitting CFA models (DiStefano & Morgan, 2014; Rhemtulla et al., 2012). Furthermore, we used discretized normal data because in CFA studies it is more common to model discrete ordinal data (i.e., responses to Likert-type items) than continuous data proper (i.e., test scores). Finally, categorizing continuous variables is employed as a widely used method to generate nonnormally distributed data (DiStefano & Morgan, 2014; Maydeu-Olivares, 2017b; Muthén & Kaplan, 1985).

Study Conditions

The simulation conditions were obtained by manipulating the following three factors: (a) sample size, (b) model size, and (c) level of nonnormality.

Sample Size. Sample sizes included 100, 200, 500, and 1,000 observations. The sample sizes were selected to reflect a range of small to large samples commonly used in psychological research.

Model Size. Model size refers to the total number of observed variables, p (Moshagen, 2012; Shi, Lee, & Terry, 2018). We used three different levels for the number of observed variables: small ($p = 10$), medium ($p = 30$), and large ($p = 60$) models.

Level of Nonnormality. Three levels of nonnormality were obtained by manipulating the population values of the skewness and (excess) kurtosis: (a) skewness = 0.00, kurtosis = 0.00 (i.e., normal data), (b) skewness = 0.00, kurtosis = 3.30, and (c) skewness = -2.00, kurtosis = 3.30. To achieve the designed skewness and kurtosis, the continuous data were discretized using selected threshold values (Maydeu-Olivares, 2017b; Muthén & Kaplan, 1985). The threshold values used for data generation and the expected area under the curve for each response category are presented in Table 1. The technical details for computing the population skewness and kurtosis given a set of thresholds can be found in Maydeu-Olivares et al. (2007).

In sum, the simulation study consisted of a fully crossed design including four sample sizes, three distributional shapes, and three model sizes. Thirty-six conditions were created in total ($4 \times 3 \times 3$). For each of the 36 simulated conditions, 1,000

Table 1. Target Item Category Probabilities and Corresponding Threshold Values Used to Generate the Data.

								Expected area under the curve						
Kurtosis	Skewness			Thresholds				0	1	2	3	4	5	6
0	0	−1.64,	−1.08,	−0.52,	0.52,	1.08,	1.64	5%	9%	16%	40%	16%	9%	5%
3.3	0	−2.33,	−1.64,	−1.04,	1.04,	1.64,	2.33	1%	4%	10%	70%	10%	4%	1%
3.3	−2.0	−2.33,	−1.88,	−1.55,	−1.17,	−0.84,	−0.55	1%	2%	3%	6%	8%	10%	70%

replications were generated with the *simsem* package in R (Pornprasertmanit et al., 2013; R Core Team, 2019).

Estimation

For each simulated data set, we fitted a one-factor CFA model with the maximum likelihood estimation method using the *lavaan* package in R (Rosseel, 2012). In the supplementary materials to this article, we provide R code for computing the exact fit test using SRMR. The SRMR test statistic Equation (14) was obtained under both NT and ADF) assumptions. Different values of this statistic based on the SRMR to assess the exact fit of the model are obtained under NT and ADF assumptions because the asymptotic covariance matrix of the standardized residual covariances, Ξ is computed differently. For computational details of the two SRMR test statistics the reader is referred to Maydeu-Olivares (2017a).

To benchmark the performance of the SRMR as a test of exact fit, we used the likelihood ratio (Jöreskog, 1969) test, also commonly known as the chi-square test (χ^2). The chi-square test statistic was also obtained both NT and ADF assumptions. The χ^2 statistic computed under normality is the likelihood ratio test. The χ^2 statistic computed under ADF is the mean and variance adjusted likelihood ratio test statistic proposed by Asparouhov and Muthén (2010; see also Satorra & Bentler, 1994). For both χ^2 and SRMR statistics, we evaluated the empirical rejection rates, that is, Type I error rates using nominal alpha levels of 5%.

Results

For all the study conditions all replications successfully converged. Accordingly, results for each of the 36 conditions under investigation were based on all 1,000 replications.

We provide in Table 2 the empirical rejection rates at the 5% significance level of the χ^2 and SRMR tests of exact fit. Following Bradley (1978), and taking into account that we used only 1,000 replications, we considered Type I error rates in [.02, .08] to be adequate. Conditions that fall outside this range are highlighted in Table 2.

The results presented in Table 2 for the χ^2 statistic were consistent with previous findings in the literature. Specifically, the χ^2 computed under normality assumption

Table 2. Empirical Rejection Rates at the 5% Significance Level of the Chi-Square and SRMR Tests of Exact Fit.

Kurtosis	Skewness	p	N	NT		ADF	
				χ^2	SRMR	χ^2	SRMR
0.0	0.0	10	100	0.08	0.03	0.08	0.03
			200	0.08	0.05	0.08	0.04
			500	0.05	0.05	0.04	0.03
			1,000	0.07	0.06	0.06	0.05
		30	100	0.67	0.03	0.68	0.01
			200	0.28	0.04	0.26	0.02
			500	0.13	0.06	0.10	0.05
			1,000	0.10	0.07	0.07	0.06
		60	100	1.00	0.01	1.00	0.00
			200	0.99	0.04	0.99	0.00
			500	0.50	0.07	0.43	0.00
			1,000	0.25	0.08	0.18	0.00
3.3	0.0	10	100	0.38	0.24	0.10	0.05
			200	0.32	0.26	0.07	0.03
			500	0.29	0.27	0.05	0.04
			1,000	0.31	0.30	0.06	0.05
		30	100	1.00	0.86	0.74	0.00
			200	0.99	0.92	0.25	0.00
			500	0.96	0.94	0.09	0.03
			1,000	0.97	0.95	0.07	0.05
		60	100	1.00	1.00	1.00	0.00
			200	1.00	1.00	0.98	0.00
			500	1.00	1.00	0.38	0.00
			1,000	1.00	1.00	0.15	0.00
3.3	−2.0	10	100	0.87	0.75	0.12	0.04
			200	0.85	0.79	0.09	0.04
			500	0.85	0.83	0.06	0.07
			1,000	0.84	0.83	0.05	0.07
		30	100	1.00	1.00	0.87	0.00
			200	1.00	1.00	0.34	0.00
			500	1.00	1.00	0.12	0.00
			1,000	1.00	1.00	0.07	0.00
		60	100	1.00	1.00	1.00	0.00
			200	1.00	1.00	1.00	0.00
			500	1.00	1.00	0.54	0.00
			1,000	1.00	1.00	0.20	0.00

Note. Highlighted are conditions with adequate Type I errors. The asymptotic covariance matrix of the residual covariances used to compute p values for the SRMR is computed differently under normality and ADF assumptions. SRMR = standardized root mean squared residual; p = number of variables; NT = under normality; ADF = asymptotically distribution free; χ^2 = likelihood ratio (LR) test (under normality) and mean and variance LR under ADF assumptions.

(NT in the table) overrejected the true model when data were nonnormal. Furthermore, the rejection rates increased as the model size increased. For the nonnormal conditions investigated, as soon as $p = 30$, the test almost always rejected the model. In fact, the only conditions investigated for which the test maintained adequate Type I error rates involved normal data and a small model ($p = 10$). For normal data and larger models ($p \geq 30$), the NT χ^2 statistic converged slowly to its asymptotic distribution, but even the largest sample size considered (1,000) was insufficient to obtain accurate Type I error rates.

We also see in Table 2 that with the increasing number of variables, the robust χ^2 (ADF in the table) converged faster than the NT χ^2 to its reference distribution, that is, it was more robust to the model size effect. This is consistent with previous findings in the literature (e.g., Maydeu-Olivares, 2017b). Under normality, the robust χ^2 achieved adequate Type I errors when $p = 30$ with 1,000 observations. However, sample sizes larger than 1,000 are needed for this statistic to yield accurate Type I error rates when $p = 60$. As expected, the ADF χ^2 was also more robust to the effect of nonnormality. Specifically, p values were acceptable for $p = 10$ and the minimum sample size needed to achieve them varied depending on the level of kurtosis and skewness in the data. A minimum of 100 observations was needed when the data shows neither (excess) kurtosis nor skewness (i.e., normal data), 200 observations when the data showed only excess kurtosis, and of 500 observations when both kurtosis and skewness were present. For $p = 30$, larger sample sizes (i.e., 1,000 observations) were needed for the test to yield nominal Type I error rates. Finally, for $p = 60$, not even the largest sample sizes (i.e., 1,000) were sufficient to obtain accurate Type I error rates.

Results for the test of exact fit using the SRMR revealed a pattern different from the one observed for the χ^2 test statistic. When performed under normality assumptions (NT in Table 2), the SRMR test yielded adequate Type I error rates for all conditions involving normally distributed data and smaller models ($p \leq 30$). These findings were in line with the results reported by Maydeu-Olivares (2017a). The Type I error rates were inaccurate (i.e., the test was underrejecting) only when the largest model and smallest sample size were considered ($p = 60$, $N = 100$). Overall, with normal data, the NT SRMR test statistic clearly outperformed the NT χ^2 (i.e., the likelihood ratio test). On the other hand, with nonnormal data, the NT SRMR test of exact fit consistently overrejected and its behavior closely resembles the behavior of the NT χ^2 statistic.

When data were normal and $p = 10$, the robust SRMR (ADF in Table 2) and robust χ^2 yielded comparable and adequate results. Conversely, when $p = 30$, a sample of 200 observations sufficed to obtain adequate p values using the robust SRMR, whereas 1,000 observations were needed using the robust χ^2. When $p = 60$, the robust SRMR underrejected the null hypothesis even at the largest sample size considered.

When data showed excess kurtosis but no skewness, the SRMR provided more accurate Type I error rates than the robust χ^2 in small models and small samples ($p = 10$, $N = 100$), slightly better results in medium size models and large samples ($p = 30$,

$N \geq 500$) but was consistently underrejecting when the largest model size considered ($p = 60$). Most interestingly, the behavior of the SRMR exact fit test was adversely affected by the skewness of data. When data showed both (excess) kurtosis and skewness, even though it was performing adequately in conditions with small models ($p = 10$), the robust SRMR was underrejecting the model in all conditions involving $p \geq 30$ observed variables. In these conditions ($p \geq 30$), the Type I error rates of the robust χ^2 were gradually returning to their nominal levels with the increasing sample size, while the same effect was not observed for the robust SRMR.

Discussion

In the present study, we have examined the accuracy of the asymptotic mean and variance correction to the distribution of the sample SRMR proposed by Maydeu-Olivares (2017a) to assess the exact fit of SEM models. Several model sizes, sample sizes, and levels of nonnormality were considered, and the SRMR was computed under both normal theory (NT) and ADF assumptions. In addition, the SRMR accuracy was pitted against the gold standard for the exact goodness-of-fit assessment, the likelihood ratio test (e.g., Jöreskog, 1969), and its robust (ADF) version obtained by adjusting the likelihood ratio statistic by its asymptotic mean and variance (Asparouhov & Muthén, 2010; Satorra & Bentler, 1994).

Overall, the results revealed that the mean and variance corrected SRMR statistic provides reasonably accurate Type I errors when data shows neither excess kurtosis nor skewness in small samples and even in large models ($p = 60$, $N = 200$), in which the likelihood ratio test statistic fails. In other words, when data are normal, the mean and variance corrected SRMR outperforms the current standard. When data shows excess kurtosis, Type I errors of the mean and variance corrected SRMR are accurate only in small models ($p = 10$), or in medium-sized models ($p = 30$) if no skewness is present and sample is large enough ($N \geq 500$). Overall, it seems that the current standard, that is, the mean and variance corrected likelihood ratio test statistic, outperforms the mean and variance corrected SRMR when data are not normal.

The robust χ^2 and SRMR test statistics considered in this article are both mean and variance corrected statistics of the type

$$T_a = a + bT, \tag{15}$$

where T_a denotes the mean and variance corrected statistic used for testing, and T denotes the original sample statistic. In the case of the robust χ^2, we write $\chi_a^2 = a + b\chi^2$, where χ_a^2 denotes the mean and variance adjusted chi-square statistic and χ^2 is the likelihood ratio test statistic. a and b are constants such that χ_a^2 agrees asymptotically in mean and variance with a reference chi-square distribution with the model's degrees of freedom. However, the asymptotic distribution of the robust χ^2 is not chi-square; it is a mixture of one degree of freedom chi-squares (Satorra & Bentler, 1994). This implies that as sample size increases, the behavior of the robust χ^2 p values need not improve.

As our results show, with nonnormal data, the approximation's behavior improves with increasing sample size. However, it is important to note that our simulation involved discretized normal data. With other algorithms to generate nonnormal data, this need not be the case (for instance, see Gao et al., 2019). In fact, one should rather expect the accuracy of the robust χ^2 p values to improve up to a sample size, and slightly worsen after that, reflecting that the reference distribution to obtain the p values is not the actual asymptotic distribution of the χ^2 statistic.

In the case of the robust SRMR, we write $z = a + b \; \widehat{SRMR}$, where a and b are constants such that z agrees asymptotically with a standard normal reference distribution. Obviously, in this case, $a = -\mu_{\widehat{SRMR}} \sigma^{-1}_{\widehat{SRMR}}$, $b = \sigma^{-1}_{\widehat{SRMR}}$, and this is the solution proposed in Equation (14). The mean and variance adjustment is also used to obtain p values for the SRMR in the normal case, and the difference between the normal and robust SRMR options lies in how the asymptotic covariance matrix of the standardized residual covariances is estimated (see Maydeu-Olivares, 2017a). It is important to note here that the use of normal distribution as a reference distribution is heuristic, and it remains to be proved that the sampling distribution of the sample SRMR converges to normality. Nevertheless, the approximation seems to work very well in practice.
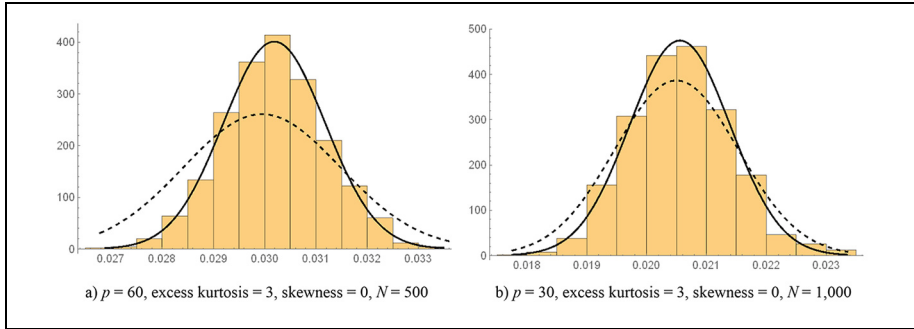
Why do p values for the robust SRMR fail to be accurate in many of the nonnormal conditions investigated in this study? One plausible explanation is that the asymptotic approximation proposed by Maydeu-Olivares (2017a) to the empirical standard deviation of the $\widehat{SRMR}$ is not sufficiently accurate. To explore this, for each simulated condition, we calculated the average SRMR estimates and empirical variances across replications and compared them to the values based on the theoretical normal reference distributions. In Table 3, we provide the empirical mean and standard deviation of the $\widehat{SRMR}$ for each of the conditions of our simulation study, that is, the mean and standard deviation of the $\widehat{SRMR}$ across the 1,000 replications for each condition. We also provide in this table the expected mean and standard deviation for each condition computed using Equations (12) and (13) under both NT and ADF assumptions. It may be observed in Table 3 that under NT, the asymptotic approximation to the empirical mean is quite accurate for all conditions involving normally distributed data. Conversely, it underestimates the empirical mean for all nonnormal conditions. The asymptotic approximation underestimates the empirical standard deviation but, as expected, it improves as sample size increases. Under ADF assumptions, the asymptotic approximation to the empirical mean is fairly accurate for all conditions investigated (the relative bias is 5% at most). Nevertheless, it overestimates the empirical standard deviation of the $\widehat{SRMR}$, and it does not improve swiftly as sample size increases. As a result, for many nonnormal conditions, the mean and variance corrected $\widehat{SRMR}$ statistic provides inaccurate p values.

We illustrate this issue in Figure 2. In this figure, we provide histograms of the $\widehat{SRMR}$ across all 1,000 replications for two selected conditions. For each condition, we have plotted a normal reference distribution using the empirical mean and standard deviation (solid line) and using the expected mean and standard deviation

Table 3. Accuracy of the Asymptotic Approximation to the Samplig Distribution of the Sample SRMR Across 1,000 Replications. Test of Normality, Observed Versus Expected Mean (*M*) and Standard Deviation (*SD*).

| | | | | SRMR | | | | | | Test of normality | |
| | | | | Observed | | Expected (NT) | | Expected (ADF) | | | |
Kurtosis	Skewness	*p*	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*SW*	*p* value
0.0	0.0	10	100	0.044	.0060	0.043	.0056	0.043	.0067	.9975	.13
			200	0.031	.0040	0.030	.0038	0.030	.0043	.9979	.25
			500	0.019	.0024	0.019	.0024	0.019	.0025	.9978	.20
			1,000	0.014	.0017	0.014	.0017	0.014	.0017	.9951	<.01
		30	100	0.051	.0038	0.051	.0023	0.050	.0045	.9979	.25
			200	0.036	.0020	0.036	.0014	0.036	.0024	.9984	.46
			500	0.023	.0010	0.023	.0008	0.023	.0012	.9976	.15
			1,000	0.016	.0007	0.016	.0006	0.016	.0007	.9988	.78
		60	100	0.053	.0034	0.053	.0014	0.052	.0041	.9962	.01
			200	0.037	.0018	0.037	.0008	0.037	.0021	.9981	.31
			500	0.024	.0008	0.023	.0005	0.023	.0009	.9984	.47
			1,000	0.017	.0004	0.017	.0003	0.017	.0005	.9990	.88
3.3	0.0	10	100	0.056	.0072	0.050	.0066	0.054	.0098	.9942	<.01
			200	0.039	.0050	0.035	.0044	0.038	.0060	.9951	<.01
			500	0.025	.0030	0.022	.0027	0.025	.0033	.9974	.10
			1,000	0.017	.0022	0.016	.0019	0.017	.0022	.9944	<.01
		30	100	0.066	.0041	0.058	.0028	0.063	.0074	.9979	.26
			200	0.046	.0023	0.041	.0017	0.045	.0040	.9980	.30
			500	0.029	.0012	0.026	.0010	0.029	.0018	.9989	.83
			1,000	0.021	.0008	0.018	.0007	0.020	.0010	.9979	.26
		60	100	0.068	.0035	0.060	.0018	0.066	.0070	.9976	.16
			200	0.048	.0018	0.042	.0010	0.047	.0036	.9980	.28
			500	0.030	.0009	0.027	.0005	0.030	.0015	.9990	.89
			1,000	0.021	.0005	0.019	.0004	0.021	.0008	.9993	.99
3.3	−2.0	10	100	0.068	.0101	0.051	.0070	0.065	.0129	.9955	<.01
			200	0.048	.0064	0.036	.0046	0.047	.0077	.9932	<.01
			500	0.030	.0037	0.022	.0028	0.030	.0042	.9976	.14
			1,000	0.021	.0026	0.016	.0019	0.021	.0028	.9979	.25
		30	100	0.079	.0067	0.059	.0032	0.077	.0104	.9969	.05
			200	0.056	.0035	0.042	.0018	0.055	.0055	.9973	.10
			500	0.035	.0018	0.026	.0010	0.035	.0024	.9981	.33
			1,000	0.025	.0011	0.019	.0007	0.025	.0014	.9984	.50
		60	100	0.082	.0060	0.061	.0022	0.080	.0100	.9975	.14
			200	0.058	.0032	0.043	.0011	0.057	.0052	.9990	.85
			500	0.037	.0013	0.027	.0006	0.036	.0021	.9979	.24
			1,000	0.026	.0007	0.019	.0004	0.026	.0011	.9978	.20

Note. SRMR = standardized root mean squared residual; *p* = number of variables; *N* = sample size; NT = under normality; ADF = asymptotically distribution free; SW = Shapiro–Wilk test statistic.

a) $p = 60$, excess kurtosis = 3, skewness = 0, $N = 500$ b) $p = 30$, excess kurtosis = 3, skewness = 0, $N = 1,000$

Figure 2. Empirical distribution of the sample standardized root mean squared residual (SRMR) and reference normal distributions. (a) $p = 60$, excess kurtosis = 3, skewness = 0, $N = 500$; (b) $p = 30$, excess kurtosis = 3, skewness = 0, $N = 1,000$.
Note. Solid line = empirical mean (*M*) and standard deviation (*SD*) across 1,000 replications; dotted line = expected *M* and *SD* using asymptotic methods; *p* values computed using the expected *M* and *SD* and yield adequate Type I errors in (b), but not in (a).

(dotted line). In both cases, it may be observed that a normal distribution using the empirical mean and standard deviation provides a very good approximation to the sampling distribution of the $\widehat{\text{SRMR}}$. In the condition with $N = 500$, $p = 60$, (excess) kurtosis = 3, and skewness = 0, the expected mean of the $\widehat{\text{SRMR}}$ underestimates the empirical mean by approximately 1%, but the expected standard deviations overestimates the empirical standard deviation by roughly 54% (see Table 3). The latter has quite a dramatic effect on the accuracy of the *p* values obtained using the mean and variance corrected $\widehat{\text{SRMR}}$. These are obtained using the dotted line in Figure 2, and as the figure reveals, *p* values in the lower tail will be quite inflated.

In the other condition displayed in Figure 2, with $N = 1,000$, $p = 30$, (excess) kurtosis = 3, and skewness = 0, the relative bias of the expected mean of the $\widehat{\text{SRMR}}$ is less than 1%, and the relative bias of the expected standard deviation is ''only'' 23%. Nevertheless, despite the substantial bias, the left tail probabilities are reasonably accurate.

As depicted in Figure 2, distribution of the sample SRMR appears to be quite normal. To further assess the quality of the normal approximation to the distribution of the sample SRMR, we performed Shapiro and Wilk's (1965) test of normality for each of the investigated conditions. We chose this particular test as it has been shown to be the most powerful to detect departures from normality (Yap & Sim, 2011). The test statistic ranges from 0 to 1, with 1 indicating perfect fit. In out study, the statistic ranged from .993 to .999 across conditions (see Table 3), indicating that a normal distribution provides a good fit to the sampling distribution of the SRMR. We have also provided in Table 3 *p* values for this test statistic because they may more clearly pinpoint conditions under which the normal approximation works best. As it may be observed in the table, the main driver of the accuracy of the normal approximation is

model size. Specifically, the normal approximation is somewhat poorer when the number of observed variables is small (i.e., $p = 10$).

Concluding Remarks

In the current study, we investigated whether a recently proposed test statistic (based on the SRMR) outperforms the current standard tests to evaluate the exact fit of structural equation models in terms of Type I errors. We conclude that the answer is negative. Because the current standard test statistics are a side product of the computations involved in obtaining maximum likelihood parameter estimates and standard errors, the current test statistics are to be preferred to the new proposal. We have not compared the power of both approaches as it only makes sense to compare the power of test statistics when accurate Type I errors are obtained, which was not the case in many of the conditions investigated.

The accuracy of the SRMR test of exact fit depends on the accuracy of the reference nomal distribution to the sampling distribution of the SRMR, and on the accuracy of the asymptotic approximation to the empirical mean and standard deviation of the sampling distribution of the SRMR. We found that the proposed reference normal distribution provides a good approximation to the sampling distribution of the SRMR when the model fits exactly, but additional statistical theory is needed to support the use of this reference distribution. We also found that the asymptotic approximation to the mean of the SRMR sampling distribution is quite accurate, but that the asymptotic approximation to the standard deviation is not. Under normality assumptions, the asymptotic approximation underestimates the empirical standard deviation; under asymptotically distribution free assumptions, it overestimates it. The reason for the differential accuracy of the asymptotic approximations to the empirical mean and standard deviation is that two terms are used to approximate the mean, but only one term is used to approximate the standard deviation (for technical details, see Maydeu-Olivares, 2017a). The present study sugests that a two-term approximation is needed also for the standard deviation. Further statistical theory is required to obtain a better asymptotic approximation to the empirical sampling distribution of the SRMR and to support the use of a reference normal distribution.

Declaration of Conflicting Interests

Funding

ORCID iDs

Goran Pavlov 🔟 https://orcid.org/0000-0001-8769-8904
Dexin Shi 🔟 https://orcid.org/0000-0002-4120-6756

Supplemental Material

Supplemental material for this article is available online.

Note

1. To be able to pit the Mplus results against the SRMR cutoff values published in the literature, Mplus users should use MODEL=NOMEANSTRUCTURE in the ANALYSIS command. In this case, Mplus computes the SRMR given by Equation (4; Asparouhov & Muthén, 2018).

References

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

Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus*. https://www.statmodel.com/download/SRMR2.pdf

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS 5* [Computer program]. Multivariate Software Inc.

Bentler, P. M. (2004). *EQS 6* [Computer program]. Multivariate Software Inc.

Bentler, P. M., & Bonett, D. D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588-606. https://doi.org/10.1037//0033-2909.88.3.588

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

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301-328). Springer. https://doi.org/10.1007/978-94-007-6094-3

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

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*(1), 1-24. https://journals.co.za/content/sasj/8/1/AJA0038271X_175

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72-141). Cambridge University Press.

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

DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling*, *25*(3), 453-466. https://doi.org/10.1080/10705511.2017.1390394

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, *21*(3), 425-438. https://doi.org/10.1080/10705511.2014.915373

Gao, C., Shi, D., & Maydeu-Olivares, A. (2020). Estimating the maximum likelihood root mean square error of approximation (RMSEA) with non-normal data: A Monte-Carlo study. *Structural Equation Modeling*, *27*(2), 192-201. https://doi.org/10.1080/10705511.2019.1637741

Hayakawa, K. (2018). Corrected goodness-of-fit test in covariance structure analysis. *Psychological Methods*, *24*(3), 371-389. https://doi.org/10.1037/met0000180

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., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453. https://doi.org/10.1037//1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hu, L., 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

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443-482. https://doi.org/10.1007/BF02289658

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202. https://doi.org/10.1007/BF02289343

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7. A guide to the program and applications* (2nd ed.). International Education Services.

Jöreskog, K. G., & Sörbom, D. (2017). LISREL (Version 9.3) [Computer program]. Scientific Software International.

Lai, K. (2019). A simple analytic confidence interval for CFI given nonnormal data. *Structural Equation Modeling*, *26*(5), 757-777. https://doi.org/10.1080/10705511.2018.1562351

Lee, T., Cai, L., & MacCallum, R. C. (2012). Power analysis for tests of structural equation models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 181-194). Guilford Press.

Lim, S., & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, *24*(4), 452-467. https://doi.org/10.1037/met0000230

MacCallum, R. C. (1990). The need for alternative measures of fit in covariance structure modeling. *Multivariate Behavioral Research*, *25*(2), 157-162. https://doi.org/10.1207/s15327906mbr2502_2

MacCallum, R. C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research*, *32*(2), 193-210. https://doi.org/10.1207/s15327906mbr3202_5

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490-504. https://doi.org/10.1037/0033-2909.111.3.490

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*(1), 185-199. https://doi.org/10.1037/0033-2909.114.1.185

Maiti, S. S., & Mukherjee, B. N. (1990). A note on distributional properties of the Jöreskog-Sörbom fit indices. *Psychometrika*, *55*(4), 721-726. https://doi.org/10.1007/BF02294619

Maydeu-Olivares, A. (2017a). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533-558. https://doi.org/10.1007/s11336-016-9552-7

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

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*(2), 157-176. https://doi.org/10.1037/1082-989X.12.2.157

Maydeu-Olivares, A., Shi, D., & Fairchild, A. J. (2020). Estimating causal effects in linear regression models with observational data: The instrumental variables regression model. *Psychological Methods*, *25*(2), 243-258. https://doi.org/10.1037/met0000226

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling*, *25*(3), 389-402. https://doi.org/10.1080/10705511.2017.1389611

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, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171-189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, L. K., & Muthén, B. O. (2017). *Mplus 8*. Muthén & Muthén.

Ogasawara, H. (2001). Standard errors of fit indices using residuals in structural equation modeling. *Psychometrika*, *66*(3), 421-436. https://doi.org/10.1007/BF02294443

Pornprasertmanit, S., Miller, P., & Schoemann, A. (2013). simsem: Simulated structural equation modeling. *R Package Version 0.5-3*.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rhemtulla, M., Brosseau-Liard, P. É. E., & 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. https://doi.org/10.18637/jss.v048.i02

Saris, W., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181-204). Sage.

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., & 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). Sage.

Savalei, V. (2018). On the computation of the RMSEA and CFI from the mean-and-variance corrected test statistic with nonnormal data in SEM. *Multivariate Behavioral Research*, *53*(3), 1-11. https://doi.org/10.1080/00273171.2018.1455142

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611. https://doi.org/10.2307/2333709

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310-334. https://doi.org/10.1177/0013164418783530

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

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research*, *53*(5), 676-694. https://doi.org/10.1080/00273171.2018.1476221

Shi, D., Maydeu-Olivares, A., & Rosseel, Y. (2020). Assessing fit in ordinal factor analysis models: SRMR vs. RMSEA. *Structural Equation Modeling*, *27*(1), 1-15. https://doi.org/10.1080/10705511.2019.1611434

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH*. Systat, Inc.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173-180. https://doi.org/10.1207/s15327906mbr2502_4

Steiger, J. H., & Lind, J. C. (1980, May 30). *Statistically-based tests for the number of common factors* [Paper presentation]. Annual Meeting of the Psychometric Society, Iowa City, IA, United States.

Stelzl, I. (1986). Changing the causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, *21*(3), 309-331. https://doi.org/10.1207/s15327906mbr2103_3

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 197-201. https://doi.org/10.1111/j.2044-8317.1985.tb00834.x

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141-2155. https://doi.org/10.1080/00949655.2010.520163

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, *92*(438), 767-774. https://doi.org/10.1080/01621459.1997.10474029

Yuan, K.-H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, *24*(3), 225-243. https://doi.org/10.3102/10769986024003225

Yuan, K. H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika*, *80*(2), 379-405. https://doi.org/10.1007/s11336-013-9386-5