Summer 2020

Bollen-Stine Bootstrapping of the Chi-Square Statistic in Structural Equation Models: The Effect of Model Size

Raul Corrêa Ferraz

BOLLEN-STINE BOOTSTRAPPING OF THE CHI-SQUARE STATISTIC IN STRUCTURAL EQUATION MODELS: THE EFFECT OF MODEL SIZE

by

Raul Corrêa Ferraz

Bachelor of Science
Federal University of Santa Maria, 2018

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Arts in

Experimental Psychology

College of Arts and Sciences

University of South Carolina

2020

Accepted by:

Alberto Maydeu-Olivares, Director of Thesis

Dexin Shi, Committee Member

John Richards, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

ABSTRACT

Previous research on the accuracy of *p*-values for the chi-square test of model fit has been limited to small models (around 10 variables), revealing that they are accurate provided sample size is not too small. At small sample sizes ($N < 100$), the usual *p*-values, obtained using asymptotic methods, are more accurate. However, asymptotic *p*-values incorrectly suggest that models fit poorly when the number of variables is large. We investigate whether Bollen-Stine (1992) bootstrap *p*-values are accurate in large models (up to 30 variables) for continuous outcomes using both normal and non-normal data. We found that as model size increases bootstrap *p*-values become too conservative (rejection rates are too small) and remarkably less accurate than asymptotic *p*-values obtained using robust methods (i.e., mean and variance corrected chi-square statistics). Further, there is a significant interaction between model size and sample size such that *p*-values for bootstrap are less accurate when the model is large and the sample size is small. Bollen-Stine *p*-values cannot be recommended to assess the fit of large models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A key element of Structural Equation Modeling (SEM) is the assessment of the fit of the estimated model to the data at hand. Model fit evaluation should be performed before any interpretation of parameter estimates, since any conclusion based on a poorly fitted model could be misleading (Maydeu-Olivares, 2017a). A number of test statistics can be used to assess whether a SEM model fits exactly (M. W. Browne, 1984; Satorra & Bentler, 1994; Yuan & Bentler, 1997, 1998, 1999). However, in applications involving maximum likelihood (ML) estimation under normality assumptions, the most widely used test statistic to assess the goodness of fit in SEM is the likelihood ratio (LR) test statistic (Maydeu-Olivares, 2017b). If the data are multivariate normal, the LR follows asymptotically a chi-square distribution when the model is correctly specified, and a non-central chi-square when the model is not correctly specified (Hoyle, 2012; Wang & Wang, 2012). As a result of its popularity, the LR test statistic is commonly referred to in the SEM literature as the chi-square test. For non-normal data, the most widely used test statistics when ML estimation is employed involves using a mean correction, or a mean and variance correction, to the likelihood ratio test (Satorra & Bentler, 1994). Generically, these test statistics robust to non-normality are generally referred to as "robust" chi-square tests.

Whether researchers should test the null hypothesis of exact fit has been and still is hotly debated in the SEM literature. A special issue of *Personality and Individual Differences* with a leading paper by Barrett (2007), is a good starting point for readers interested in this topic. Some of the arguments put forth for not assessing the exact fit of the model are (a) power may be excessive, leading to model rejection for misfits that have no practical significance, (b) it is unlikely that models fit perfectly because human knowledge in any field is not at the point where a perfect model is possible, and by definition models are essentially approximations in the first place, or (c) the focus should be on comparing alternative models and not on model fit.

One of the main discussion points in the literature on evaluating fit in SEM is whether the null hypothesis of exact fit should be replaced by a null hypothesis of approximate fit. Assuming that no constraints are imposed on the mean structure, the null hypothesis of exact fit in SEM states that the covariance structure implied by the fitted model matches exactly the unknown population covariance matrix (Bentler, 1990; Bollen, 1989; Hoelter, 1983). In a null hypothesis of approximate fit, this null hypothesis is relaxed, being replaced by a hypothesis that the covariance matrix implied by the fitted model matches the unknown population covariance matrix by a pre-specified margin of error. There are different ways to measure the discrepancy between population covariance matrix implied by the fitted model and the unknown data-generating population covariance matrix – effect sizes of model misfit in Maydeu-Olivares (Maydeu-Olivares, 2017a) terminology. The most widely used effect size of model misfit (a population parameter used to assess approximate fit) is the Root Mean Squared Error of Approximation (RMSEA: Browne & Cudeck, 1992; Steiger, 1990). In the RMSEA,

the discrepancy between the two population covariance matrices is unstandardized and adjusted for model parsimony. Alternatively, a standardized measure of population misfit may be used, such as the Standardized Root Mean Squared Residual (SRMR: Bentler, 1995; Jöreskog & Sörbom, 1988; Maydeu-Olivares, 2017a).

Regardless of whether exact or approximate fit is assessed, it is critical that the performance of the test statistic used to assess model fit be adequate, namely, that empirical rejection rates match Type I errors and that the statistic has sufficient power to detect models that are substantially misspecified. Unfortunately, it is well know that empirical rejection rates for the likelihood ratio test statistic, possibly robustified to address data non-normality, are in some situations inflated, leading to rejecting well-fitting models (Anderson & Gerbing, 1984; Bentler & Yuan, 1999; P. Curran, West, & Finch, 1996; Fouladi, 2000; Herzog, Boomsma, & Reinecke, 2007; Hu, Bentler, & Kano, 1992; Moshagen, 2012; Nevitt & Hancock, 2004). Similar results of over-rejection of well-fitting models in some setups is also found when the RMSEA is used to assess approximate model fit (Curran, Bollen, Chen, Paxton, & Kirby, 2003; Fan, Thompson, & Wang, 1999; Hu & Bentler, 1998; Kenny, Kaniskan, & McCoach, 2015; Maydeu-Olivares, Shi, & Rosseel, 2018; Nevitt & Hancock, 2000). That is to be expected, as the sample RMSEA is a function of the likelihood ratio test statistic. As a result, if the empirical performance of the former is poor, the empirical performance of the latter is also likely to be poor.

One of the main drivers of the performance of the likelihood ratio test statistic and of the RMSEA is model size, that is, the number of observed variables. A number of studies (Herzog et al., 2007; Moshagen, 2012; Shi, Lee, & Terry, 2018) have reported

that the empirical sampling distribution of the likelihood ratio statistic, possibly robustified to account for non-normality, is poorly approximated by its reference asymptotic distribution when the model involves a large number of observed variables. Similarly, the empirical sampling distribution of the estimated RMSEA is not well approximated by its reference asymptotic distribution when model size is large (Maydeu-Olivares et al., 2018). The largest number of observed variables at which the likelihood ratio test statistic or the RMSEA yield accurate *p*-values is around 30 (Maydeu-Olivares, 2017b; Maydeu-Olivares et al., 2018); beyond that, the use of both statistics leads to over-rejection of well-fitting models. Of course, the performance of these statistics depends on several additional factors (see the references above), such as sample size and average R^2 of the observed variables (i.e., factor loading in factor analysis models). Yet, a number of observed variables way above 30 appears to be an unsurmountable barrier for the adequate performance of both the likelihood ratio statistic and the RMSEA.

Of course, one approach to overcome the limitations of the likelihood ratio test statistic (and of the RMSEA) when the number of observed variables is large is to use alternative test statistics. For instance, Hayakawa (Hayakawa, 2019) has recently shown that a statistic originally proposed by Browne (1982) for normally distributed data performs well in large models under normality, and that when robustified adjusting it by its asymptotic mean and variance, it performs well under non-normality. Similarly, Maydeu-Olivares, Shi and Rosseel (Maydeu-Olivares et al., 2018) show that approximate fit can be reliably assessed in large models using the SRMR, for both normal and non-normal data.

In this article, because of the popularity of the likelihood ratio test we take on a different path. Instead of identifying alternative test statistics that may yield accurate *p*-values when assessing fit in large models, we investigate by simulation whether accurate *p*-values for the likelihood ratio test statistic can be obtained in conditions of large model size using the bootstrap (Bradley Efron & Tibshirani, 1993; Stine, 1989). In particular, we will focus on *p*-values obtained using the most widely used bootstrap procedure in the SEM literature, that proposed by Bollen and Stine (Bollen & Stine, 1992), and compare its performance to *p*-values obtained using asymptotic methods. Previous research (Grønneberg & Foldnes, 2018; Nevitt & Hancock, 2001) has shown that this procedure yields accurate *p*-values in small models but the behavior of this approach with large models remains to be investigated.

The remaining of this article is organized as follows: First we describe the likelihood ratio test statistic and the robust version used for non-normal data that we will use as benchmark in our simulations. Next, we describe the bootstrap procedure proposed by Bollen and Stine (1992) to obtain *p*-values for a test of exact fit using the likelihood ratio test. In this section, we review previous research on the performance of this method. Next, we describe the simulation conditions employed and summarize the results obtained. We conclude with a discussion of the results, offer guidelines for applied researchers and outline future lines of research.

CHAPTER 2

THE LIKELIHOOD RATIO TEST STATISTIC AND ITS ROBUST VERSIONS TO ACCOUNT FOR NON-NORMALITY

When no structure is imposed on the intercepts of the model (i.e., in covariance structure analysis), the null and alternative hypotheses of model fit can be written as $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ and $H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$, where $\boldsymbol{\Sigma}$ denotes the population covariance matrix, and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes the covariance matrix implied by the theoretical model under consideration, expressed as a function of the model parameters $\boldsymbol{\theta}$. When ML estimation is used, almost invariably, the test statistic used to test the null hypothesis is the likelihood ratio test

$$\begin{aligned} & X^2_{ML} = (N-1)\hat{F}_{ML} \\ & \textit{where} \\ & \hat{F}_{ML} = \ln\left|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right| - \ln\left|\mathbf{S}\right| + tr(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})) - p \end{aligned} \tag{1}$$

where $\mathbf{S}$ is the sample covariance matrix, $\hat{\boldsymbol{\theta}}$ denote the estimated parameters, *N* denotes sample size, and *p* is the number of observed variables. We use X^2_{ML} to refer to this statistic as it is commonly denoted the chi-square test in the SEM literature. When data is normally distributed and the model is correctly specified, X^2_{ML} follows asymptotically a chi-square distribution with $p(p+1)/2 - q$ degrees of freedom, where *q* is the number of mathematically independent elements in the parameter vector $\boldsymbol{\theta}$.

When the data are not normally distributed, X^2_{ML} will not follow a chi-square distribution even when the model is correctly specified, and the use of X^2_{ML} leads to over-rejecting well-fitting models (e.g., West, Finch, & Curran, 1995). The most widely used approach to solve this problem is to adjust X^2_{ML} so that the resulting test statistic matches asymptotically a chi-square distribution either in its mean (e.g., Asparouhov & Muthén, 2005; Satorra & Bentler, 1994; Yuan & Bentler, 2000), or in its mean and its variance mean (Asparouhov & Muthén, 2010b; Satorra & Bentler, 1994). Here, we focus on the mean and variance corrected statistics as they have been shown to provide more accurate *p*-values than mean corrected statistics (Foldnes & Olsson, 2015; Maydeu-Olivares, 2017b) at the expense of additional computations. In particular, Maydeu-Olivares (2017b) showed that in large models ($p \geq 32$) the mean corrected statistics over-reject the model unless $N > 1{,}000$ whereas the mean and variance corrected statistic examined maintained its nominal rates except in small samples ($N \leq 200$). Of the two mean and variance corrected statistics proposed in the literature mean (Asparouhov & Muthén, 2010b; Satorra & Bentler, 1994), the former slightly outperforms the latter (Foldnes & Olsson, 2015; Savalei & Rhemtulla, 2013), and will be our choice of robust likelihood ratio test statistic. We refer to the mean and variance corrected likelihood ratio statistic proposed by Asparouhov and Muthén (2010) as X^2_{MLMV} using the nomenclature used in the widely used software M*plus* (Muthén & Muthén, 2017) and we note that $X^2_{MLMV} = a + X^2_{ML} / c$ where *a* and *c* are shift and corrections to X^2_{ML}. See Asparouhov and Muthén (2010) for the expressions of these constants.

CHAPTER 3

OBTAINING BOOTSTRAP *P*-VALUES FOR THE CHI-SQUARE TEST: BOLLEN AND STINE'S (1992) PROCEDURE

The bootstrap is a technique introduced by Bradley Efron in the 1970s (Efron, 1979) as a more dependable and widely applicable version of the jackknife method; differences between the bootstrap and the jackknife are detailed in Efron and Tibshirani (1993). Stine (Stine, 1989) and Hartmann (2005) are two additional suitable introductions to these methods. Generally speaking, the bootstrap is a computer intensive procedure used to obtain confidence intervals for parameter estimates, or *p*-values for hypothesis testing. Standard procedures for obtaining confidence intervals and *p*-values rely on the use of distributional assumptions (e.g., normality of the data), or large sample assumptions (i.e., that the sample size is large enough to rely on the central limit theorem). The bootstrap provides an alternative elegant solution to the problem of approximating the sampling distribution of a statistic when the population distribution is unknown. Bootstrapping draws repeated samples with replacement (called bootstrap samples) from the parent sample; the parameter of interest is then determined for each bootstrap sample and the empirical distribution of each parameter's bootstrap may be used for statistical inference. There are different approaches that can be used; they differ in what data set is used as the parent sample.

3.1 Naive (aka non-parametric) bootstrapping

The simplest form of bootstrapping involves using the original data set as the parent sample (Bradley Efron, 1979). Confidence intervals (CIs) for parameters can be obtained, for instance, by taking the appropriate percentiles of the bootstrap sampling distribution of the parameter estimates. This method is referred to as percentile bootstrapping (Bradley Efron & Tibshirani, 1993). For instance, if a 95% CI is desired for a parameter, the 2.5^{th} and 97.5^{th} percentiles of the bootstrap sampling distribution of the parameter estimates are used as endpoints for the CI. This method has been successfully used in SEM to obtain CIs for parameter estimates in both complete (Hancock & Liu, 2012; Nevitt & Hancock, 2001; Yuan & Hayashi, 2003) and incomplete data (Enders, 2001) scenarios. Bootstrap methods can also be used to obtain confidence intervals for functions of parameter estimates, such as indirect effects, and have become the method of choice for obtaining CIs in mediation analysis (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Williams & Mackinnon, 2008).

3.2 Model-based bootstrapping

The traditional naive bootstrapping just described cannot be used for the likelihood ratio test that is the focus of our research. Bollen and Stine (1992) have shown that the mean and variance of the bootstrap distribution of the likelihood ratio statistic are larger than mean and variance of the sampling distribution of the test statistic in the original sample. In other words, the naive bootstrap sampling distribution of X^2_{ML} "would contain noncentrality reflective of the degree to which the model under scrutiny is misspecified" (Hancock & Liu, 2012). The solution proposed by Bollen and Stine

(1992) is to transform the matrix of centered observed variables, **Y**, using the parent sample covariance matrix, **S**, and the estimated covariance matrix, $\hat{\boldsymbol{\Sigma}}$, using

$$\mathbf{Z} = \mathbf{Y}\mathbf{S}^{-1/2}\hat{\boldsymbol{\Sigma}}^{1/2}. \quad (2)$$

The bootstrap is then performed by resampling rows of **Z** instead of resampling rows of the original data matrix **Y**. This transformed parent sample has the same distribution than the original parent sample but with a perfect model fit (Hancock & Liu, 2012). We note that an earlier and more technical description of this model-based approach to bootstrapping was introduced by Beran and Srivastava (1985), including analytic proofs; nonetheless, we will use the Bollen-Stine reference for consistency within the field of SEM.

The Bollen-Stine approach to bootstrap *p*-values of the chi-square test statistic has been implemented in the M*plus* software (Muthén & Muthén, 2017) under the name *residual parametric option*, and is available only for continuous outcomes using ML estimation. The R (R Core Team, 2019) package *lavaan* (Rosseel, 2012) implements the Bollen-Stine bootstrap under the "bootstrapLavaan" command, with argument = "bollen.stine". Finally, AMOS (Arbuckle, 2017) uses the Bollen-Stine approach to bootstrap *p*-values of the chi-square test statistic; in addition, it implements the Linhart and Zucchini (1986) bootstrap method for model comparison.

3.3 Previous research on the performance of the Bollen-Stine method

Despite having been proposed in 1992, few articles have examined the performance of the Bollen-Stine approach to bootstrap *p*-values of the chi-square of fit.

Fouladi (1998) investigated the performance of these *p*-values in two models involving $p = 6$ variables: an independence model, and a simplex model. Additional

conditions were obtained by crossing three different levels of skewness (0, 1, 2) and four different levels of kurtosis (-1, 0, 1, 3, 6). Sample sizes ranged from 30 to 750 observations, 5,000 replications were used in each condition, and 1,000 bootstrap samples were used per replication. Results showed that the Bollen-Stine *p*-values performed well provided sample size was at least 150 observations.

Nevitt and Hancock (2001) used models with $p = 9$ variables to compare the performance of Bollen-Stine bootstrap *p*-values to asymptotic *p*-values for X^2_{ML} and for the mean corrected chi-square proposed by Satorra and Bentler (1994), X^2_{MLM} in M*plus* terminology. Using 200 replications per condition, they considered samples of sizes 100 to 1,000 observations and the use of between 250 to 2,000 bootstrap samples. Three distributional conditions were employed: multivariate normal, moderately non-normal (skewness = 2, kurtosis = 7), and extremely non-normal (skewness = 3, kurtosis = 21). Bootstrap *p*-values yielded accurate rejection rates across all the models considered regardless of the number of bootstrap samples employed. In particular, they were more accurate than X^2_{ML} *p*-values for all non-normal conditions, and more accurate than X^2_{MLM} *p*-values at the smallest samples considered ($N < 200$ in the normal and moderately non-normal conditions, and $N < 500$ in the highly non-normal condition).

Ichikawa and Konishi published a series of articles on bootstrap methods in structural equation models (Ichikawa & Konishi, 1995, 1997, 2001). The first paper considered an unrotated exploratory factor analysis model with $p = 9$ observed variables, normal and elliptical data, and sample sizes of 150 and 300. The Bollen-Stine bootstrap method provided robust rejection rates across conditions with normal data, but not with

elliptical data. The following paper (Ichikawa & Konishi, 1997) included more complex models with independent or just uncorrelated common and unique factors and sample sizes varying between 50 and 800, although the number of indicators was still small (between $p = 6$ and 10). They found that when common and unique factors were independent and sample size was small, the Bollen-Stine bootstrap retained the null hypothesis too often. When common factors and unique factors were just uncorrelated, the bootstrap rejected the null hypothesis too often. While the Bollen-Stine bootstrap performed better than ML overall, in both cases it was not clear whether model complexity affects the performance of the former. Finally, Ichikawa and Konishi (Ichikawa & Konishi, 2001) proposed an efficient bootstrap method to deal with the problem of non-convergence, which had promising results in terms of rejection rates. We will discuss this work in more detail in the discussion section.

Enders (2002) considered the behavior of Bollen-Stine *p*-values with missing data using a very similar model to that of Nevitt and Hancock (Nevitt & Hancock, 2001): a confirmatory factor analysis (CFA) model with three latent variables and three indicators each, with higher factor intercorrelation (.40 as opposed to .30). Conditions were different combinations of three sample sizes ($N = 100$, 250 and 500), two missing data rates (10% and 20%) with missing completely at random (MCAR) pattern and three distributional forms: skewness = 0 and kurtosis = 7, skewness = 2.25 and kurtosis = 7.0, and skewness = 3.25, kurtosis = 20.0. He found that while overall the bootstrap had more accurate results than full-information maximum likelihood, it did not perform well in small sample conditions ($N = 100$), in which case this method is too conservative. This behavior was more pronounced as non-normality and missing data rate increased.

Grønneberg and Foldnes (2018) considered models with $p = 11$ observed variables, samples ranging from $N = 100$ to 900 observations and three levels of non-normality: normal, moderately non-normal (skewness = 1, kurtosis = 7), highly non-normal (skewness = 2, kurtosis = 7). Consistent with Nevitt and Hancock's results, they found that Bollen-Stine bootstrapped *p*-values maintained nominal rates across the board, whereas X^2_{ML} *p*-values over-reject the model in non-normal conditions and that X^2_{MLM} *p*-values over-reject the model in non-normal conditions and small samples ($N = 100$ for moderately non-normal data, and $N \leq 300$ for highly non-normal data).

CHAPTER 4

METHODS

Using maximum likelihood estimation for continuous outcomes as implemented in M*plus* (Muthén & Muthén, 2017), we performed a Monte Carlo simulation study to investigate the performance of *p*-values for the chi-square test of model fit (i.e., the likelihood ratio statistic) across 36 different conditions using three methods: Bollen-Stine bootstrapping using 1,000 bootstrap draws, asymptotic *p*-values under normality assumptions (choice ML in M*plus* nomenclature), and asymptotic *p*-values robust to normality adjusting the chi-square statistic by its mean and variance (choice MLMV in M*plus*, Asparouhov & Muthén, 2010). A total of 1,000 replications were generated for each possible condition. We then compared the empirical 1, 5 and 10% rejection rates of the bootstrap, ML and MLMV to their expected Type I error rates, that is, the proportion of replications in which the model is rejected at the $\alpha = 1$, 5 and 10% levels of significance.

The underlying populational model is a unidimensional CFA model with varying number of indicator variables ($p = 10$, 20, and 30). Population parameter values are such that the factor variance is set to 1.0, the factor mean is set to zero, all factor loadings are set to .70, and all error variances are set to .51 (i.e., $1-.70^2$).

Four sample sizes were included in the study: extremely small (100), small (200), moderate (500) and large (1,000). Three different distributional shapes were created by

varying values of skewness and kurtosis: normal (skewness = 0, kurtosis = 0); moderate non-normal (skewness = 0, kurtosis = 3.3); and severe non-normal (skewness = -2, kurtosis = 3.3). In all cases, we generated multivariate normal data with mean zero and a covariance structure conforming to the population model. Then, the continuous data were discretized into five categories coded 0 to 4. It has been shown that when the number of categories is large, it is appropriate to treat the discretized data as continuous (Bollen, 1989; Dolan, 1994; Muthén & Kaplan, 1985; Savalei & Rhemtulla, 2013) and this practice has been routinely employed by substantive researchers (e.g., Skule et al., 2014). To introduce non-normality, we manipulated the threshold values to create targeted distributional properties for the population (Muthén & Kaplan, 1985). Population skewness and kurtosis were computed as described by Maydeu-Olivares, Coffman, and Hartmann (2007). We provide in Table 4.1 the threshold values used to generate the data. All code is provided in the Appendices section.

Table 4.1

Target item category probabilities and corresponding threshold values used to generate the data

			Expected area under the curve				
Kurtosis	Skewness	Thresholds	0	1	2	3	4
0	0	-1.55, -0.64, 0.64, 1.55	6%	20%	48%	20%	6%
3.3	0	-1.64, -1.04, -1.04, 1.64	5%	15%	60%	15%	5%
3.3	-2.0	-2.05, -1.55, -1.08, -0.52	2%	4%	8%	16%	70%

CHAPTER 5

RESULTS

The quantitative measure of robustness of empirical model rejection rates was the criterion suggested by Bradley (1978). For a properly specified model, the empirical rejection rates will be considered adequate if they range within the interval [.5α, 1.5α]. For α = 1, 5 and 10%, the intervals are [.5, 1.5] %, [2.5, 7.5] % and [5, 15] % respectively. For all simulation conditions, results are summarized in Table 5.1. Model rejection in percentage values are presented for the ML and MLMV statistics as well as for the Bollen and Stine bootstrap *p*-values for α = .01, .05 and .10. Rejection rates falling within the interval criteria are shaded. Results in Table 5.2 show that overall, bootstrapping performs well only in the smallest model considered, $p = 10$, but it is too conservative (i.e., it does not reject enough) for $p = 20$ and 30. In contrast, asymptotic *p*-values computed under normality (i.e., choice ML in M*plus*), are generally too liberal, whereas the asymptotically robust *p*-values (i.e., choice MLMV) computed are marginally better than the ML estimator, but the MLMV method yields accurate *p*-values across most conditions. Results for ML and MLMV are consistent with previous findings in the literature (Ichikawa & Konishi, 1995, 1997; Maydeu-Olivares, 2017b). Results for bootstrap *p*-values are also consistent with previous findings (Enders, 2002; Grønneberg & Foldnes, 2018; Nevitt & Hancock, 2001).

5.1 Bootstrap

In particular, the results for the bootstrap reveal that it provides somewhat better *p*-values when data is normally distributed. First, let us consider the smallest model under study ($p = 10$). accurate *p*-values can be obtained provided the sample size is large enough ($N \geq 200$). At smaller sample sizes it tends to be conservative, leading to the conclusion that the model fits better than it actually does. Sample sizes of at least 500 observations may be needed to obtain accurate *p*-values provided that the data is not too far off from normality. With 1,000 observations, 5 and 10% rejection rates were robust regardless of normality, whereas the 1% rejection rate was always conservative.

Second, in the intermediate size model ($p = 20$), when data is normal at least 500 observations are needed for accurate 5 and 10% rejection rates, and a 1,000 for $\alpha = 1\%$. Finally, at the largest model size considered ($p = 30$), rejection rates were too small across the board, suggesting that the model fits better than what it does. The only exception are 10% rejection rates with at least 500 observations, and even in this case they are borderline small. In summary, assuming the typical choice of $\alpha = .05$, the bootstrap can be used when a) $p = 10$ and data is normal, b) $p = 10$ and $N = 500$ or more, and c) $p = 20$, data is normal and $N = 500$ or more.

5.2 Maximum Likelihood

As expected, the ML estimator had poor performance in most conditions. Given a small model ($p = 10$), data must be normal and sample size should be 200 or greater when the choice of alpha is $\alpha = .05$ or .10. In larger models, rejection rates are at best borderline robust for a few select conditions with large sample size (at least $N = 500$) and normality. When the number of indicators is 20 or 30 and data is severely non-normal

(skewness = -2, kurtosis = 3.3), rejection rates are close to 100% no matter the sample size.

5.3 Maximum Likelihood with Mean and Variance correction

The MLMV test statistic, on the other hand, yielded robust rejection rates for most conditions tested. Rejection rates are mostly within the robust intervals even in the larger models with 20 and 30 observed variables, with the exception of the conditions with the smallest sample size (N = 100). In the more extreme case, with 30 indicators, the smallest sample size and severe non-normality, rejection rates are on average 33.8, 84 and 96.7% instead of the expected 1, 5 and 10%, respectively.

5.4 Comparisons

Further insight into the performance of the Bollen-Stine bootstrap *p*-values relative to the *p*-values obtained using asymptotic methods via the mean and variance corrected chi-square (MLMV) can be obtained by plotting them. We provide in Figure 5.1 a scatter plot of the *p*-values for the condition involving p = 20 observed variables, sample size N = 500, skewness = 0 and kurtosis = 3.3. Clearly, the bootstrap *p*-values are uniformly larger than mean MLMV values, and from Table 5.1 we gather that the rejection rates at α = 1, 5, 10% for MLMV are .5, 4.5, and 9.1% respectively, whereas rejection rates for the bootstrap are .1, 1, and 4.10% respectively. In turn, the comparison of the two asymptotic methods (ML and MLMV) provided in Figure 5.2 clearly shows that on average ML *p*-values tend to be much lower than the mean and variance corrected alternative: rejection rates are 23.3%, 48.7% and 61.4% for the expected α = 1, 5, 10%, respectively.

Using Table 5.2, we can have an in-depth view of the behavior of the *p*-values across all conditions. Even when the model is small ($p = 10$), normality (or lack thereof) and sample size can have a substantial impact on the decision to retain a model or not, no matter the chosen α. If data are normal and sample size is small ($N = 100$), bootstrap *p*-values are on average .08 greater than MLMV *p*-values, and this difference can be as large as .19 for some conditions. If data presents excess kurtosis, this difference goes up to an average of .15 and a maximum of .31. For this small model, the average difference only goes down to the third decimal place when sample size is 1,000 and normality holds – if data is non-normal, the average is .08 and up to a ceiling of .16.

Let us now consider the conditions involving a larger number of observed variables. When the number of indicators is $p = 20$, the average difference in *p*-values has a minimum magnitude of .02, obtained with sample size N = 1,000 and normal data, and a substantially higher sample size is required to obtain similar performance to that of the smaller model. For example, for the same condition exhibited in Figures 5.1 and 5.2, bootstrap *p*-values are on average .12 greater than MLMV *p*-values, although this difference can be as large as .21.

In the largest model tested ($p = 30$), the differences in performance are even more accentuated. The condition with the most extreme differences was the one with sample size $N = 100$, skewness = 0 and kurtosis = 3.3, the average difference is .45 with a maximum of .57; furthermore, the *minimum* difference is of .14. Differences are unsurprisingly less dramatic as sample size increases, although the mean difference between the bootstrapped and MLMV *p*-values is still a substantial .04 with the largest sample size tested and normally distributed data.

For further insight into how *p*-values vary across methods, we used an ANOVA model to investigate the drivers of the *p*-value difference obtained using the bootstrap and robust chi-square methods. Results are summarized in Table 5.3. While neither of the two-way interactions between normality condition and a) number of indicators or b) sample size was significant, the interaction between number of indicators and sample size was significant, $p < .001$. Figures 5.3 and 5.4 are both visualizations of this interaction effect where *p*-values for the Bollen-Stine bootstrap get considerably higher as model size increases and sample size decreases.

Table 5.1

Results for the Chi-Square Test of Model Fit. Empirical Rejection Rates at 1%, 5% and 10% Significance Levels

			Bootstrap			ML			MLMV		
p	*N*	Normal	1%	5%	10%	1%	5%	10%	1%	5%	10%
10	100	1	.20	2.90	7.60	2.30	10.30	17.60	1.70	7.00	13.60
		2	.10	.70	3.20	10.40	26.60	37.90	.90	6.70	15.40
		3	.10	1.90	4.20	70.30	86.20	91.40	1.40	8.30	19.10
	200	1	.60	4.00	8.20	1.70	7.40	12.50	.80	5.70	10.70
		2	.10	1.30	3.90	6.00	20.10	32.20	.60	5.00	10.70
		3	.20	1.90	5.40	66.30	83.70	90.00	.60	5.20	11.50
	500	1	.80	5.40	11.10	1.50	7.10	12.60	1.10	6.10	11.60
		2	.30	2.40	5.60	6.80	18.90	29.50	.70	5.40	11.70
		3	.50	3.60	8.10	62.70	81.90	88.90	1.30	5.70	11.20
	1000	1	1.60	6.50	11.60	2.00	7.20	13.40	1.60	6.50	12.30
		2	.20	3.10	6.30	6.60	19.30	31.50	1.30	6.00	11.50
		3	.40	3.10	7.70	62.30	80.00	87.20	.40	5.30	9.80
20	100	1	.10	.30	1.90	12.40	30.20	42.20	1.10	8.70	21.40
		2	.00	.10	.40	54.50	77.70	86.20	1.10	13.00	28.40
		3	.00	.10	.90	99.90	100.00	100.00	3.40	27.80	52.30
	200	1	.00	.90	3.90	3.70	12.70	21.20	.20	4.90	11.30
		2	.00	.50	2.20	35.60	59.70	72.80	.60	5.70	13.60
		3	.00	.20	2.70	100.00	100.00	100.00	.30	6.30	18.20
	500	1	.30	3.30	9.00	1.80	9.10	14.70	.70	4.80	10.80
		2	.10	1.00	4.10	23.30	48.70	61.40	.50	4.50	9.10
		3	.10	1.30	5.40	99.60	100.00	100.00	.30	3.90	9.60

Table 5.1 (cont.)

Results for the Chi-Square Test of Model Fit. Empirical Rejection Rates at 1%, 5% and 10% Significance Levels

			Bootstrap			ML			MLMV		
p	*N*	Normal	1%	5%	10%	1%	*5%*	10%	1%	5%	10%
30	1000	1	.80	4.20	10.60	2.00	8.10	14.90	1.00	5.10	11.80
		2	.40	1.80	5.10	23.00	46.00	58.50	1.00	5.20	10.70
		3	.40	2.00	6.10	99.80	100.00	100.00	.90	3.80	9.50
	100	1	.00	.10	.10	40.60	66.60	78.10	1.00	11.60	30.90
		2	.00	.00	.00	96.40	99.30	99.90	2.30	30.00	60.50
		3	.00	.00	.00	100.00	100.00	100.00	33.80	84.00	96.70
	200	1	.00	.30	.80	11.30	27.60	39.60	.40	4.50	14.60
		2	.00	.10	.50	80.00	92.50	96.30	.20	7.00	20.30
		3	.00	.00	1.50	100.00	100.00	100.00	1.40	19.50	45.60
	500	1	.30	2.10	5.50	3.80	12.90	20.70	.80	4.30	9.90
		2	.10	.60	2.60	60.40	79.70	87.70	.40	4.10	9.90
		3	.00	1.10	4.30	100.00	100.00	100.00	.30	4.50	13.30
	1000	1	.30	2.00	5.60	1.50	8.40	16.10	.60	3.90	9.50
		2	.00	1.30	3.80	51.10	75.50	85.70	.50	5.00	10.00
		3	.20	1.20	4.40	100.00	100.00	100.00	.30	3.50	9.00

Notes: *p* = number of observed variables, *N* = sample size, Normal = normality condition: 1 → skewness = 0, kurtosis = 0 (normal), 2 → skewness = 0, kurtosis = 3.3, 3 → skewness = -2, kurtosis = 3.3, boots. = Bollen-Stine bootstrap *p*-values, ML = asymptotic *p*-values under normality, MLMV = asymptotic *p*-values for the mean and variance corrected statistic.

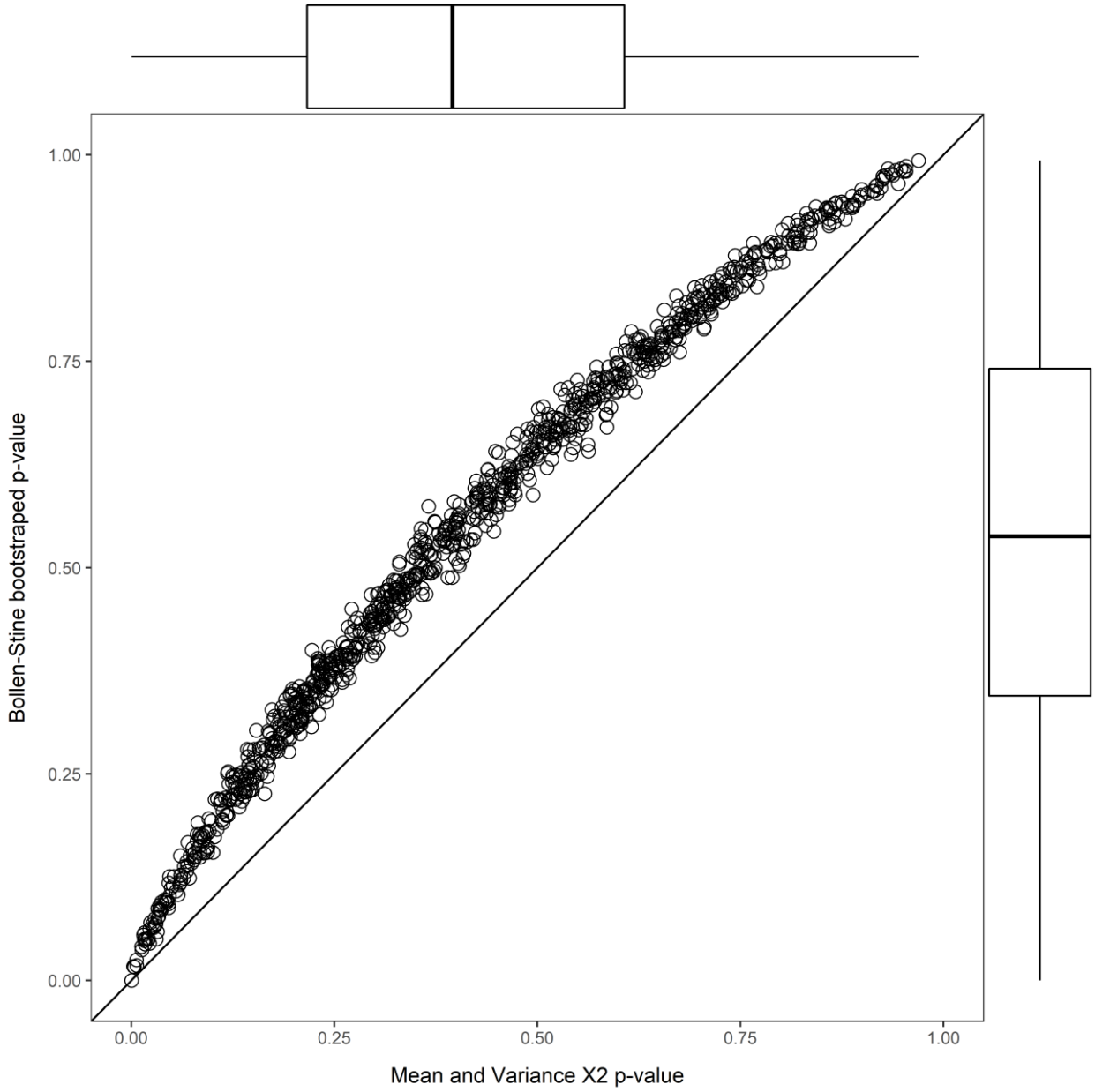

Figure 5.1. *Plot of Bollen-Stine bootstrap vs. MLMV p-values: p = 20, N = 500, skewness = 0 and kurtosis = 3.3*

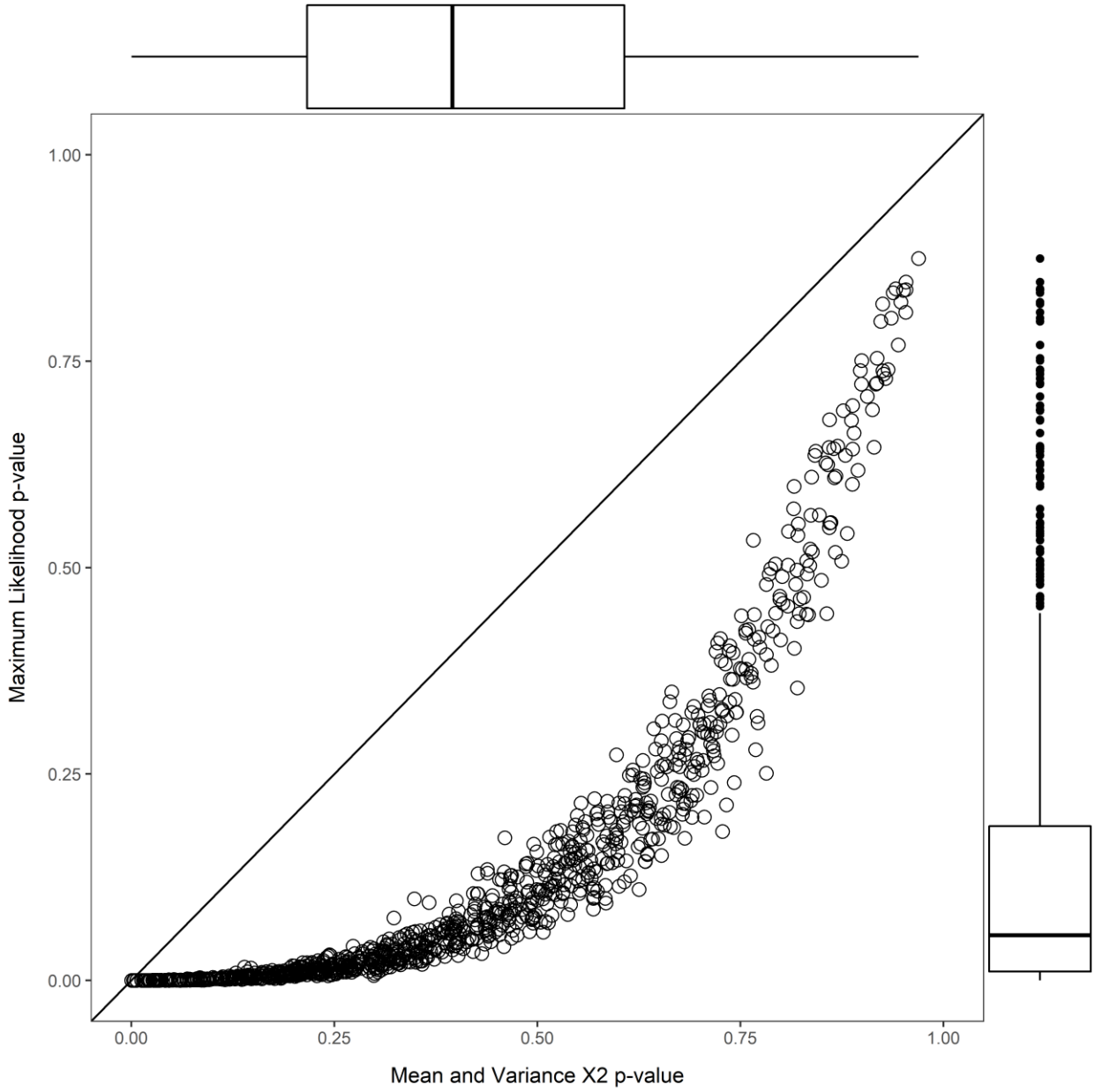

Figure 5.2. *Plot of ML vs. MLMV p-values: p = 20, N = 500, skewness = 0 and kurtosis = 3.3*

Table 5.2

Chi-square p-value difference: Bollen-Stine bootstrap versus MLMV

p	*N*	Normal	Range	Minimum	Maximum	Mean	Standard Deviation
10	100	1	.195	-.010	.185	.077	.033
		2	.306	.006	.312	.146	.048
		3	.300	-.010	.290	.133	.050
	200	1	.176	-.038	.138	.037	.024
		2	.228	.006	.235	.108	.038
		3	.231	-.027	.203	.074	.035
	500	1	.121	-.052	.069	.010	.018
		2	.196	.000	.196	.081	.034
		3	.162	-.039	.123	.040	.026
	1000	1	.093	-.042	.051	.004	.015
		2	.160	-.002	.158	.075	.032
		3	.141	-.030	.112	.034	.022
20	100	1	.324	.007	.331	.234	.055
		2	.391	.032	.423	.290	.066
		3	.442	.037	.479	.302	.079
	200	1	.186	.011	.196	.118	.032
		2	.243	.028	.271	.183	.043
		3	.246	.030	.276	.164	.041
	500	1	.119	-.009	.110	.041	.020
		2	.208	.000	.208	.118	.035
		3	.145	.002	.147	.074	.026
	1000	1	.092	-.023	.069	.021	.016
		2	.183	.000	.183	.099	.035
		3	.121	-.001	.120	.052	.022
30	100	1	.476	.041	.517	.437	.063
		2	.434	.138	.573	.447	.083
		3	.526	.125	.651	.431	.108
	200	1	.296	.028	.324	.217	.043
		2	.327	.026	.353	.259	.053
		3	.336	.047	.383	.244	.064
	500	1	.144	.001	.145	.085	.027
		2	.224	.006	.230	.148	.037
		3	.171	.013	.184	.114	.032
	1000	1	.102	-.001	.102	.043	.017
		2	.187	.003	.190	.115	.034
		3	.137	-.001	.136	.071	.022

Table 5.2 (cont.)

Notes: p = number of observed variables, N = sample size, Normal = normality condition: 1 $\rightarrow$ skewness = 0, kurtosis = 0 (normal), 2 $\rightarrow$ skewness = 0, kurtosis = 3.3, 3 $\rightarrow$ skewness = -2, kurtosis = 3.3, MLMV = mean and variance corrected statistic. All differences were computed so that tabled values are results from the subtraction (bootstrap p-value) – (MLMV p-value).

Table 5.3

Analysis of Variance for Bollen-Stine bootstrap vs. MLMV p-value mean differences

Source	df	SS	Mean Square	*F*-value	*p*-value
Model	23	.498	.022	209.768	.000
Intercept	1	.730	.730	7067.524	.000
p	2	.134	.067	647.907	.000
N	3	.267	.089	863.217	.000
Normal	2	.023	.012	112.324	.000
N : Normal	6	.001	.000	2.399	.093
p : N	6	.071	.012	115.135	.000
p : Normal	4	.001	.000	2.334	.115
Error	12	.001	.000		
Total	36	1.229			
Corrected total	35	.499			

Notes: *p* = number of observed variables, *N* = sample size, Normal = normality condition, df = degrees of freedom, SS = type III sum of squares.

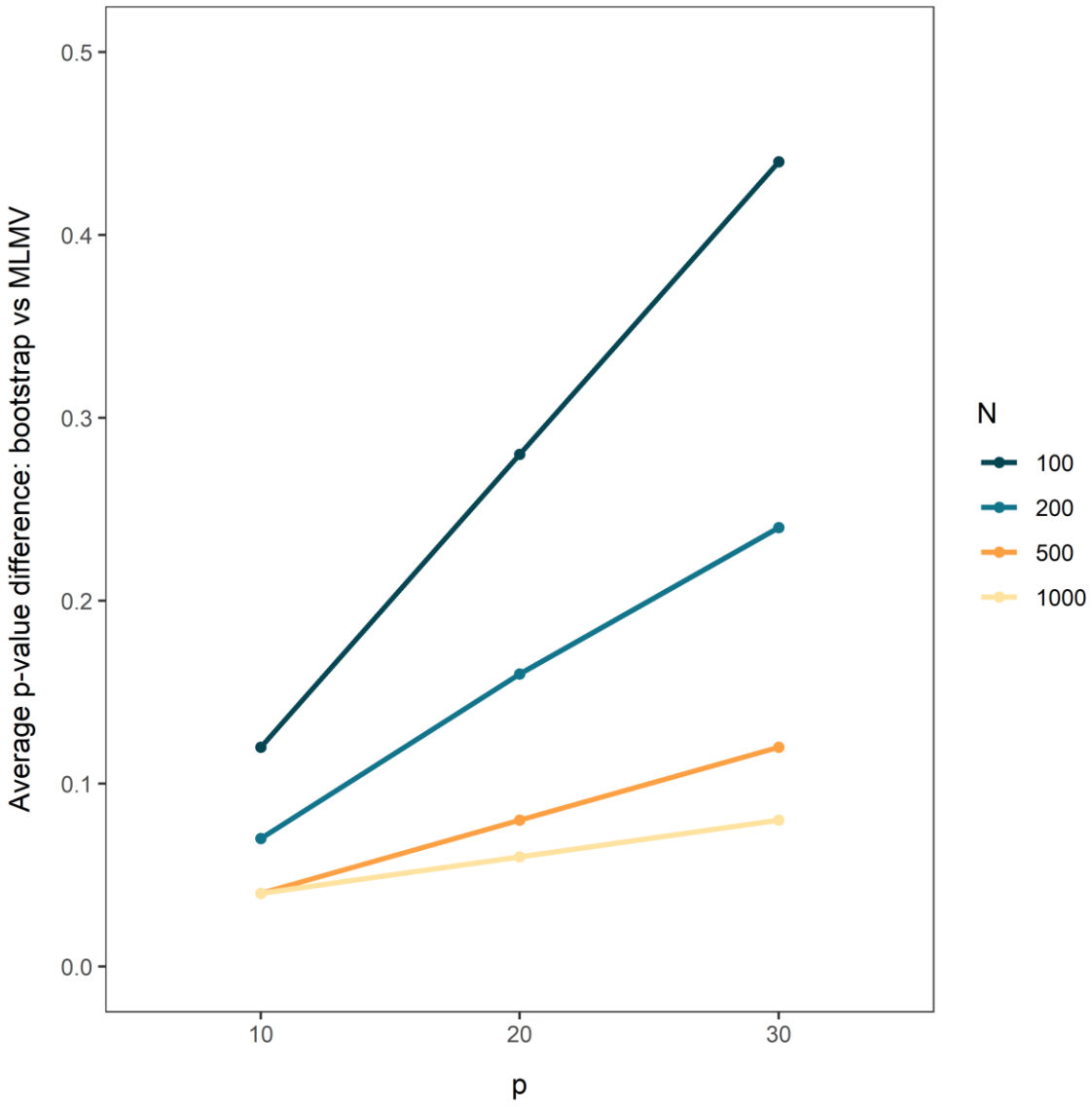

Figure 5.3. *Two-way interactions between sample size and number of indicators*

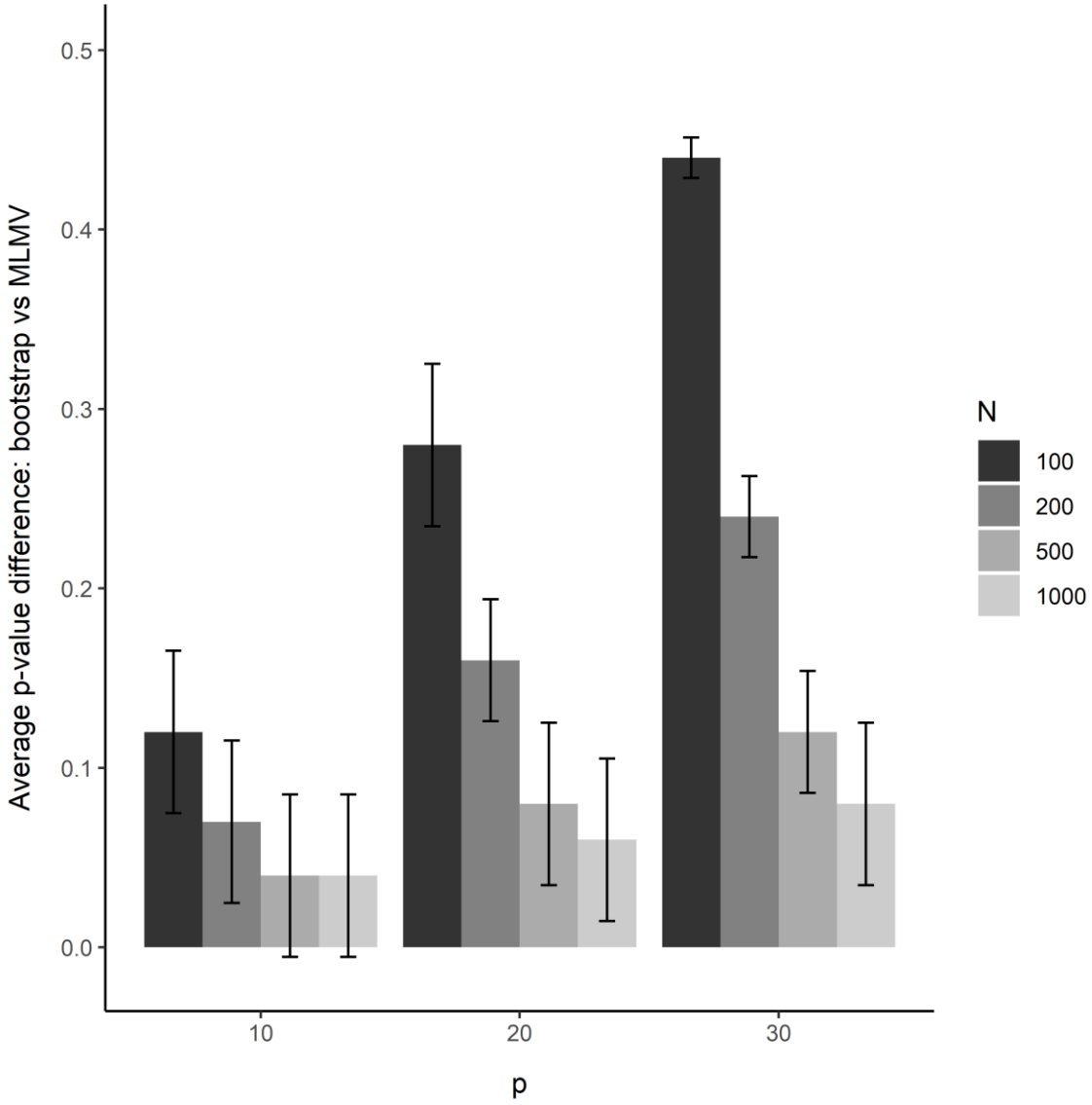

Figure 5.4. *Mean differences in p-values: bootstrap vs. MLMV*

CHAPTER 6

AN EXAMPLE: FITTING AN EXPLORATORY FACTOR ANALYSIS MODEL TO THE RATIONAL PROBLEM-SOLVING SCALE

The Social Problem Solving Inventory-Revised (SPSI-R: D'Zurilla, Nezu, & Maydeu-Olivares, 2002) is the most widely used instrument to assess social problem solving skills, that is, problem solving as it occurs in the natural environment or "real world" (D'Zurilla, Nezu, & Maydeu-Olivares, 2004). It consists of five scales aimed at measuring three different problem-solving styles (rational, impulsive/careless, and avoidant) and two different albeit related orientations towards problems (positive and negative). In turn, rational problem solving, a constructive problem-solving style that is defined as the rational, deliberate, and systematic application of effective problem-solving skills includes four major skills: (a) problem definition and formulation (PDF), (b) generation of alternative solutions (GAS), (c) decision making (DM), and (d) solution implementation and verification (SIV).

To illustrate the effect of the choice of method to obtain *p*-values for the chi-square test of fit, we used a random sample of 200 females from the Spanish normative sample (Maydeu-Olivares et al., 2000). These data consist of 5 items for each subscale (PDF, GAS, DM, and SIV) for a total of 20 rating items. The items are scored in 5 categories and are quite normally distributed: skewness is at most |.3| and (excess) kurtosis is at most |.9|, a single item shows a kurtosis of 1.5.

We fitted an exploratory factor model with four factors to match the theoretical model underlying this scale. The estimated likelihood ratio statistic is $X^2 = 147.43$ on 116 df. The asymptotic *p*-value obtained under normality assumptions is 0.03, whereas the Bollen-Stine bootstrapped *p*-value is 0.25. Our simulation results indicate that the *p*-value obtained under normality is too small, as rejection rates for a similar condition at $\alpha = 1$, 5, 10% are 3.70, 12.70, and 21.20%, respectively. Our simulation results also suggest that the Bollen-Stine *p*-value is too large, as rejection rates for a similar condition at $\alpha = 1$, 5, 10% are <.01, .90, and 3.90%, respectively. The asymptotic *p*-value we obtain using the Asparouhov and Muthén (2010a) mean and variance corrected is 0.21. Our simulation results suggest that this is the most accurate *p*-value for our example as rejection rates for a similar condition at $\alpha = 1$, 5, 10% are .20, 4.90, and 11.30%, respectively. We recognize that the difference between the *p*-values obtained using the mean and variance adjusted statistic and using bootstrap methods is small.

CHAPTER 7

DISCUSSION AND CONCLUSIONS

Bootstrapping has been in use in structural equation models for quite some time. For instance, Chatterjee (1984), Boomsma (1986), Bollen and Stine (Bollen & Stine, 1990) and Yuan and Hayashi (Yuan & Hayashi, 2006) used bootstrap to study standard errors in covariance structure models. Yung and Bentler (1996), Yuan and Hayashi (2003) and Yuan and Marshall (2004) used bootstrap to estimate power and lack of fit in these models.

In particular, bootstrapping procedures (Beran & Srivastava, 1985; Bollen & Stine, 1992; Stine, 1989; Yung & Bentler, 1996) provide an alternative to the use of asymptotic methods for obtaining *p*-values for tests of exact fit in SEM models. Despite having been around for over 25 years, few studies had investigated the performance of these *p*-values and all of the studies focused on small models (up to 11 observed variables). In this article, we have investigated the accuracy of Bollen-Stine bootstrapped *p*-values in larger models (up to 30 observed variables). Consistent with previous studies, we found that Bollen-Stine *p*-values are accurate in small models ($p = 10$); in particular, we highlight the results of Ichikawa and Konishi (1995) who also obtained robust rejection rates across small sample size and normality conditions. In Table 7.1, rejection rates are provided based on the results reported on their Table 5 (e.g., in *R*, we can use the "*pchisq*" function with 19 degrees of freedom). However, as model size increases,

Bollen-Stine *p*-values become conservative (they reject the model less than they should) and that asymptotic *p*-values obtained robust chi-square statistics (in this study, the mean and variance chi-square) are considerably more accurate. Since obtaining a *p*-value for the chi-square test of exact fit using the Bollen-Stine bootstrap is computationally more intensive than using asymptotically robust methods, the latter (i.e., an asymptotic mean and variance correction to the likelihood ratio statistic) is preferred.

Asymptotic methods have another advantage over the use of the Bollen-Stine bootstrap in that they yield a) *p*-values for RMSEA tests of approximate fit, and b) standard errors for parameter estimates a byproduct. In contrast, if bootstrapping is to be performed, naïve bootstrapping is to be performed to obtain standard errors, model-based bootstrapping is to be performed to obtain *p*-values for tests of exact fit, and an alternative bootstrapping is needed to obtain confidence intervals for goodness of fit indices.

7.1 Bootstrap confidence intervals for goodness of fit indices

The earliest attempt to obtain confidence intervals for goodness of fit indices is Bone et al. (1989) who suggested using the bootstrap to compute the standard error of goodness of fit indices (and significance tests) using a fully specified alternative model. Kim and Millsap (Kim & Millsap, 2014) used a somewhat similar procedure. First, they consider an alternative model similar to the fitted model and estimated it using the sample data. Using the estimated covariance matrix under this alternative model, they use the Bollen-Stine method to transform the observed data into the parent sample used for bootstrapping. This procedure enables them to determine how plausible are the observed goodness of fit indices under this alternative model. Thus, Kim and Millsap method

requires that bootstrapping is performed once if a bootstrapped *p*-value for the test of exact fit is desired, and again if a *p*-value for the goodness of fit indices is desired. Of course, if a researcher considers several alternative plausible models, for each of them a bootstrap run is needed using this procedure.

Yuan, Hayashi and Yanagihara (YHY: 2007) added a very useful tool to the bootstrapping arsenal to assess model fit. Whereas naïve bootstrapping involves a parent population with covariance matrix **S** and Bollen-Stine bootstrapping involves a parent population with covariance matrix $\hat{\boldsymbol{\Sigma}}$, their procedure involves transforming the observed data using a covariance matrix that is between **S** and $\hat{\boldsymbol{\Sigma}}$. More specifically, the data is transformed so that the "population noncentrality parameter in the transformed data is equal to the estimated sample noncentrality in the original data" (Zhang & Savalei, 2016). In other words, the covariance matrix of the parent population is chosen so that the model-based bootstrapped CIs for goodness of fit indexes based on non-centrality parameter estimates such as the RMSEA and the CFI (Comparative Fit Index: Bentler, 1990) can be obtained. Zhang and Savalei (2016) performed a simulation study to compare the performance of the naïve, Bollen-Stine and YHY CIs for a number of goodness of fit indices including the RMSEA, CFI, SRMR and GFI (Goodness of Fit Index: Jöreskog & Sörbom, 1988). Both correctly specified and mispecified conditions were included but model size was at most 18 observed variables. As expected, the YHY procedure provided more accurate covarage rates for the RMSEA and the CFI than for the SRMR and GFI, the use of Bollen-Stine procedure yielded unacceptable coverage rates for all conditions. Of particular interest in the Zhang and Savalei (2016) study is the comparison between analytic (i.e., based on asymptotic methods) and YHY bootstrapped

CIs for the RMSEA. They found that there is not a clear advantage of bootstrap CIs over analytic ones.

In closing, in this study we have investigated whether bootstrapped *p*-values are more accurate than analytic *p*-values (obtained using aysmptotic methods) to solve the thorny issue of assessing the exact fit of SEM models when the number of observed variables is large. Within the conditions investigated, analytic *p*-values provide substantially better results than bootstrap *p*-values. Further research is needed to develop alternative bootstrapping schemes that successfully adress this issue. A promising venue of research is the efficient bootstrap method proposed by Ichikawa and Konishi (Ichikawa & Konishi, 2001): by avoiding fitting the model to each bootstrap sample, their procedured is computationally more efficient and avoids problems of convergence in smaller samples. Altough they reported that their bootstrap method accepted the null hypothesis to often, calculating rejection rates from the data they made available shows robust rejection rates across all conditions of sample size (N = 150, 250, 500, 1000) and normality (mixture parameter ε = 0.0, 0.1 and 0.3) with p = 15 observed variables, for a nominal 5% rejection rate. The shorter computational time makes it a particularly interesting method for studying large models.

Table 7.1

Chi-Square Test of Model Fit results in Ichikawa & Konishi (1995): Bollen-Stine p-values

		Rejection rates (nominal *versus* empirical)			
ε	*N*	1%	2.5%	5%	10%
0	150	0.80	2.14	4.51	9.49
0	300	0.85	2.26	4.63	9.61
0.1	150	1.55	4.00	8.17	16.19
0.1	300	2.11	5.14	9.92	18.74
0.3	150	2.41	6.02	11.70	21.92
0.3	300	3.34	7.72	14.16	25.30

Notes: ε = normality condition, where zero represents normal data and the mixture parameter epsilon represents increasing degrees of nonnormality, *N* = sample size. Shaded values indicate robust rejection rates.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*(2), 155–173. https://doi.org/10.1007/BF02294170

Arbuckle, J. L. (2017). *Amos (Version 25.0)*. Chicago: IBM SPSS.

Asparouhov, T., & Muthén, B. (2010a). *Simple Second Order Chi-Square Correction*.

Asparouhov, T., & Muthén, B. O. (2005). Multivariate statistical modeling with survey data. *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*.

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

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. (1995). *EQS 5*. Retrieved from http://files/973/Bentler - EQS 6 Structural Equations Program Manual.pdf

Bentler, Peter 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, Peter M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*(2), 181–197. https://doi.org/10.1207/S15327906Mb340203

Beran, R., & Srivastava, M. S. (1985). Bootstrap Tests and Confidence Regions for Functions of a Covariance Matrix. *The Annals of Statistics*, *13*(1), 95–115. https://doi.org/10.1214/aos/1176346579

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

Bollen, K. A., & Stine, R. (1990). Direct and Indirect Effects: Classical and Bootstrap Estimates of Variability. *Sociological Methodology*, *20*, 115–140. https://doi.org/10.2307/271084

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/https://doi.org/10.1177/0049124192021002004

Bone, P. F., Sharma, S., & Shimp, T. A. (1989). A Bootstrap Procedure for Evaluating Goodness-of-Fit Indices of Structural Equation and Confirmatory Factor Models. *Journal of Marketing Research*, *26*(1), 105. https://doi.org/10.2307/3172673

Boomsma, A. (1986). On the Use of Bootstrap and Jackknife in Covariance Structure Analysis. In *COMPSTAT* (pp. 205–210). https://doi.org/10.1007/978-3-642-46890-2_30

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., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

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. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Chatterjee, S. (1984). Variance estimation in factor analysis: An application of the bootstrap. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 252–262. https://doi.org/10.1111/j.2044-8317.1984.tb00803.x

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite Sampling Properties of the Point Estimates and Confidence Intervals of the RMSEA. *Sociological Methods & Research*, *32*(2), 208–252. https://doi.org/10.1177/0049124103256130

Curran, P., West, S., & Finch, J. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(l), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

D'Zurilla, T. J., Nezu, A. M., & Maydeu-Olivares, A. (2002). *Manual of the Social Problem-Solving Inventory-Revised*. North Tonawanda, NY: Multi-Health Systems, Inc.

D'Zurilla, T. J., Nezu, A. M., & Maydeu-Olivares, A. (2004). Social problem solving: Theory and assessment. In E. C. Chang, T. J. D'Zurilla, & L. J. Sanna (Eds.), *Social*

problem solving: Theory, research, and training (pp. 11–27). https://doi.org/10.1037/10805-000

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309–326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. Retrieved from http://www.jstor.org/stable/2958830

Efron, Bradley. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, Bradley, & Tibshirani, R. (1993). An introduction to the bootstrap. In *Monographs on statistics and applied probability CN - QA276.8 .E3745 1993*. Retrieved from http://files/336/Efron and Tibshirani - 1993 - An introduction to the bootstrap.pdf

Enders, C K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*(4), 352–370.

Enders, Craig K. (2002). Applying the Bollen-Stine bootstrap for goodness-of-fit measures to structural equation models with missing data. *Multivariate Behavioral Research*, *37*(3), 359–377. https://doi.org/10.1207/S15327906MBR3703_3

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56–83. https://doi.org/10.1080/10705519909540119

Foldnes, N., & Olsson, U. H. (2015). Correcting Too Much or Too Little? The Performance of Three Chi-Square Corrections. *Multivariate Behavioral Research*, *50*(5), 533–543. https://doi.org/10.1080/00273171.2015.1036964

Fouladi, R. T. (1998). Covariance Structure Analysis Techniques under Conditions of Multivariate Normality and Nonnormality - Modified and Bootstrap Based Test Statistics. *Annual Meeting of the American Educational Research Association*, 1–22. San Diego, CA, CA.

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, *7*, 356–410.

Grønneberg, S., & Foldnes, N. (2018). Testing Model Fit by Bootstrap Selection. *Structural Equation Modeling*, *00*(00), 1–9. https://doi.org/10.1080/10705511.2018.1503543

Hancock, G. R., & Liu, M. (2012). Bootstrapping Standard Errors and Data-Model Fit Statistics in Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 296–306). London: Guilford Press.

Hartmann, W. M. (2005). *Resampling Methods in Structural Equation Modeling*. (Ml).

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

Hoelter, J. (1983). The Analysis of Covariance Structures: Goodness-of-Fit Indices. *Sociological Methods & Research*, *11*(3), 325–344. https://doi.org/https://doi.org/10.1177/0049124183011003003

Hoyle, R. H. (2012). *Handbook of Structural Equation Modeling* (1st Editio). Retrieved from http://files/499/Hoyle - Handbook of Structural Equation Modeling.pdf

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

Hu, L., Bentler, P. M., & 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

Ichikawa, M., & Konishi, S. (1995). Application of the bootstrap methods in factor analysis. *Psychometrika*, *60*(1), 77–93. https://doi.org/10.1007/BF02294430

Ichikawa, M., & Konishi, S. (1997). Bootstrap Tests for the Goodness of Fit in Factor Analysis. *Behaviormetrika*, *24*(1), 27–38. https://doi.org/10.2333/bhmk.24.27

Ichikawa, M., & Konishi, S. (2001). Efficient Bootstrap Tests for the Goodness of Fit in Covariance Structure Analysis. *Behaviormetrika*, *28*(2), 103–110. https://doi.org/10.2333/bhmk.28.103

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

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

Kim, H., & Millsap, R. (2014). Using the Bollen-Stine Bootstrapping Method for Evaluating Approximate Fit Indices. *Multivariate Behavioral Research*, *49*(6), 581–596. https://doi.org/10.1080/00273171.2014.947352

Linhart, H., & Zucchini, W. (1986). Model selection. In *Model selection*. Retrieved from http://files/984/1986-98758-000.html

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104. https://doi.org/10.1037/1082-989X.7.1.83

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., 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., Rodriguez-Fornells, A., Gomez-Benito, J., D'Zurilla, T. J., Gómez-Benito, J., & D'Zurilla, T. J. (2000). Psychometric properties of the Spanish adaptation of the Social Problem-Solving Inventory-Revised (SPSI-R). *Personality and Individual Differences*, *29*(4), 699–708. https://doi.org/10.1016/S0191-8869(99)00226-3

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

Nevitt, J., & Hancock, G. (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. https://doi.org/10.1207/S15328007SEM0803_2

Nevitt, J., & Hancock, G. R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *Journal of Experimental Education*, *68*(3), 251–268. https://doi.org/10.1080/00220970009600095

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

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.r-project.org

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(1), 1–36. https://doi.org/10.18637/jss.v048.i02

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., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 201–223. https://doi.org/10.1111/j.2044-8317.2012.02049.x

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

Skule, C., Ulleberg, P., Lending, H. D., Berge, T., Egeland, J., Brennen, T., & Landrø, N. I. (2014). Depressive symptoms in people with and without alcohol abuse: Factor structure and measurement invariance of the beck depression inventory (BDI-II) across groups. *PLoS ONE*, *9*(2), e88321. https://doi.org/10.1371/journal.pone.0088321

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

Stine, R. (1989). An Introduction to Bootstrap Methods: Examples and Ideas. *Sociological Methods & Research*, *18*(2–3), 243–291. https://doi.org/10.1177/0049124189018002003

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. https://doi.org/10.4135/9781412956253.n563

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Williams, J., & Mackinnon, D. P. (2008). Resampling and Distribution of the Product Methods for Testing Indirect Effects in Complex Models. *Structural Equation Modeling*, *15*(1), 23–51. https://doi.org/10.1080/10705510701758166

Yuan, K.-H., & Bentler, P. M. (1997). Mean and Covariance Structure Analysis: Theoretical and Practical Improvements. *Journal of the American Statistical Association*, *92*, 767–774. https://doi.org/10.1080/01621459.1997.10474029

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 289–309. https://doi.org/10.1111/j.2044-8317.1998.tb00682.x

Yuan, K.-H., & Bentler, P. M. (1999). 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 nonnormal missing data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *The British Journal of Mathematical and Statistical Psychology*, *56*(Pt 1), 93–110. https://doi.org/10.1348/000711003321645368

Yuan, K.-H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 397–417. https://doi.org/10.1348/000711005X85896

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., & Marshall, L. L. (2004). A new measure of misfit for covariance structure models. *Behaviormetrika*, *31*(1), 67–90. https://doi.org/10.2333/bhmk.31.67

Yung, Y.-F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). Mahwah, NJ, NJ: Erlbaum.

Zhang, X., & Savalei, V. (2016). Bootstrapping Confidence Intervals for Fit Indexes in Structural Equation Modeling. *Structural Equation Modeling, 23*(3), 392–408. https://doi.org/10.1080/10705511.2015.1118692

APPENDIX A

MPLUS CODE FOR DATA GENERATION OF THE FIRST CONDITION

Specifications: $p = 10$, $N = 100$, skewness = 0 and kurtosis = 0. The underlying populational model is a unidimensional CFA model parameter values are such that the factor variance is set to 1.0, the factor mean is set to zero, all factor loadings are set to .70, and all error variances are set to .51.

```
MONTECARLO:
      names=x1-x10;
      generate = x1-x10(4);
      nobservations=100;
      nreps=1000;
      seed=123;
      repsave=all;
      save=C1.*.dat;
MODEL POPULATION:
      f by x1-x10@0.7;
      f@1;
      [f@0];
      x1-x10@0.51;
      [
      x1$1-x10$1*-1.55477
      x1$2-x10$2*-0.643345
      x1$3-x10$3*0.643345
      x1$4-x10$4*1.55477
      ];
```

APPENDIX B

MPLUS CODE FOR ANALYZING THE 1,000 GENERATED DATASETS UNDER CONDITION 1 USING ML OR MLMV

```
DATA:
      file=C1.list.dat;
      type=montecarlo;
VARIABLE:
      names are x1-x10;
ANALYSIS:
      ! alternatively, for MLMV estimation use "ESTIMATOR=MLMV;"
      ESTIMATOR=ML;
MODEL:
      f by x1-x10*.70;
      f@1;
      x1-x10*;
SAVEDATA:
      ! save all results including chi-square statistic p-values
      RESULTS=C.SAV;
```

APPENDIX C

R CODE FOR COMPILING MLMV RESULTS

```
CON=c(rep(NA,36))
one=c(rep(NA,36))
five=c(rep(NA,36))
ten=c(rep(NA,36))
for(i in 1:36){
setwd("C:\\")
library(stringr)

read<-paste0('c',i,'Condition',i,'MLMV.out')
con=file(read)
line=readLines(con)
ten[i]<-unlist(strsplit(line[grep("Chi-Square Test of Model
Fit",line)+19]," "))[unlist(strsplit(line[grep("MODEL FIT
INFORMATION",line)+24]," "))!=''][2]

five[i]<-unlist(strsplit(line[grep("Chi-Square Test of Model
Fit",line)+20]," "))[unlist(strsplit(line[grep("MODEL FIT
INFORMATION",line)+24]," "))!=''][2]

one[i]<-unlist(strsplit(line[grep("Chi-Square Test of Model
Fit",line)+22]," "))[unlist(strsplit(line[grep("MODEL FIT
INFORMATION",line)+24]," "))!=''][2]
CON[i]=i
}

Fresult=data.frame(CON,
                   one,five,ten)
setwd("C:\\ ")
write.table(Fresult, "mlmv.csv",col.names=TRUE,row.names=FALSE, sep =
",")
```

APPENDIX D

R CODE FOR BOOTSTRAP ANALYSES

```
# Compile p-values for each repetition of each condition
SIM10=function(c){
      path=sprintf("C:/ c%d",c)
      setwd(path)
      CHI=c(rep(NA,1000))
      REP=c(rep(NA,1000))
      library(stringr)

# Generate Mplus code for data analysis of models with p = 10 using the
Bollen-Stine bootstrap
for(r in 1:1000){
imp<-paste0('DATA:
             file=C',c,'.',r,'.dat;
             variable:
             names are x1-x10;
             analysis:
             model=nomeanstructure;
             ESTIMATOR=ML;
             BOOTSTRAP = 1000 (RESIDUAL);
             MODEL:
             f by x1-x10*.70;
             f@1;
             x1-x10*;')

write.table(imp, "boot.inp", sep="",row.names=F,col.names = F,quote = F)
batch<-paste0('C:\\Program Files\\Mplus','
             ',
             'Mplus ', path,'\\boot.inp', ' ', path,'\\boot.out','
             EXIT')
```

```
write.table(batch, "BATCH.bat", sep="",row.names=F,col.names = F,quote =
F)
shell ("BATCH.bat")

# Compile p-values
con=file("boot.out")
line=readLines(con)
p<-as.numeric(unlist(as.character(line[grep("Chi-Square  Test  of  Model
Fit",line)+5])%>% str_match_all("[0-9.^-]+"))[2])

CHI[r]=p
REP[r]=r
}

result=data.frame(REP,CHI)
write.table(result, "bootresult.csv",col.names=TRUE,row.names=FALSE, sep
= ",")
}

#Execute condition 1
SIM10(c=1)
```

APPENDIX E

R CODE FOR COMPILING BOOTSTRAP RESULTS

```
con=c(rep(NA,36))
one=c(rep(NA,36))
five=c(rep(NA,36))
ten=c(rep(NA,36))

for(i in 1:36){

  path=sprintf("C: /c%d",i) #define path using condition number
  setwd(path) # Specify the dictionary

  data<-read.csv("bootresult.csv")

  one[i]=mean(data$CHI<0.01)
  five[i]=mean(data$CHI<0.05)
  ten[i]=mean(data$CHI<0.10)
  con[i]=i
}

Fresult=data.frame(con,
                   one,five,ten)

setwd("C:/ ")
write.table(Fresult,
"all.bootresults.csv",col.names=TRUE,row.names=FALSE, sep = ",")
```

APPENDIX F

R (VERSION 3.6.3) CODE AND SPSS (VERSION 25) SYNTAX FOR MULTIWAY ANOVA OF P-VALUE DIFFERENCES

```
# Three-way Factorial Design using Table 5.2 data
options(contrasts=c("contr.sum", "contr.poly"))
fit <- aov(mean ~ p + N + normal + N:normal + N:p + normal:p, data=table3)
drop1(fit,~.,test="F") # type III SS and F Tests
```

```
UNIANOVA mean BY p N normal
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/POSTHOC=p N normal(TUKEY)
/PLOT=PROFILE(p*N)
/CRITERIA=ALPHA(0.05)
/DESIGN=p N normal N*normal N*p normal*p.
```