UNIVERSITAT DE BARCELONA

Testing for exact model fit and model comparison in structural equation modeling under non-normality

Goran Pavlov

UNIVERSITAT DE
BARCELONA

Testing for exact model fit and model comparison in structural equation modeling under non-normality

Goran Pavlov

University of Barcelona
Department of Clinical Psychology and Psychobiology
Section of Personality and Psychological Assessment

Ph.D. Program in Brain, Cognition and Behavior
Research Line: Quantitative Psychology

Supervisor: Prof. Dr. Alberto Maydeu-Olivares

April 2023

Acknowledgements

This doctoral thesis has taken an enormous amount of support from my supervisor, my family, and my colleagues, without whom I would not have been able to complete this work. Thank you!

INDEX

ABSTRACT

Structural equation modeling (SEM) is a versatile framework that allows researchers to estimate systems of equations and test theoretical models. A significant portion of the literature on SEM focuses on model fit and selection, where researchers are interested in evaluating the goodness of fit of a theoretical model (absolute fit) or comparing multiple plausible models (relative fit). Evaluating exact or approximate fit is possible in both cases. The current doctoral thesis is a compilation of two published studies that contribute to the literature on both absolute and relative fit.

The first study aimed to compare the accuracy of assessing exact model fit using two tests, namely the mean and variance adjusted chi-square test and the recently developed robust version of the Standardized Root Mean Squared Residual (SRMR) test, in situations where data is not normal. Through simulation, the study examined the impact of factors such as (non)normality, sample size, and model size on test accuracy. The results showed that the robust chi-square test outperformed the robust SRMR test with respect to Type I error rates and was less affected by model size.

The second study investigated the accuracy of evaluating relative model fit using several versions of chi-square difference tests that are robust to violations of normality. The study manipulated levels of (non)normality, sample size, model size, and degrees of freedom for the difference test through simulation. The results showed that the mean and variance adjusted chi-square difference test performed accurately across all investigated conditions and outperformed its mean-adjusted competitors, which required larger samples to perform adequately.

In summary, the two studies in the doctoral thesis contribute to the literature on both absolute and relative fit in SEM. The findings suggest that the robust chi-square test is more accurate in assessing exact model fit than the robust SRMR test, and the mean and variance adjusted chi-square difference test is a reliable method for evaluating relative model fit in SEM.

RESUMEN

La modelización de ecuaciones estructurales (SEM, por sus siglas en inglés) es un marco general para estimar sistemas de ecuaciones. Debido a su generalidad y flexibilidad, SEM puede utilizarse para evaluar modelos teóricos y existe una cantidad sustancial de literatura centrada en la bondad de ajuste y en la selección de modelos. Específicamente, dado un modelo teórico, los investigadores están interesados en evaluar su bondad de ajuste (también conocida como ajuste del modelo a los datos o ajuste absoluto). Cuando hay varios modelos teóricamente plausibles, también están interesados en la selección del modelo (también conocida como ajuste relativo o comparación de modelos). En ambos casos, es posible evaluar el ajuste exacto o el aproximado. La presente tesis doctoral es una compilación de dos estudios publicados que contribuyen a la literatura de ajuste absoluto y relativo.

El primer estudio tuvo como objetivo comparar la precisión de la evaluación del ajuste exacto del modelo utilizando dos pruebas: la prueba chi-cuadrado ajustada por media y varianza y la versión robusta recientemente desarrollada de la prueba de la raíz cuadrada media estandarizada residual (SRMR por sus siglas en inglés), en situaciones donde los datos no son normales. A través de simulaciones, el estudio examinó el impacto de factores como la (no)normalidad, el tamaño de la muestra y el tamaño del modelo en la precisión de las pruebas. Los resultados mostraron que la prueba chi-cuadrado robusta superó a la prueba SRMR robusta en términos de las tasas de error de Tipo I y fue menos afectada por el tamaño del modelo.

El segundo estudio investigó la precisión de la evaluación del ajuste relativo del modelo utilizando varias versiones de pruebas de diferencia chi-cuadrado que son robustas a las violaciones de la normalidad. Utilizando simulaciones, el estudio manipuló los niveles de (no)normalidad, el tamaño de la muestra, el tamaño del modelo y los grados de libertad. Los resultados mostraron que la prueba de diferencia chi-cuadrado ajustada por media y varianza fue precisa en todas las condiciones investigadas y superó a sus competidores ajustados por media, los cuales requirieron muestras más grandes para funcionar adecuadamente.

En resumen, los dos estudios en la tesis doctoral contribuyen a la literatura tanto sobre el ajuste absoluto como sobre el ajuste relativo en modelos SEM. Los hallazgos sugieren que la prueba chi-cuadrado robusta es más precisa en la evaluación del ajuste exacto del modelo que la prueba SRMR robusta, y que la prueba de diferencia chi-cuadrado ajustada por media y varianza es un método confiable para evaluar el ajuste relativo de modelos SEM.

GENERAL INTRODUCTION

Structural equation modeling (SEM) refers to a general approach to estimate systems of equations, generally involving latent variables to account for measurement error. Popular SEM software programs include AMOS (Arbuckle, 2014), EQS (Bentler, 2004), LISREL (Jöreskog & Sörbom, 2017), Mplus (Muthén & Muthén, 2017) and the Lavaan package (Rosseel, 2012) in R (R Core Team, 2019). Many models can be subsumed as special cases of the general SEM framework, including regression, instrumental variables regression, models for experimental data, factor analysis, path analysis, random effects models for panel data (Bollen & Curran, 2006), etc. The generality of the modeling approach and the availability of user-friendly software has made SEM of the most widely used data modeling techniques across behavioral, social, medical, and management sciences.

The SEM modeling and estimation toolkit was originally developed for models involving continuous outcomes and with no mean structure (i.e., covariance structure modeling), and for estimation under normality assumptions. However, over the years, the SEM toolkit has considerably expanded to include standard errors and tests of model fit robust to non-normality (Satorra & Bentler, 1994), models with structured means (Browne & Arminger, 1995), multiple-population models (Muthén, 1989), methods for missing data (Arbuckle, 1996), models with discrete outcomes (Muthén, 1984) and item response theory (IRT) models (Embretson & Reise, 2000), models for clustered data (e.g., multilevel models: Hox, Moerbeek & Van de Schoot, 2017; Skrondal & Rabe-Hesketh, 2004), etc. When the observed outcomes are continuous, maximum likelihood (ML) has emerged as the estimator of choice (Maydeu-Olivares, 2017b); when the outcomes are discrete, multi-stage estimators involving the computation of polychoric and polyserial correlations are more popular (Finney & DiStefano, 2013) due to superior computational efficiency (Forero & Maydeu-Olivares, 2009).

Because of its generality and flexibility, SEM is often used for testing theoretical propositions (e.g., theoretical models). As a result, a substantial body of literature to date has focused on model goodness of fit and model selection in SEM.

Goodness of Fit in SEM

Given a theoretical model, researchers are interested in evaluating its goodness of fit. The goodness of fit of a statistical model refers to the extent to which a statistical model captures the data generating mechanism. Thus, goodness of fit refers to assessing the discrepancy between the proposed model and the data, and it is often referred to as model-data fit, or absolute model fit. These terms will be used interchangeably in this work.

However, in applications, researchers oftentimes propose several competing models (i.e., hypotheses) that are all plausible based on the available theory. In such cases, it is not only important to assess how well each of these models fits the observed data (i.e., models' absolute fit or model-data fit) but also to select the best fitting model among the competing models under consideration. Model selection refers to assessing the discrepancy between two models, and it is often referred to as relative model fit, or model comparison. These terms will be used interchangeably in this work.

Two different perspectives exist regarding goodness of fit and model selection. The first perspective can be labeled "*models as approximations*". Within this perspective, all models are simply approximations to real life phenomena (i.e., all models are wrong) and the researcher's aim is to choose the model that provides the "best" approximation to the data generating process (Cudeck & Henly, 1991; MacCallum, 2003). From this perspective, it makes more sense to focus on model selection and take model parsimony into account. The second perspective can be labeled "*models as structural relations*". Within this perspective, modeling involves estimating structural or "causal" relations even when data is observational (Bollen & Pearl, 2013; Pearl, 2009). From this perspective, it makes more sense to focus on model fit because inferences are to be as precise and valid as possible.

It is important to realize the "*models as structural relations*" is nested within the "*models as approximations*" perspective. In the models as approximations perspective, a discrepancy between the unknown data generating model and the fitted model is chosen, for instance the Root Mean Square Error of Approximation (RMSEA: Browne & Cudeck, 1993; Steiger & Lind, 1980), and a model is retained if the RMSEA is "sufficiently small" (5% is usually the chosen cut-off value). In contrast, the models as structural relations framework amounts to testing whether the RMSEA is zero, that is, to testing whether the fitted model is the data generating model. Similarly, within the models as approximations framework, a model nested within a model with more parameters is selected when the difference in RMSEA is "small" (MacCallum, Browne, & Cai, 2006), that is, when the difference in fit is "sufficiently small". In contrast, from the models as structural relations perspective, a nested model is selected when the difference in RMSEA is zero, i.e., when both models yield the same fit to the data. The table below summarizes the relationship between tests of exact fit (*models as structural relations*) and the tests of close fit (*models as approximations*).

This doctoral thesis focuses on exact fit, that is, on means of evaluating if a model fits the data exactly, or whether two models are statistically indistinguishable from each other. Because tests of exact fit are more stringent than tests of close fit, they are of interest particularly in models with few observed variables or few degrees of freedom. In such situations, tests of close fit can be misleading (see e.g., Kenny et al., 2015). On the other hand, testing for zero difference in fit has been used extensively, particularly within the measurement invariance literature (e.g., Guhn et al., 2018; Hawes et al., 2018; Huhtala et al., 2018; Jenkins et al., 2018), but also in a variety of other applications involving several theoretically equally justified hypotheses (e.g., Elkins et al., 2018; Lai et al., 2015; Pappu, & Quester, 2016; Schivinski & Dabrowski, 2016; Shams et al., 2017).

Assessing exact model fit: The likelihood ratio (chi-square) test statistic

In classical SEM (i.e., in covariance structure analysis), the maximum likelihood (ML) model estimation procedure involves minimizing

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

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*, *p* is the number of observed variables, and *tr* is the trace of the matrix.

When ML estimation is used for model estimation, model fit can be (and most commonly is) statistically evaluated relying on the likelihood ratio (LR) test (Jöreskog, 1969). This is because the LR test statistic is the product of the minimum of the fit function in Equation (1) multiplied by the sample size:

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

where *N* denotes sample size, and $\hat{F}_{ML}$ is the minimum of the fit function in Equation (1). If the model is correctly specified and data is multivariate normal, the *T* statistic asymptotically follows a central chi-square distribution with degrees of freedom (*df*) equal to the number of free sample variances and covariances minus the number of parameters estimated (*p(p+1)/2-q*), hence providing statistical basis for evaluating the overall model fit. Specifically, *T* statistic can be conveniently pitted against the reference sampling distribution to obtain a *p*-value. The LR test evaluates if the model fits exactly and may be considered the only substantive test of fit for SEM (see e.g., Barrett, 2007). Given that its reference distribution is chi-square, *T* statistic is commonly referred to as the "chi-square" test statistic in the applied literature and the corresponding goodness of fit test, a "chi-square test" (χ^2).

The LR test can also be used to evaluate a relative fit of two competing models, given that the two models are nested (i.e., hierarchical). The most common type of model nesting is parameter nesting (Bentler & Bonett, 1980). A model (let's denote it with M_0) is nested within another model if the covariance structures implied by it can be reproduced exactly by fitting the other model (let's denote it with M_1). The LR difference test assesses the null hypothesis that the model with fewer estimated parameters fits no worse than a model with more parameters. Under the normality assumptions, and for ML estimation, the difference in fit between two nested models can be tested simply by subtracting the two LR absolute 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 et al., 1985). Given that its reference distribution is chi-square, the D statistic is commonly referred to as the "chi-square difference" test statistic in the applied literature and the corresponding test a "chi-square difference test" ($\Delta\chi^2$).

It is important to observe that in this set-up, the LR test for the absolute model fit and the LR difference test for relative model fit are directly related. Specifically, the former may be considered simply a special case of a latter because the LR test of overall model fit evaluates the null hypothesis that the proposed model (M_0) fits no worse than a saturated model (M_1), for which $T_1 = 0$ and $df_1 = 0$, hence $D = T_0$.

Factors affecting performance of chi-square tests

The multivariate normality assumption

The critical assumption underlying the chi-square test is that data are multivariate normal (normal theory – NT). If data are not normal, the reference sampling distribution of the test statistic may no longer be chi-square and, consequently, the accuracy of the p-values may deteriorate. Given that the assumption of multivariate normality is often untenable in empirical research (see e.g., Cain et al., 2017; Micceri, 1989), considerable research efforts have been directed towards estimating the effects of violations of normality on the performance of the chi-square test. Overall, it has been well documented in the literature to date that chi-square test rejects the null hypothesis (i.e., the correctly specified model) too often (that is, above the specified alpha level), even under relatively small violations of the normality assumption (e.g., Fouladi, 2000; Hu et al., 1992; Satorra, 1990; Satorra & Bentler, 1994).

When data are not normal, *D* statistic in Equation (3) may also not be chi-square distributed (Satorra, 2000). Like the chi-square test for overall fit, it has been convincingly shown in simulation research that the chi-square difference test also results in inflated Type I error rates (i.e., overrejection) when data are non-normal (e.g., Brace & Savalei, 2017; Chuang et al., 2015). Put differently, under violations of the normality assumption, the chi-square difference test will tend to favor the more complex model under investigation too often when, in fact, the two compared models are indistinguishable in terms of fit.

Sample size

The statistical theory underpinning the chi-square test is asymptotic, meaning that the reference sampling distribution of the statistic is known to be chi-square in very large samples. However, for a variety of reasons, researchers often rely on small sample sizes. For instance, Jiang and Yuan (2017) report that among publications that have used SEM methodology between 2010 and 2016, sample sizes are less than 283 in half of the studies, less than 164 in a quarter of the studies, and less than 100 in 11% of the studies.

Given that small samples are a reality in research, the performance of the chi-square test in 'smaller' samples has been intensively investigated in simulations including the minimum sample size requirements for the statistical results to be valid. In general, key simulation findings indicate that at small sample sizes, the chi-square test's Type I error rates may become inflated even under ideal conditions, that is, when normality assumption holds and model is correctly specified (see e.g., Nevitt & Hancock, 2004), and that inflation in Type

I error rates is exacerbated further with introduced data nonnormality. Minimal sample size requirements for the chi-square test to be accurate vary depending on several factors such as model size, missing data, reliability of the variables, and strength of the relations among the variables. In a given tentative application, the influence of these factors can be estimated by simulation (e.g., Muthén & Muthén, 2002) to determine minimum sample size requirements.

Given that the chi-square difference statistic is also chi-square distributed asymptotically, its performance when comparing difference in fit between two nested models may also be compromised in small samples. The research on this issue has been scarce and somewhat mixed. For instance, Chuang and colleagues (2015) found that chi-square difference test might tend to slightly overreject in small samples even if the normality assumption holds. On the other hand, Brace and Savalei (2017) reported accurate Type I rates and high power of the chi-square difference test in small samples under normality. Under violations of normality, both studies reported that larger sample sizes did not help rectify the problem of inflated Type I error rates of the chi-square difference test.

Model size

Model size has been operationally defined in the literature in several ways, that is, as the number of observed variables (i.e., the size of the covariance matrix), the number of free parameters being estimated, or a combination of these, such as the model degrees of freedom (*df*). It has been repeatedly shown in simulations that increasing the size of a model is associated with a decreasing accuracy of the chi-square *p*-values. It is still unclear why this happens, but essentially, the problem is that the asymptotic chi-square approximation of the test statistic is not good enough when models are large.

In general, it has been shown in simulations that the chi-square test tends to overreject correctly specified models (i.e., has the inflated Type I error rates) with increasing model size (e.g., Herzog et al., 2007; Hoogland & Boomsma, 1998; Jackson, 2003; Kenny & McCoach, 2003). More recent research efforts aimed at identifying the primary source of the suboptimal performance of the chi-square test with respect to the model size (e.g., Moshagen, 2012; Shi et al., 2018). In the first systematic investigation of this kind, Moshagen (2012) reported that the primary contributor to the overrejection effect is the number of observed variables, while neither the number of estimated model parameters nor *df* affect the chi-square test accuracy when the number of variables is held constant. In a follow-up to this study, Shi and colleagues (2018) confirmed that the number of variables is the primary source of the overrejection. In addition, the authors also reported that the number of estimated parameters has its unique contribution to the accuracy of the test, such that increasing the number of estimated parameters somewhat rectifies the overrejection problem.

Model size may also be discussed in the context of testing for difference in fit of two nested models. Like in the case of testing for overall model fit, model size may also be operationalized in terms of the number of observed variables of the two models, yet it is important to note that the size of the two covariance matrices will always be the same given that the models are nested. In addition, because fit of two nested models is being compared, it seems more reasonable to discuss model size in this set-up in terms of the relative model size, that is, the difference in the number of estimated parameters (the *df* for the difference test). Simulation studies investigating performance of the chi-square difference statistic has traditionally relied on the two abovementioned definitions of model size when designing their simulation conditions (e.g., Brace & Savalei, 2017; Chuang et al., 2015). The results of these studies do suggest that model size may have an affect the accuracy of the chi-square difference test. Specifically, it appears that increasing the difference in the estimated model parameters may exacerbate the effects of non-normality on Type I error inflation rates (Chuang et al., 2015). In addition, increasing the number of observed variables may also lead to increasing overrejection rates, especially in small samples (Brace & Savalei, 2017).

Potential solutions under the violation of normality assumption

In the literature to date, several approaches have been suggested to overcome problems associated with chi-square tests when data are not normal. The first approach involves developing corrected chi-square tests that are robust to violations of normality. The second approach involves resampling methods which do not assume normally distributed data. Finally, the third approach involves development of alternative tests robust to nonnormality.

Regarding the first approach, several corrected chi-square tests, robust to violations of normality assumption, have been proposed. These corrections can be either mean adjustments (modify test statistic so that it agrees in mean with the chi-square reference distribution) or mean and variance adjustments so that it agrees in both mean and variance (Satorra & Bentler, 1994). Specifically, the test statistic can be corrected so that in large samples it agrees in mean with the 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). Traditionally, mean adjustments have been more popular among substantive researchers primarily because mean and variance adjustments have only recently been implemented into popular statistical software packages (e.g., Mplus, R). However, in large samples, the mean and variance corrected chi-square test should be superior to the less computationally expensive mean corrected chi-square (Asparouhov & Muthén, 2013), which has been also confirmed in simulations (e.g., Maydeu-Olivares, 2017b).

If the normality assumption is violated, the difference statistic *D,* in Equation (3) will not be chi-square distributed even if it is computed based on the two corrected (i.e., robust) LR-based statistics (Satorra, 2000). Several corrected chi-square difference tests 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 the Satorra-Bentler mean-adjusted chi-square difference tests (Satorra & Bentler, 2001; Satorra & Bentler, 2010). The results of simulation research (Brace & Savalei, 2017; Chuang et al., 2015) provided limited support for the robustness of the Satorra and Bentler (2001) and (2010) corrections gently favoring the more recently proposed one. Even though the mean and variance adjusted chi-square difference statistic (Asparouhov & Muthén, 2006, 2010) should perform better than the mean corrected difference options and are currently implemented in some of the most popular statistical software packages (e.g., R, Mplus), its performance has not yet been thoroughly evaluated in simulations.

Regarding the second approach, the most used resampling method for evaluating overall model fit is the Bollen and Stine (1992) model-based solution to bootstrap *p*-values of the chi-square tests statistic (Bollen & Stine, 1992; Yuan et al., 2007). Although the Bollen-Stine approach has been implemented in several widely used SEM packages, only a few studies to date have been investigating its performance. Overall, these studies reported higher accuracy of the method over the mean corrected versions of the chi-square test (e.g., Grønneberg & Foldnes, 2019; Nevitt & Hancock, 2001). The Bollen-Stine approach may also be used for comparing nested models, although its performance has not been systematically investigated in simulations (see e.g., Grønneberg & Foldnes, 2019).

Regarding the third approach, several tests appropriate for non-normal data have been proposed as alternatives to the chi-square test. For instance, Wu and Lin (2016) developed a scaled *F* test by matching 3 moments (i.e., the mean, variance, and skewness) simultaneously. Maydeu-Olivares (2017a) provided statistical theory for utilizing the standardized root mean square residual (SRMR: Browne & Cudeck, 1993) as a test of exact fit, both under normality assumptions and when data is not normal. The SRMR can be used to assess exact fit by simply setting the value of the approximate fit to 0 and using a normal reference distribution approximation to obtain *p*-values. More recently, Hayakawa (2019) proposed a goodness of fit test and its scaled version based on reweighted least squares (RLS), originally proposed by Browne (1974). RLS relies on the ML discrepancy function to obtain ML estimates, which are then plugged into the generalized least squares (GLS) loss function to calculate the test statistic (T_{RLS}). Although T_{RLS} is based on the ML estimator, it does not rely on the conventional discrepancy function in Equation (1) to compute the test statistic. Although these newly proposed alternatives to the scaled chi-square tests show promise, additional research is required to ascertain their performance under violations of normality and potential advantages to the current standard. In addition, at the time of this writing, these tests for assessing overall model fit do not include statistical theory for their comparative fit counterparts.

OBJECTIVE AND MOTIVATION

The objective of this doctoral thesis is twofold, and it aims at contributing to both exact model fit and model comparison literature. Regarding the former, as discussed above, the chi-square test has been repeatedly found to underperform when the normality assumption is not met. Although the proposed adjustments to the chi-square statistic clearly outperform the unadjusted version, their performance remains negatively affected under some conditions. On the other hand, the robust version of the SRMR may outperform the robust chi-square tests and solve the problem of assessing the exact fit under nonnormal data. Although the initial simulation test of the performance of the SRMR statistics showed promise, it included only normal data and small models (Maydeu-Olivares, 2017a). Thus, evaluation of the robust version of SRMR for assessing exact model fit under conditions of nonnormality remains an outstanding need in the literature. Accordingly, the first objective of this thesis is to: 1) replicate the findings on the performance of chi-square tests under conditions of nonnormality and other suboptimal conditions such as, small samples and large models, and 2) to pit the performance of these tests against the performance of the SRMR.

Regarding the latter, when comparing fit of two nested models using the chi-square difference test, previous simulation research has been primarily focused on investigating the performance of the unadjusted test and potential benefits of the mean adjustments. Although it has been clearly shown that the mean adjusted chi-square difference tests perform considerably better under nonnormality and other suboptimal conditions, their performance remained inadequate under circumstances. On the other hand, while the mean and variance adjustments work better and they do when assessing absolute fit, their performance when assessing comparative fit has not been systematically explored to date. Accordingly, the second objective of this thesis is to: 1) replicate findings on the performance of unadjusted and mean adjusted chi-square difference tests under nonnormality and other suboptimal conditions, and 2) to pit the performance of these statistics to the mean and variance adjusted chi-square difference tests.

Thesis structure

This doctoral thesis is a compilation of two published studies. In the remainder of this document, I present each of these studies in the following sections below. Specifically, in order to preserve the organizing logic of the thesis document, I first present the study in the exact model fit followed by the study on model comparison. Presentation of the two studies is then followed by a general summary and discussion of main findings. At the end of this document, I discuss the main limitations of the current work, provide some viable suggestions for further research, and end with brief concluding remarks.

Publications included in the thesis

STUDY I

Pavlov, G., Maydeu-Olivares, A., & Shi, D. (2021). Using the standardized root mean squared residual (SRMR) to assess exact fit in structural equation models. *Educational and Psychological Measurement*, *81*(1), 110-130 https://doi.org/10.1177/0013164420926231

STUDY II

Pavlov, G., Shi, D., & Maydeu-Olivares, A. (2020). Chi-square difference tests for comparing nested models: An evaluation with non-normal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 908-917. https://doi.org/10.1080/10705511.2020.1717957

STUDY I: USING THE STANDARDIZED ROOT MEAN SQUARED RESIDUAL (SRMR) TO ASSESS EXACT FIT IN STRUCTURAL EQUATION MODELS

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 (2017a) to assess the exact fit of SEM models. In a simulation study, we found that under normality, the mean and variance 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 is 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.

Introduction

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 regards 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, Shi, & Fairchild, 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

$H_0^*: RMSEA \leq RMSEA_0$ vs. $H_1^*: RMSEA > RMSEA_0$ where $RMSEA_0$ is an arbitrary population value of the RMSEA. When data is normally distributed, a p-value for a test of close fit can be obtained using

$$1 - F_{\chi^2}\left(X^2; df, N \times df \times RMSEA_0^2\right), \tag{4}$$

where N denotes sample size, $F_{\chi^2}(\cdot; df, \lambda)$ denotes the non-central chi-square distribution with df degrees of freedom and non-centrality 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 (4) can also be used to assess the exact fit of the model, i.e., $RMSEA_0 = 0$. In this case, the non-centrality parameter $N \times df \times RMSEA_0^2$ becomes zero, and (4) 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 non-normality is used, (4) is replaced by

$$1 - F_{\chi^2}\left(X^2; df, N \times df \times RMSEA_0^2 / c\right), \tag{5}$$

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, (5) reduces to the usual chi-square testing in the special case of examining exact fit, e.g., $H_0^*: RMSEA = 0$.

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 is 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 8 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 et al., 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 non-normal 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 non-normal 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 (SRMR)

The sample SRMR

Let the standardized residual variances and covariances be

$$\hat{\varepsilon}_{ij} = \frac{s_{ij} - \hat{\sigma}_{ij}}{\sqrt{s_{ii} s_{jj}}}, \tag{6}$$

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

$$\widehat{SRMSR} = \sqrt{\frac{1}{t}\sum_{i \le j} \hat{\varepsilon}_{ij}^2} = \sqrt{\frac{1}{t}\hat{\varepsilon}'\hat{\varepsilon}}, \tag{7}$$

where $t = p(p+1)/2$ denotes the number of non-redundant variances and covariances, and $\hat{\varepsilon}$ denotes the vector of *t* standardized residual covariances (6).

Equation (7) 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:

$$\widehat{SRMSR^*} = \sqrt{\frac{1}{t+p}\left(\sum_{i \le j} (\hat{\varepsilon}_{ij}^*)^2 + \sum_{i} (\hat{\varepsilon}_{i}^*)^2\right)}. \tag{8}$$

$$\hat{\varepsilon}_{ij}^{*} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}, \qquad \hat{\varepsilon}_{i}^{*} = \frac{m_i}{\sqrt{s_{ii}}} - \frac{\hat{\mu}_i}{\sqrt{\hat{\sigma}_{ii}}}, \tag{9}$$

where m_i and $\hat{\mu}_i$ denote the sample and expected mean of variable i. It needs to be noted that in (9), $\hat{\sigma}_{ii}$ is used when standardizing the expected covariance. In contrast, in (6), the unrestricted estimate s_{ii} is used for standardization. This need not impact considerably the SRMR values because, in many applications, the estimated variances equal the sample variances, i.e., $s_{ii} = \hat{\sigma}_{ii}$. However, the inclusion of the mean structure components $\hat{\varepsilon}_{i}^{*}$ in the Mplus version of the SRMR statistic may have a non-negligible 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 (8) will result in a *lower value* than the value computed using the SRMR in (7). 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 (7).

Confidence intervals for the population SRMR

The sample SRMR provided in equation (7) is an estimator of the population SRMR:

$$SRMR = \sqrt{\frac{1}{t}\sum_{i \le j} \varepsilon_{ij}^{2}} = \sqrt{\frac{1}{t}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}, \qquad \varepsilon_{ij} = \frac{\sigma_{ij} - \sigma_{ij}^{0}}{\sqrt{\sigma_{ii}\sigma_{jj}}}. \tag{10}$$

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.

[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 (7) (Asparouhov & Muthén, 2018).

The sample SRMR provided in (7), however, is a 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 and colleagues (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 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 and colleagues (Table 2: 2018) 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

$$\widehat{SRMR_u} = \hat{k}^{-1}\sqrt{\frac{\max\left(\hat{\varepsilon}'\hat{\varepsilon} - \mathrm{tr}(\hat{\Xi}),0\right)}{t}}\,, \qquad \hat{k} = 1 - \frac{\mathrm{tr}(\hat{\Xi}^2) + 2\hat{\varepsilon}'\hat{\Xi}\hat{\varepsilon}}{4\left(\hat{\varepsilon}'\hat{\varepsilon}\right)^2}\,, \tag{11}$$

where Ξ denotes the asymptotic covariance matrix of the sample standardized residuals (6), 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).

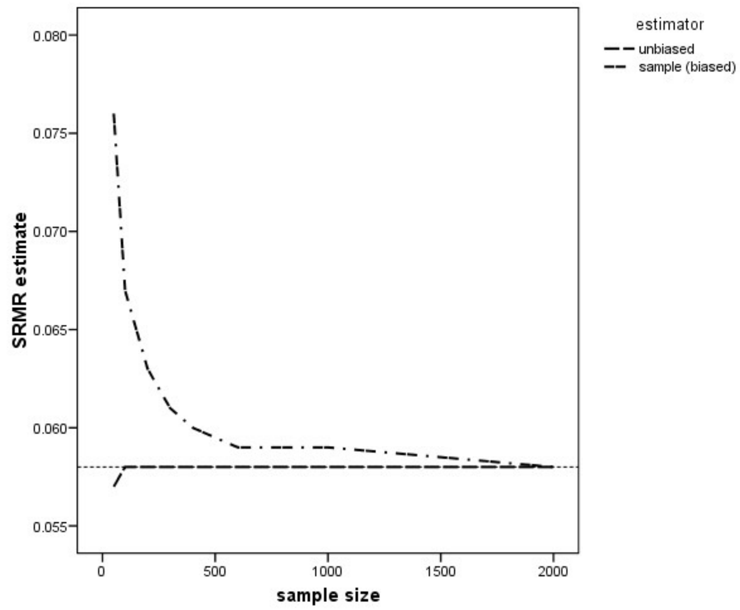

Figure 1. Average sample (i.e., biased) SRMR and unbiased SRMR estimates of the population SRMR of .058 across 1,000 replications as a function of sample size.

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

$$\Pr\left(\widehat{SRMR_u} - z_{\alpha/2}SE(\widehat{SRMR_u}) \le SRMR \le \widehat{SRMR_u} + z_{\alpha/2}SE(\widehat{SRMR_u})\right) = 1 - \alpha\,, \qquad (12)$$

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

$$SE(\widehat{SRMR_u}) = \sqrt{k^{-2}\,\frac{\mathrm{tr}(\hat{\Xi}^2) + 2\hat{\varepsilon}'\hat{\Xi}\hat{\varepsilon}}{2t\;\hat{\varepsilon}'\hat{\varepsilon}}}\,. \qquad (13)$$

Finally, p-values for a null hypothesis of close fit, $H_0 : SRMR \leq SRMR_0$ vs. $H_1 : SRMR > SRMR_0$, where SRMR$_0$ denotes an arbitrary value of the population SRMR, can be obtained using

$$p = 1 - \Phi\left(\frac{\widehat{SRMR_u} - SRMR_0}{\mathrm{SE}(\widehat{SRMR_u})} \right), \tag{14}$$

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 $\hat{\varepsilon}'\hat{\varepsilon} - \mathrm{tr}\left(\hat{\Xi}\right) < 0$, and the unbiased SRMR estimate is set to zero; see equation (11). Put differently, when the model fits exactly, the sampling distribution of the $\widehat{SRMR_u}$ must be zero inflated and a normal distribution must provide a poor approximation. See Figure 1 of Shi and colleagues (2019) 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: $H_0 : \Sigma = \Sigma_0$ vs. $H_1 : \Sigma \neq \Sigma_0$, 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, 2019), 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 (7) can be approximated in large samples using

$$\mu_{\widehat{SRMR}} = \sqrt{\frac{\mathrm{tr}(\Xi)}{t}}\left(1 - \frac{\mathrm{tr}(\Xi^2)}{4\,\mathrm{tr}(\Xi)^2}\right), \tag{15}$$

$$\sigma_{\widehat{SRMR}} = \sqrt{\frac{\mathrm{tr}(\Xi^2)}{2t\,\mathrm{tr}(\Xi)}}\,. \tag{16}$$

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

$$p = 1 - \Phi(z), \qquad z = \frac{\widehat{SRMR} - \mu_{\widehat{SRMR}}}{\sigma_{\widehat{SRMR}}}. \tag{17}$$

To investigate the performance of the method above, Maydeu-Olivares (2017a) performed a simulation study involving a CFA model with 8 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 et al., 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 is 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 non-normal 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 were a confirmatory factor analysis (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 non-normality 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 7 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 non-normally 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 non-normality.

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, et al., 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 non-normality. Three levels of non-normality 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, and colleagues (2007).

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

Kurt	**Skew**	**Thresholds**	**Expected Area Under the Curve**						
			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%

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 × 3 × 3). For each of the 36 simulated conditions, one thousand (1,000) replications were generated with the *simsem* package in R (Pornprasertmanit et al., 2013; R Core Team, 2019).

Estimation

For each simulated dataset, we fitted a one-factor CFA model with the Maximum Likelihood (ML) estimation method using the *lavaan* package in R (Rosseel, 2012). The SRMR test statistic (17) was obtained under both normality (NT) and asymptotically distribution free (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 (NT in the table) overrejected the true model when data were non-normal. Furthermore, the rejection rates increased as the model size increased. For the non-normal 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 non-normality. 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 showed 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.

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

Kur.	Skew.	*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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	1.00	1.00	0.20	0.00

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

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 non-normal 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 non-normality were considered, and the SRMR was computed under both normal theory (NT) and asymptotic distribution free (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 is 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 is not normal.

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

$$T_a = a + bT \, , \tag{18}$$

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 $X_a^2 = a + bX^2$, where X_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 non-normal 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 non-normal 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, $a = -\mu_{\widehat{SRMR}}\sigma_{\widehat{SRMR}}^{-1}$, $b = \sigma_{\widehat{SRMR}}^{-1}$ and this is the solution proposed in (17). 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 non-normal 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 (15) and (16) 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 non-normal 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 non-normal conditions, the mean and variance corrected $\widehat{SRMR}$ statistic provides inaccurate *p*-values.

Table 3. Accuracy of the Asymptotic Approximation to the Sampling Distribution of the Sample SRMR Across 1,000 Replications. Test of Normality, Observed vs. Expected Mean and Standard Deviation

				SRMR							
				Observed		Expected (NT)		Expected (ADF)		Test of normality	
Kur.	Skew.	*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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	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
			1000	0.026	.0007	0.019	.0004	0.026	.0011	.9978	.20

Note: *p* = number of variables; *N* = sample size; M = mean; SD = standard deviation; NT = under normality; ADF = asymptotically distribution free. SW = Shapiro-Wilk test statistic.

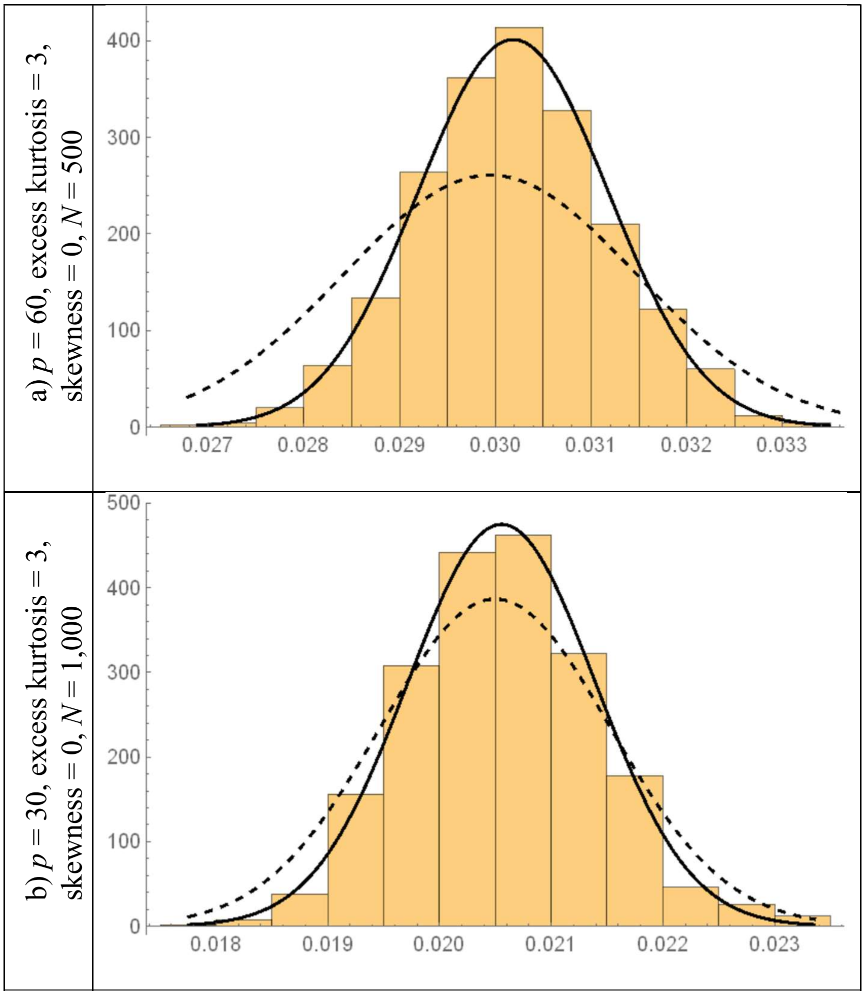

Figure 2. Empirical distribution of the sample SRMR across 1,000 replications and reference normal distributions. The solid and dotted lines are obtained using the empirical and asymptotic mean and standard deviations, respectively.

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 $\widehat{SRMR}$ 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-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 our 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 normal 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 suggests 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.

References

Asparouhov, T., & Muthén, B. (2010). *Simple second order chi-square correction scaled chi-square statistics* (Technical appendix). Los Angeles, CA.

Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus*. Los Angeles, CA. Retrieved from www.statmodel.com/download/SRMR2.pd

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]. Encino, CA: Multivariate Software Inc.

Bentler, P. M. (2004). EQS 6 [Computer Program]. Encino, CA: 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). Dordrecht: Springer Netherlands. 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. Retrieved from 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, UK: 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). Newbury Park, CA: Sage.

DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling: A Multidisciplinary Journal*, *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: A Multidisciplinary Journal*, *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: A Multidisciplinary Journal*, *27*(2), 192–201. https://doi.org/10.1080/10705511.2019.1637741

Hayakawa, K. (2019). 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: A Multidisciplinary Journal*, *14*(3), 361–390. https://doi.org/10.1080/10705510701301602

Hu, Li-tze, & 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, Li-tze, 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

Hu, Li-tze, & 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

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

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

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

Lai, K. (2019). A simple analytic confidence interval for CFI given nonnormal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *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). London: 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: A Multidisciplinary Journal*, *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: A Multidisciplinary Journal*, *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. Los Angeles, CA: 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. Vienna, Austria, Austria: 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). Newbury Park, CA: 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). Thousand Oaks, CA: 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), 419–429. 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: A Multidisciplinary Journal*, *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: A Multidisciplinary Journal*, *27*(1), 1–15. https://doi.org/10.1080/10705511.2019.1611434

Steiger, J. H. (1989). EzPATH: A supplementary module for SYSTAT and SYGRAPH. Evanston, IL: 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). Statistically-based tests for the number of common factors. Iowa: Paper presented at the Annual Meeting of the Psychometric Society.

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.

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

Supplementary materials

```
###########################################################################
#This R function conducts the exact fit test using SRMR#
###########################################################################

#install and load the lavaan package
install.packages("lavaan")
library(lavaan)

#Set the working dictionary and Read the Data
setwd("C:/EXACTFITSRMR")
Data=read.csv("DATA.csv",header = F,sep = ",")

# Fitting the SEM model using lavaan
cfa.model <- ' f  =~ NA*V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10
f~~1*f'
fit <- cfa(cfa.model, data = Data, estimator = "ML",meanstructure =
FALSE,fixed.x = FALSE)

#Read the code/function for computation, put "lav_fit_usrmr_ucrmr.R" in the
same working dictionary
source("lav_fit_usrmr_ucrmr.R")

# Obtain Normal-Theory Results
lav_fit_usrmr_ucrmr (lavobject = fit, ADF = FALSE)

# Obtain ADF Results
lav_fit_usrmr_ucrmr (lavobject = fit, ADF = TRUE)

###########################################################################
# Notes #
#This function works for models with a single group, no mean structure, and
continuous outcomes #
#listwise deletion should be applied when there is missing data #
###########################################################################
```

STUDY II: CHI-SQUARE DIFFERENCE TESTS FOR COMPARING NESTED MODELS: AN EVALUATION WITH NON-NORMAL DATA

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.

Introduction

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 et al., 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 et al., 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 to 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 et al., 2018; Hawes et al., 2018; Huhtala et al., 2018; Jenkins et al, 2018; Krieg et al., 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 and colleagues (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) the mean and variance adjusted difference chi-square to the 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 the chi-square difference statistic

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:

$$F_{ML}(\mathbf{S},\Sigma(\theta)) = \log|\Sigma(\theta)| - \log|\mathbf{S}| + tr(\mathbf{S}\Sigma^{-1}(\theta)) - p, \tag{19}$$

where $\mathbf{S}$ is the sample covariance matrix, $\Sigma(\theta)$ is the model implied covariance matrix, θ 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,

$$T = (N-1)F_{ML}, \tag{20}$$

where N denotes sample size, and $\hat{F}_{ML}$ is obtained by minimizing the ML fit function with respect to θ. 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 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) to be used with complete data. In 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 may be 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:

$$D = T_0 - T_1, \tag{21}$$

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 et al, 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

$$\bar{D}_{01} = \frac{D}{c_{01}}, \qquad c_{01} = \frac{df_0 c_0 - df_1 c_1}{df_0 - df_1}, \tag{22}$$

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 (22) can take a negative value leading to a negative estimate of the test statistic. To avoid this shortcoming of the scaling correction in (22), 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 (22) except that c_1 in (22) is replaced by $c^* = \frac{T^*}{\bar{T}^*}$, where T^*, $\bar{T}^*$ 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, and it is always 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 as 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 as DSB10$_{\text{MLR}}$. We note that what we refer to in this paper as DSB10$_{\text{MLM}}$ corresponds to the difference statistic D$_{\text{R2}}$ evaluated by Chuang and colleagues (2015), and to the D$_{\text{SB10}}$ statistic evaluated by Brace and Savalei (2017).

Of focal interest in the current study is the second order (i.e., the mean and variance) adjusted difference statistics (Asparouhov & 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. To match the empirical mean and variance of the difference statistic with those of a chi-square distribution, a and b need to meet $E(\overline{D}) = df$ and $Var(\overline{D}) = 2df$. The second order adjustment (Asparouhov & Muthén, 2010) is given by

$$\overline{\overline{D}} = \sqrt{\frac{df}{tr(\mathbf{M}^2)}}D + df - \sqrt{\frac{df\,tr(\mathbf{M})^2}{tr(\mathbf{M}^2)}}, \tag{23}$$

where $\mathbf{M}$ is given in formula (9) in Asparouhov and Muthén (2006). We refer to the difference statistic in (23) as D$_{\text{MLMV}}$. In Table 1, we summarize the choices of statistics available to substantive researchers to test differences in fit between nested models.

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).

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$_{\mathrm{MLM}}$ to underreject and DSB10$_{\mathrm{MLM}}$ 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$_{\mathrm{MLM}}$ and DSB10$_{\mathrm{MLM}}$ 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$_{\mathrm{MLM}}$ were accurate except for a few conditions involving the smallest sample sizes (N = 220). The behavior of DSB1$_{\mathrm{MLM}}$ 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., $\mathrm{DSB1}_{\mathrm{MLR}}$ and $\mathrm{DSB10}_{\mathrm{MLR}}$). 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: $\mathrm{DSB1}_{\mathrm{MLM}}$, $\mathrm{DSB1}_{\mathrm{MLR}}$, $\mathrm{DSB10}_{\mathrm{MLM}}$, $\mathrm{DSB10}_{\mathrm{MLR}}$, and $\mathrm{D}_{\mathrm{MLMV}}$. 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 et al., 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 points was set to 0.15. Finally, the error variances were set such that the population variances of the observed variables were equal to one.

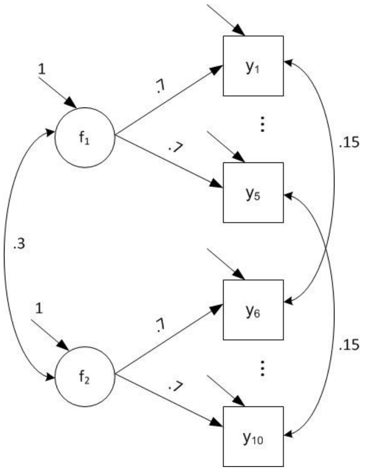

Figure 1. Small model used in the simulations.

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 were 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 et al., 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 × 3 × 4 × 2 × 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$_{\text{MLM}}$, DSB1$_{\text{MLR}}$, DSB10$_{\text{MLM}}$, DSB10$_{\text{MLR}}$, and D$_{\text{MLMV}}$) 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}, $\text{DSB10}_{\text{MLR}}$, 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 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., $SB10_{MLR}$).

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).

Table 2. Correctly Specified Small Model (p = 10). Type I Error Rates at the 5% Significance Level

Distribution			*df = 1*					
Kurt	Skew	N	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
0.0	0.0	200	0.03	0.03	0.03	0.03	0.03	0.03
0.0	0.0	500	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
7.0	2.0	100	0.28	0.06	0.12	0.07	0.08	0.07
7.0	2.0	200	0.24	0.06	0.09	0.06	0.07	0.06
7.0	2.0	500	0.27	0.05	0.07	0.06	0.06	0.06
7.0	2.0	1,000	0.25	0.06	0.06	0.06	0.06	0.06
10.0	2.0	100	0.25	0.06	0.13	0.07	0.10	0.07
10.0	2.0	200	0.27	0.05	0.09	0.06	0.08	0.06
10.0	2.0	500	0.29	0.05	0.08	0.05	0.07	0.05
10.0	2.0	1,000	0.31	0.04	0.06	0.05	0.05	0.04

Distribution			*df = 5*					
Kurt	Skew	N	D	DSB1$_{MLM}$	DSB1$_{MLR}$	DSB10$_{MLM}$	DSB10$_{MLR}$	D$_{MLMV}$
0.0	0.0	100	0.03	0.03	0.03	0.03	0.02	0.03
0.0	0.0	200	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
0.0	0.0	1,000	0.02	0.02	0.02	0.02	0.02	0.02
7.0	2.0	100	0.27	0.08	0.14	0.10	0.12	0.06
7.0	2.0	200	0.26	0.05	0.09	0.07	0.09	0.04
7.0	2.0	500	0.27	0.06	0.08	0.07	0.08	0.05
7.0	2.0	1,000	0.26	0.07	0.08	0.07	0.07	0.05
10.0	2.0	100	0.27	0.08	0.17	0.12	0.14	0.06
10.0	2.0	200	0.28	0.05	0.11	0.07	0.09	0.04
10.0	2.0	500	0.32	0.06	0.11	0.08	0.10	0.04
10.0	2.0	1,000	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). Type I Error Rates at 5% Significance Level

Distribution			*df = 1*					
Kurt	Skew	N	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
0.0	0.0	200	0.02	0.03	0.03	0.03	0.03	0.03
0.0	0.0	500	0.02	0.02	0.02	0.03	0.02	0.03
0.0	0.0	1,000	0.03	0.03	0.03	0.03	0.03	0.03
7.0	2.0	100	0.28	0.09	0.12	0.08	0.09	0.08
7.0	2.0	200	0.25	0.06	0.09	0.06	0.08	0.07
7.0	2.0	500	0.30	0.05	0.06	0.06	0.06	0.06
7.0	2.0	1,000	0.30	0.06	0.07	0.06	0.06	0.06
10.0	2.0	100	0.26	0.08	0.13	0.08	0.09	0.08
10.0	2.0	200	0.29	0.07	0.10	0.07	0.08	0.07
10.0	2.0	500	0.30	0.05	0.08	0.05	0.07	0.05
10.0	2.0	1,000	0.32	0.05	0.07	0.06	0.06	0.06

Distribution			*df = 15*					
Kurt	Skew	N	D	DSB1$_{MLM}$	DSB1$_{MLR}$	DSB10$_{MLM}$	DSB10$_{MLR}$	D$_{MLMV}$
0.0	0.0	100	0.02	0.02	0.02	0.02	0.01	0.01
0.0	0.0	200	0.02	0.02	0.02	0.02	0.01	0.01
0.0	0.0	500	0.02	0.02	0.01	0.02	0.01	0.01
0.0	0.0	1,000	0.02	0.02	0.01	0.01	0.01	0.01
7.0	2.0	100	0.26	0.13	0.22	0.17	0.18	0.08
7.0	2.0	200	0.26	0.09	0.15	0.11	0.13	0.05
7.0	2.0	500	0.26	0.06	0.08	0.06	0.07	0.05
7.0	2.0	1,000	0.29	0.07	0.08	0.07	0.08	0.04
10.0	2.0	100	0.28	0.14	0.22	0.17	0.18	0.07
10.0	2.0	200	0.30	0.10	0.18	0.13	0.15	0.06
10.0	2.0	500	0.32	0.08	0.12	0.10	0.11	0.05
10.0	2.0	1,000	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 5% Significance Level

	Distribution						*df = 1*		
Δλ	Kurt	Skew	N	D	DSB1$_{\text{MLM}}$	DSB1$_{\text{MLR}}$	DSB10$_{\text{MLM}}$	DSB10$_{\text{MLR}}$	D$_{\text{MLMV}}$
0.2	0.0	0.0	100	0.31	0.32	0.31	0.33	0.30	0.33
	0.0	0.0	200	0.58	0.58	0.57	0.58	0.57	0.58
	0.0	0.0	500	0.94	0.94	0.94	0.94	0.94	0.94
	0.0	0.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00
	7.0	2.0	100	0.47	0.25	0.30	0.27	0.27	0.29
	7.0	2.0	200	0.58	0.33	0.36	0.35	0.35	0.35
	7.0	2.0	500	0.83	0.59	0.59	0.61	0.59	0.61
	7.0	2.0	1,000	0.97	0.88	0.86	0.88	0.87	0.88
	10.0	2.0	100	0.46	0.23	0.30	0.25	0.27	0.25
	10.0	2.0	200	0.62	0.34	0.37	0.36	0.35	0.36
	10.0	2.0	500	0.83	0.59	0.59	0.60	0.60	0.60
	10.0	2.0	1,000	0.95	0.83	0.81	0.83	0.82	0.84
0.4	0.0	0.0	100	0.83	0.84	0.84	0.85	0.83	0.85
	0.0	0.0	200	0.99	0.99	0.99	0.99	0.99	0.99
	0.0	0.0	500	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
	7.0	2.0	100	0.83	0.67	0.63	0.68	0.64	0.69
	7.0	2.0	200	0.96	0.88	0.84	0.89	0.86	0.89
	7.0	2.0	500	1.00	1.00	0.99	1.00	0.99	1.00
	7.0	2.0	1,000	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
	10.0	2.0	200	0.96	0.90	0.86	0.89	0.87	0.90
	10.0	2.0	500	1.00	0.99	0.98	0.99	0.98	0.99
	10.0	2.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00
							df = 5		
0.2	0.0	0.0	100	0.16	0.17	0.15	0.17	0.14	0.16
	0.0	0.0	200	0.32	0.34	0.33	0.33	0.31	0.32
	0.0	0.0	500	0.78	0.78	0.77	0.78	0.77	0.77
	0.0	0.0	1,000	0.99	0.99	0.99	0.99	0.99	0.99
	7.0	2.0	100	0.46	0.21	0.28	0.25	0.25	0.17
	7.0	2.0	200	0.56	0.29	0.34	0.33	0.33	0.25
	7.0	2.0	500	0.81	0.58	0.59	0.60	0.59	0.54
	7.0	2.0	1,000	0.96	0.87	0.88	0.88	0.88	0.85
	10.0	2.0	100	0.46	0.20	0.30	0.25	0.25	0.18
	10.0	2.0	200	0.60	0.30	0.35	0.34	0.34	0.27
	10.0	2.0	500	0.82	0.56	0.58	0.58	0.58	0.52
	10.0	2.0	1,000	0.96	0.84	0.83	0.84	0.83	0.80
0.4	0.0	0.0	100	0.58	0.59	0.58	0.60	0.56	0.58
	0.0	0.0	200	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
	0.0	0.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00
	7.0	2.0	100	0.79	0.58	0.59	0.61	0.58	0.53
	7.0	2.0	200	0.95	0.83	0.82	0.85	0.83	0.80
	7.0	2.0	500	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
	10.0	2.0	100	0.79	0.59	0.62	0.64	0.59	0.54
	10.0	2.0	200	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	0.98
	10.0	2.0	1,000	1.00	1.00	1.00	1.00	1.00	1.00

Notes: highlighted cells have incorrect Type I errors; *p* = nr. of indicators; Δλ = noninvariance; Kurt = Kurtosis; Skew = Skewness; N = sample size; *df* = degrees of freedom. D = uncorrected ML $\Delta\chi^2$; DSB1$_{\text{MLM}}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Satorra-Bentler χ^2 (1994); DSB1$_{\text{MLR}}$ = Satorra-Bentler $\Delta\chi^2$ (2001) with Asparouhov-Muthén χ^2 (2005); DSB10$_{\text{MLM}}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Satorra-Bentler χ^2 (1994); DSB10$_{\text{MLR}}$ = Satorra-Bentler $\Delta\chi^2$ (2010) with Asparouhov-Muthén χ^2 (2005); D$_{\text{MLMV}}$ = Asparouhov-Muthén $\Delta\chi^2$ (2010).

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

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

Notes: highlighted cells have incorrect Type I errors; *p* = nr. 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).

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 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 involving MLM outperformed choices involving 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 is 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). Because of the model size effect (Moshagen, 2012), this assumption may not be reliably tested in large models. 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

Supplementary materials

Example and Mplus Code for Computing Chi-Square Difference Tests

In the following example, we describe how to compute chi-square difference tests using Mplus software. Mplus syntax is included in the example so that applied researchers may utilize it in their own research.

The data set for the example ("pc.dat") contains four indicators measuring a unidimensional latent factor at two time-points. The total number of observed variables is thus eight. Sample size is 500.

The baseline model (M_1) in the example is the *configural invariance* model with the factor correlation and all factor loadings at each time-point freely estimated. The restricted model (M_0) is the *weak invariance* model, which introduces equality constraints on all four factor loadings between the two time-points. The number of additional constraints in the model M_0 is thus four.

We include Mplus code for computing the uncorrected difference test - D, Satorra-Bentler "original" difference test (2001) in concert with MLM - $DSB1_{MLM}$, Satorra-Bentler "strictly positive" difference test (2010) in concert with MLM - $DSB10_{MLM}$, and Asparouhov-Muthén chi-square difference test (2010) - D_{MLMV}. For options involving choice MLR, the code for MLM can be used simply by replacing the estimator in the analysis command.

We also provide results for all six options discussed in the paper.

Computing the Uncorrected Difference Test

Step 1: Estimating configural invariance model (M_1) with ML.

```
TITLE:      CONFIGURAL INVARIANCE (MODEL M1) WITH ML
DATA:       FILE IS 'pc.dat';
VARIABLE:   NAMES ARE
!four factor indicators at time 1
            pc1_1-pc1_4
!four factor indicators at time 2
            pc2_1-pc2_4;
ANALYSIS:   ESTIMATOR = ML;
MODEL:
!factor loadings freely estimated at both time points
            f1 by pc1_1-pc1_4*;
            f2 by pc2_1-pc2_4*;
!factor correlation freely estimated
            f2 with f1;
!both factor variances set to 1
            f1-f2@1;
!residual correlations between time points estimated
            pc1_1 with pc2_1;
            pc1_2 with pc2_2;
            pc1_3 with pc2_3;
            pc1_4 with pc2_4;
```

Chi-square test statistic (T_1) and degrees of freedom (df_1) are provided in the output.

Step 2: Estimating weak invariance model (M_0) with ML.

```
TITLE:      WEAK INVARIANCE (MODEL M0) WITH ML
DATA:       FILE IS 'pc.dat';
VARIABLE:   NAMES ARE
!four factor indicators at time 1
            pc1_1-pc1_4
!four factor indicators at time 2
            pc2_1-pc2_4;
ANALYSIS:   ESTIMATOR = ML;
MODEL:
!factor loadings are set to equality between time points
            f1 by pc1_1-pc1_4* (1-4);
            f2 by pc2_1-pc2_4* (1-4);
!factor correlation freely estimated
            f2 with f1;
!both factor variances set to 1
            f1-f2@1;
!residual correlations between time points estimated
            pc1_1 with pc2_1;
            pc1_2 with pc2_2;
            pc1_3 with pc2_3;
            pc1_4 with pc2_4;
```

Chi-square test statistic (T_0) and degrees of freedom (df_0) are provided in the output.

Step 3: The uncorrected chi-square difference statistic (D) is obtained with the Equation in (2). Number of degrees of freedom for the difference test is $df_0 - df_1$.

Computing Satorra-Bentler "Original" Difference Test (2001) with Choice MLM (or MLR)

Step 1: Estimating configural invariance model (M_1) with MLM.

```
TITLE:      CONFIGURAL INVARIANCE (MODEL M1) WITH MLM
DATA:       FILE IS 'pc.dat';
VARIABLE:   NAMES ARE
!four factor indicators at time 1
            pc1_1-pc1_4
!four factor indicators at time 2
            pc2_1-pc2_4;
ANALYSIS:   ESTIMATOR = MLM;
MODEL:
!factor loadings freely estimated at both time points
            f1 by pc1_1-pc1_4*;
            f2 by pc2_1-pc2_4*;
!factor correlation freely estimated
            f2 with f1;
!both factor variances set to 1
            f1-f2@1;
!residual correlations between time points estimated
            pc1_1 with pc2_1;
            pc1_2 with pc2_2;
            pc1_3 with pc2_3;
            pc1_4 with pc2_4;
```

Robust chi-square test statistic ($\overline{T}_1$), degrees of freedom (df_1), and scaling correction factor (c_1) are provided in the output.

Step 2: Estimating weak invariance model (M_0) with MLM.

```
TITLE:      WEAK INVARIANCE (MODEL M0) WITH MLM
DATA:       FILE IS 'pc.dat';
VARIABLE:   NAMES ARE
!four factor indicators at time 1
            pc1_1-pc1_4
!four factor indicators at time 2
            pc2_1-pc2_4;
ANALYSIS:   ESTIMATOR = MLM;
MODEL:
!factor loadings are set to equality between time points
            f1 by pc1_1-pc1_4* (1-4);
            f2 by pc2_1-pc2_4* (1-4);
!factor correlation freely estimated
            f2 with f1;
!both factor variances set to 1
            f1-f2@1;
!residual correlations between time points estimated
            pc1_1 with pc2_1;
            pc1_2 with pc2_2;
            pc1_3 with pc2_3;
            pc1_4 with pc2_4;
```

Robust chi-square test statistic ($\overline{T}_0$), degrees of freedom (df_0), and scaling correction factor (c_0) are provided in the output.

Step 3: The two scaling correction factors (c_1 and c_0) and two degrees of freedom (df_1 and df_0) are introduced into Equation in (3) to obtain scaling correction for the difference (c_{01}).

Step 4: The uncorrected chi-square difference statistic (D) computed earlier is divided by the scaling correction c_{01} to obtain the corrected chi-square difference statistic (DSB1$_{\text{MLM}}$). Number of degrees of freedom for the difference test is $df_0 - df_1$.

Computing Satorra-Bentler “Strictly Positive” Difference Test (2010) with Choice MLM (or MLR)

Step 1: Estimating weak invariance model (M_0) with MLM and requesting syntax for model M^* in the output.

```
TITLE:      WEAK INVARIANCE (MODEL M0) WITH MLM REQUESTING SVALUES
DATA:       FILE IS 'pc.dat';
VARIABLE:   NAMES ARE
!four factor indicators at time 1
            pc1_1-pc1_4
```

```
!four factor indicators at time 2
            pc2_1-pc2_4;
ANALYSIS:   ESTIMATOR = MLM;
MODEL:
!factor loadings are set to equality between time points
            f1 by pc1_1-pc1_4* (1-4);
            f2 by pc2_1-pc2_4* (1-4);
!factor correlation freely estimated
            f2 with f1;
!both factor variances set to 1
            f1-f2@1;
!residual correlations between time points estimated
            pc1_1 with pc2_1;
            pc1_2 with pc2_2;
            pc1_3 with pc2_3;
            pc1_4 with pc2_4;
!generating syntax for model M* in the output
OUTPUT:     SVALUES;
```

Robust chi-square test statistic ($\overline{T}_0$), degrees of freedom (df_0), scaling correction factor (c_0), and syntax for model M^* are provided in the output (under "MODEL COMMAND WITH FINAL ESTIMATES USED AS STARTING VALUES").

Step 2: Estimating configural invariance model M^*.

```
TITLE:      CONFIGURAL INVARIANCE MODEL M*
DATA:       FILE IS 'pc.dat';
VARIABLE:   NAMES ARE
!four factor indicators at time 1
            pc1_1-pc1_4
!four factor indicators at time 2
            pc2_1-pc2_4;
ANALYSIS:   ESTIMATOR = MLM;
!to set the number of iterations to 0
            CONVERGENCE=100000000
MODEL:
!introducing syntax for model M*
!obtained from the output of M1 run in step 1
            f1 BY pc1_1*0.57830;
            f1 BY pc1_2*0.56133;
            f1 BY pc1_3*0.62387;
            f1 BY pc1_4*0.48925;
            f2 BY pc2_1*0.57830;
            f2 BY pc2_2*0.56133;
            f2 BY pc2_3*0.62387;
            f2 BY pc2_4*0.48925;

            pc1_1 WITH pc2_1*0.10261;
            pc1_2 WITH pc2_2*0.03403;
            pc1_3 WITH pc2_3*0.15778;
            pc1_4 WITH pc2_4*0.17586;

            f2 WITH f1*0.84084;

            [ pc1_1*3.20400 ];
            [ pc1_2*3.61400 ];
            [ pc1_3*3.44600 ];
            [ pc1_4*3.51400 ];
            [ pc2_1*3.31400 ];
            [ pc2_2*3.72400 ];
            [ pc2_3*3.50400 ];
            [ pc2_4*3.56800 ];

            pc1_1*0.43837;
            pc1_2*0.28083;
            pc1_3*0.38353;
            pc1_4*0.46765;
            pc2_1*0.35235;
            pc2_2*0.23927;
            pc2_3*0.34973;
            pc2_4*0.48929;
            f1@1;
            f2@1;
!to confirm that the number of iterations was 0
OUTPUT:     TECH5;
```

The scaling correction factor c^* and degrees of freedom (df_1) are provided in the output.

Step 3: The two scaling corrections, c^* and c_0, and corresponding degrees of freedom (df_1 and df_0) are introduced into Equation in (4) to obtain the scaling correction c_{10}.

Step 3: The uncorrected chi-square difference (D) is divided by the scaling correction c_{10} to obtain the corrected chi-square difference statistic DSB10$_{\text{MLM}}$. Degrees of freedom for the difference test are $df_0 - df_1$.

Computing Asparouhov-Muthén (2010) Chi-Square Difference Test with Choice MLMV

Step 1: Estimating configural invariance model (M_1) with MLMV and saving data for the difference test.

```
TITLE:      CONFIGURAL INVARIANCE (MODEL M1) WITH MLMV
DATA:       FILE IS 'pc.dat';
VARIABLE:   NAMES ARE
!four factor indicators at time 1
            pc1_1-pc1_4
!four factor indicators at time 2
            pc2_1-pc2_4;
ANALYSIS:   ESTIMATOR = MLMV;
MODEL:
!factor loadings freely estimated at both time points
            f1 by pc1_1-pc1_4*;
            f2 by pc2_1-pc2_4*;
!factor correlation freely estimated
            f2 with f1;
!both factor variances set to 1
            f1-f2@1;
!residual correlations between time points estimated
            pc1_1 with pc2_1;
            pc1_2 with pc2_2;
            pc1_3 with pc2_3;
            pc1_4 with pc2_4;
```

```
!saving data for the difference test
SAVEDATA:   DIFFTEST IS diffmlmv.dat;
```

Step 2: Estimating weak invariance model (M_0) with MLMV using saved data from Step 1.

```
TITLE:      WEAK INVARIANCE (MODEL M0) WITH MLMV AND DIFFTEST
DATA:       FILE IS 'pc.dat';
VARIABLE:   NAMES ARE
!four factor indicators at time 1
            pc1_1-pc1_4
!four factor indicators at time 2
            pc2_1-pc2_4;
ANALYSIS:   ESTIMATOR = MLMV;
            DIFFTEST IS diffmlmv.dat;
MODEL:
!factor loadings are set to equality between time points
            f1 by pc1_1-pc1_4* (1-4);
            f2 by pc2_1-pc2_4* (1-4);
!factor correlation freely estimated
            f2 with f1;
!both factor variances set to 1
            f1-f2@1;
!residual correlations between time points estimated
            pc1_1 with pc2_1;
            pc1_2 with pc2_2;
            pc1_3 with pc2_3;
            pc1_4 with pc2_4;
```

The Asparouhov-Muthén (2010) chi-square difference statistic D_{MLMV}, degrees of freedom, and the corresponding p-value are available in the output under “Chi-Square Test for Difference Testing”.

Results

	Configural invariance (*df* = 15)		Weak invariance (*df* = 19)		Difference test (*df* = 4)		
Choice	χ^2	*p*-value	χ^2	*p*-value	Correction	$\Delta\chi^2$	*p*-value
ML	28.069	.0211	33.373	.0218	D (uncorrected difference test)	5.304	.2575
MLM	21.732	.1150	26.938	.1061	DSB1 (Satorra & Bentler, 2001)	5.094	.2778
	21.732	.1150	26.938	.1061	DSB10 (Satorra & Bentler, 2010)	4.917	.2960
MLR	21.979	.1084	26.983	.1050	DSB1 (Satorra & Bentler, 2001)	4.885	.2993
	21.979	.1084	26.983	.1050	DSB10 (Satorra & Bentler, 2010)	4.832	.3050
MLMV	21.004	.1367	26.099	.1275	D_{MLMV} (Asparouhov & Muthén, 2010)	4.976	.2898

Note: N = 500; *df* = degrees of freedom; χ^2 = chi-square; $\Delta\chi^2$ = chi-square difference.

GENERAL DISCUSSION

In SEM, testing for exact model fit is of critical importance because only if the model fits exactly, valid inferences regarding parameter estimates can be made. On the other hand, applied researchers are often interested in evaluating several competing theoretical models. In such cases, testing for overall model fit needs to be supplemented by evaluating relative (or comparative) fit of the models under consideration.

Both the exact model fit and model comparison are most commonly statistically evaluated using the likelihood ratio test statistic (often referred to in the literature simply as the "chi-square statistic" of exact fit). In classical SEM, when evaluating exact fit, the chi-square test statistic evaluates the fit of the proposed model against a saturated model. When evaluating the comparative fit of two competing models, a difference in fit can be evaluated as a difference of the two chi-square statistics, provided the two models are nested. Under normality assumptions and given that the sample size is sufficiently large, both the chi-square statistic and the chi-square difference statistic have known reference distributions, that is, are asymptotically chi-square distributed, thus allowing for statistical inference.

In practical applications, the normality assumption is commonly not tenable. In such cases, the reference distributions of the chi-square statistic and the chi-square difference statistic may deviate from chi-square reference distribution, thus questioning the validity of substantive conclusions. It has been consistently shown via simulations that both chi-square statistics are sensitive to departures from normality assumptions (Brace & Savalei, 2017; Chuang et al., 2015; Fouladi, 2000; Hu et al., 1992; Satorra, 1990; Satorra, 2000; Satorra & Bentler, 1994). In addition, other factors, such as sample size and model size, may also distort the accuracy of chi-square statistics' *p*-values obtained using their reference asymptotic distribution.

To deal with non-normal data, several approaches has been pursued, including adjusting the likelihood ratio statistics so that they can be better approximated by a chi-square distribution when data is non-normal, bootstrapping methods, and using alternative test statistics that are robust to violations of normality assumption.

This doctoral thesis aims at contributing to both the exact model fit and model comparison literatures within the framework of structural equation modeling when data need not be normally distributed.

Two simulation studies were presented. In the first simulation study included in this thesis, I re-examined the performance of the current gold standard for exact goodness of fit assessment of SEM models, the likelihood ratio test statistic (e.g., Jöreskog, 1969) and its asymptotic mean and variance adjustment (Asparouhov & Muthén, 2010; Satorra & Bentler, 1994) against a recently proposed SRMR statistic for exact fit and its asymptotic mean and variance adjustment (Maydeu-Olivares, 2017b). Major factors of the performance of the chi-square test *p*-values were considered in the context of a CFA population model, including model size (defined as the number of observed variables), sample size, and (non)normality. In the second simulation study included in this thesis, I compared the performance of the uncorrected, and mean adjusted chi-square difference test (Satorra & Bentler, 2001; Satorra & Bentler, 2010) to its mean and variance adjusted version (Asparouhov & Muthén, 2006, 2010) in the context of testing for weak factorial invariance across time. Major factors affecting the performance of these fit statistics' *p*-values were considered including model size (defined as the number of observed variables), relative model size (defined as *df* for the difference test), sample size, and (non)normality.

General summary of findings

The results of the first study replicated findings of previous research. Specifically, under normality, the uncorrected version of the chi-square performed accurately even in small samples when the models involved a few variables. However, when larger models were considered, the test rejected the null hypothesis too often (i.e., showed inflated Type I error rates); and the quality of the asymptotic approximation did not sufficiently improve even at the largest sample sizes considered (*N* = 1000). Under non-normality, the uncorrected version of the statistic led to a severe model over-rejection across the board, which was in line with previous research (e.g., Fouladi, 2000; Hu et al., 1992; Satorra, 1990; Satorra & Bentler, 1994). The mean and variance adjustment to the chi-square test showed considerable improvements over the unadjusted version. Specifically, the results revealed that the adjusted version was relatively robust to nonnormality. However, when data were not normal, somewhat larger sample sizes were needed for the test to be accurate. On the other hand, the accuracy of the adjusted chi-square test was susceptible to model size effects. For large models, the accuracy of the chi-square adjusted *p*-values was again compromised and could not be fully rectified even with the largest sample sizes considered.

With respect to the behavior of the SRMR, when data were normally distributed, SRMR *p*-values obtained under normality were reasonably accurate, even in small samples, and for large models, in which the chi-square test and its adjusted version failed. When data were nonnormal, and SRMR *p*-values robust to non-normality were used, they were accurate only in small models. For larger models, the use of larger sample sizes did not lead to substantially better results (except for medium models and when no skewness was present).

In sum, based on the results of the first study, it seems that the current standard, that is, the mean and variance corrected chi-square test, performs accurately under violations of normality provided sample size is not too small. The major problem with this adjustment seems to be that testing large models must involve very large sample sizes for *p*-values to be accurate. When data are normal, the SRMR computed under normality is less susceptible to the effect of model size, and it outperforms the current standard. However, for nonnormal data, the robust

SRMR outperforms the adjusted chi-square test only in small models. For larger models, the mean and variance corrected chi-square test outperforms the robust SRMR. Accordingly, based on the results of this work, the SRMR does not seem to be the solution to assess exact model fit in large models.

With respect to the behavior of the chi-square difference test, the results of the second study replicated findings of previous research. Under normality, the uncorrected chi-square difference test performed accurately regardless of the sample size, model size, and *df* for the difference test. When data were nonnormal, the uncorrected test rejected the null hypothesis too often. In the current study, the magnitude of overrejection was not notably affected by varying levels of non-normality, model size, *df* for the difference test, and sample size. The accuracy of the mean adjusted difference tests was also affected by nonnormality. In addition, the magnitude of overrejection of the mean adjustment seemed to be compounded by other factors considered in the simulation. Specifically, the performance of the mean adjustments was the worst when large models, large *df* for the difference test, and small samples were considered. In general, the use of larger samples led to Type I error rates closer to their nominal levels, but the convergence of the reference distribution to the actual sampling distribution of the test statistic is slower the larger the size and degrees of freedom of the models considered. Finally, the focal statistic considered in the study, the mean and variance chi-square adjustment, performed relatively accurately across all conditions considered.

In sum, based on the results of the second study, it seems that the mean and variance chi-square difference adjustment performs better than the mean adjustments when comparing the fit of two nested models. In general, for mean adjustments to yield accurate Type I error rates, larger sample sizes are needed than if mean and variance adjustments are used. However, it is important to note that even though the mean and variance adjusted chi-square difference test outperforms the mean adjusted versions with respect to Type I error rates, both versions require very similar sample sizes to reach adequate levels of statistical power.

Limitations and future research directions

There are several limitations to the current work. In the study on exact model fit, the performance of the SRMR and the chi-square test was evaluated only with respect to Type I error rates but not power. Limiting the investigation only to Type I errors seemed appropriate given that they were inaccurate in many investigated conditions. Second, the data generating model in the first study was a relatively simple model with a saturated mean structure, i.e., a CFA model. Future research should extend the current study, for example, by considering a greater range of number of factors and indicators and by varying magnitudes of factor loadings (e.g., see Hancock & Mueller, 2011; Ximénez et al., 2022). Mean and covariance structure models and more complex SEM models such as growth or multilevel models should also be considered in the future.

In the study on exact model fit, the SRMR was investigated as an alternative to the current standard, that is, the mean and variance adjusted chi-square test. Pending further investigation, SRMR in its current form does not seem to be a viable alternative to the current standard. That said, future research should also continue investigating other alternatives to evaluate exact model fit. One of the viable options is to use resampling methods, i.e., bootstrapping. Recently, Corrêa Ferraz and colleagues (2022) compared the Bollen-Stine bootstrapping method (Bollen & Stine, 1992) to the mean and variance adjusted chi-square test. The authors found that the Bollen-Stine method tends to underreject (i.e., yields deflated Type I error rates) under suboptimal conditions and overall performs worse than the mean and variance adjusted chi-square test. Similar to the SRMR, the Bollen-Stine bootstrapping method seems to be particularly sensitive to model size (Corrêa Ferraz et al., 2022) and it also does not seem to be a viable alternative to the current standard. Along this line of inquiry, future research should also consider investigating the performance of alternative bootstrapping schemes.

Another viable option is to rely on alternative test statistics (e.g., Foldnes & Grønneberg, 2018; Hayakawa, 2019; Wu & Lin, 2016). For example, Foldnes and Grønneberg (2018) showed that their method of *eigenvalue block averaging* performs more accurately the mean and variance adjusted chi-square test under nonnormality and small sample sizes. More recently, Zheng and Bentler (2023) found that Hayakawa's (2019) robust version of the *reweighted least squares (RLS) statistic* performs substantially better than the robust chi-square under suboptimal conditions. Although the performance of these recently proposed test statistics seems promising, additional research is required to ascertain their superiority over the current standard. In addition, such research efforts must be supported by prompt implementation of these tests into the popular statistical software packages. For instance, at the time of this writing, the RLS test statistics is only available in EQS (Bentler, 2004) and LISREL (Jöreskog & Sörbom, 2017); it has not yet been implemented in Mplus (Muthén & Muthén, 2017) and the existing R (R Core Team, 2019) packages cannot compute it.

Most of the limitations pertaining to the study of the exact model fit may also be put forth with respect to the study of comparative model fit. Specifically, in the second study, the mean and variance chi-square difference adjustment performed accurately across all investigated conditions. However, the data generating model in this study was a relatively simple CFA model and model comparison was investigated in the context of weak longitudinal invariance by constraining some of the factor loadings to be equal. That said, additional research is needed to replicate these findings in different scenarios including more complex population models, larger model sizes, and larger *df* for the difference test. In addition, although the data were generated using an advanced data generating method (Foldnes & Olsson, 2016), future research should also consider other alternatives for generating nonnormal data. Finally, nonnormality conditions generated in current work were not as extreme as in some related studies (e.g., Chuang et al., 2015), and future research could replicate current findings on the mean and variance chi-square difference adjustment in conditions involving greater departures from normality from those specified in the current work.

On a related note, the accuracy of mean adjustments to the chi-square difference statistic was the worst in conditions involving large models and large *df* for the difference test. However, with the current design, it was not possible to disentangle the effect of the model size from the effect of *df* for the difference test. These effects should be in focus and systematically investigated in future research efforts.

General conclusions and recommendations

In this doctoral thesis, I evaluated the performance of chi-square tests in the context of exact model fit and in the context of comparison in fit between two models that are nested. The findings of this thesis are of relevance for psychologists across various psychological subdisciplines such as educational, developmental, health, clinical, and industrial-organizational, to name a few, and across various settings including academia, for-profit, non-profit, and government. Application of appropriate statistical methods to evaluate and select theoretical models is essential for a robust advancement of psychological science that in turn may have important implications and facilitate psychological practice in terms of evidence-based prediction and intervention, valid and fair psychological testing and assessment, and/or ultimately, in terms of shaping public policy more broadly.

With respect to the issue of evaluating exact model fit, the mean and variance adjustment to the chi-square test outperformed its uncorrected version and performed relatively accurately under various suboptimal conditions considered in this work. It should be noted, however, that when fitting large models, results of the mean and variance adjusted chi-square test may still be suspect because very large sample sizes were needed for it to perform accurately (N > 1000) even when data were normally distributed. Accordingly, based on the findings of this thesis, researchers are encouraged to make their sample size decisions considering the size of their theoretical models.

In the current work, a recently proposed SRMR test for exact model fit performed overall worse than the mean and variance adjustment to the chi-square test. Accordingly, based on the results of this thesis and pending further investigation, I cannot advise researchers to include the SRMR in its current form in their standard toolkit for evaluating exact fit of models under consideration.

When comparing the fit of two nested models, the results of this thesis reconfirmed that the unadjusted version of the chi-square difference test does not perform accurately when data are not normal. Therefore, researchers are strongly advised to use some of the available adjustments to the chi-square difference test to increase the validity of their statistical conclusions and the confidence in their findings. In this thesis, the mean and variance adjustment to the chi-square difference test performed accurately and outperformed mean adjustments in all conditions under investigation. Overall, based on the results of the current work, researchers should consider the mean and variance adjustment to the chi-square statistic as the option of choice for testing both exact model fit and model comparison.

REFERENCES

Arbuckle, J. L. (1996). Full Information Estimation in the Presence of Incomplete Data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced Structural Equation Modeling: Issues and Techniques* (pp. 277–365). Mahwah, NJ: Lawrence Erlbaum.

Arbuckle, J. L. (2014). Amos (version 23) [Computer program]. Chicago, IL: IBM SPSS.

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 https://www.statmodel.com/download/2005FCSM_Asparouhov_Muthen_IIA.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

Asparouhov, T., & Muthén, B. (2018). SRMR in Mplus. Los Angeles, CA. Retrieved from www.statmodel.com/download/SRMR2.pd

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815-824. https://doi.org/10.1016/j.paid.2006.09.018

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]. Encino, CA: Multivariate Software Inc.

Bentler, P. M. (2004). EQS 6 [Computer Program]. Encino, CA: Multivariate Software Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3)*,* 588–606. http://dx.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., & Curran, P. J. (2006). *Latent curve analysis. A structural equations perspective*. Hoboken, NJ: Wiley.

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). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, *21*(2), 205-229. https://doi.org/10.1177/0049124192021002004

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

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://hdl.handle.net/10520/AJA0038271X_175

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511897375

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean-and covariance-structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York: Plenum. https://doi.org/10.1007/978-1-4899-1292-3

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

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*, 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

Corrêa Ferraz, R., Maydeu-Olivares, A., & Shi, D. (2022). Asymptotic is better than Bollen-Stine bootstrapping to assess model fit: The effect of model size on the chi-square statistic. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(5), 731-743. https://doi.org/10.1080/10705511.2022.2053128

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the “problem” of sample size: A clarification. *Psychological Bulletin*, *109*(3), 512–519. https://doi.org/10.1037/0033-2909.109.3.512

DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling: A Multidisciplinary Journal*, *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: A Multidisciplinary Journal*, *21*(3), 425–438. https://doi.org/10.1080/10705511.2014.915373

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

Embretson, S., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Finney, S., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (2nd ed., pp. 439–492). Charlotte, NC: Information Age. https://doi.org/10.1111/j.1744-6570.2007.00081_13.x

Foldnes, N., & Grønneberg, S. (2018). Approximating test statistics using eigenvalue block averaging. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(1), 101–114. https://doi.org/10.1080/10705511.2017.1373021

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

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, *14*(3), 275–299. https://doi.org/10.1037/a0015825

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(3), 356-410. https://doi.org/10.1207/S15328007SEM0703_2

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: A Multidisciplinary Journal*, *27*(2), 192–201. https://doi.org/10.1080/10705511.2019.1637741

Grønneberg, S., & Foldnes, N. (2019). Testing model fit by bootstrap selection. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(2), 182–190. https://doi.org/10.1080/10705511.2018.1503543

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

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

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

Hayakawa, K. (2019). 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: A Multidisciplinary Journal*, *14*(3), 361–390. https://doi.org/10.1080/10705510701301602

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

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

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

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N: Q hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 128–141. https://doi.org/10.1207/S15328007SEM1001_6

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

Jiang, G., & Yuan, K. H. (2017). Four new corrected statistics for SEM with small samples and nonnormally distributed data. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 479-494. https://doi.org/10.1080/10705511.2016.1277726

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.). Chicago, IL: International Education Services.

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

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(3), 333–351. http://dx.doi.org/10.1207/S15328007SEM1003_1

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

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

Lai, K. (2019). A simple analytic confidence interval for CFI given nonnormal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *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). London: 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. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*(1), 19–35. https://doi.org/10.1037/1082-989X.11.1.19

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: 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

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: A Multidisciplinary Journal*, *25*(3), 389–402. https://doi.org/10.1080/10705511.2017.1389611

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: A Multidisciplinary Journal*, *19*(1), 86–98. https://doi.org/10.1080/10705511.2012.634724

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132. https://doi.org/10.1007/BF02294210

Muthén, B. (1989). Multiple-group structural modelling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology*, *42*(1), 55–62. https://doi.org/10.1111/j.2044-8317.1989.tb01114.x

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. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599-620. https://doi.org/10.1207/S15328007SEM0904_8

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

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(3)*,* 353–377. http://dx.doi.org/10.1207/S15328007SEM0803_2

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439-478. https://doi.org/10.1207/S15327906MBR3903_3

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

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

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

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. Vienna, Austria, Austria: 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

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

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). Newbury Park, CA: 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. (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*, 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*, 243-248. https://doi.org/10.1007/s11336-009-9135-y

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

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), 419–429. https://doi.org/10.1080/00273171.2018.1455142

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

Shams, R., Brown, M., & Alpert, F. (2017). The role of brand credibility in the relationship between brand innovativeness and purchase intention. *Journal of Customer Behaviour*, *16*(2), 145-159. https://doi.org/10.1362/147539217X14909732699534

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. (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., 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: A Multidisciplinary Journal*, *27*(1), 1–15. https://doi.org/10.1080/10705511.2019.1611434

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: CRC Press.

Steiger, J. H. (1989). EzPATH: A supplementary module for SYSTAT and SYGRAPH. Evanston, IL: Systat, Inc. Retrieved from https://www.statpower.net/Steiger%20Biblio/EzPath%20Manual.pdf

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). Statistically based tests for the number of common factors. *Paper Presented at the Annual Meeting of the Annual Spring Meeting of the Psychometric Society, Iowa City.*

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

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

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 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

Wu, H., & Lin, J. (2016). A scaled F distribution as an approximation to the distribution of test statistics in covariance structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 409-421. https://doi.org/10.1080/10705511.2015.1057733

Ximénez, C., Maydeu-Olivares, A., Shi, D., & Revuelta, J. (2022). Assessing Cutoff Values of SEM Fit Indices: Advantages of the Unbiased SRMR Index and Its Cutoff Criterion Based on Communality. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 368–380. https://doi.org/10.1080/10705511.2021.1992596

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. & 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

Yuan, K. H., Hayashi, K., & Yanagihara, H. (2007). A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research*, *42*(2), 261-281. https://doi.org/10.1080/00273170701360662

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

Zheng, B. Q., & Bentler, P. M. (2023): RGLS and RLS in covariance structure analysis, *Structural Equation Modeling: A Multidisciplinary Journal, 30*(2), 234-244. https://doi.org/10.1080/10705511.2022.2117182