Assessing cutoff values of SEM fit indices: Advantages of the unbiased SRMR index and its cutoff criterion based on communality

Carmen Ximénez[1]

Alberto Maydeu-Olivares[2, 3]

Dexin Shi[2]

Javier Revuelta[1]

[1] *Department of Psychology*, Autonoma University of Madrid

[2] *Department of Psychology*, University of South Carolina

[3] *Department of Psychology*, University of Barcelona

Author note

Carmen Ximénez. https://orcid.org/0000-0003-1337-6309
Alberto Maydeu-Olivares. http://orcid.org/0000-0001-5790-392X
Dexin Shi. http://orcid.org/0000-0002-4120-6756
Javier Revuelta. https://orcid.org/0000-0003-4705-6282

Correspondence concerning this article should be addressed to Carmen Ximénez (e-mail: carmen.ximenez@uam.es). This work was supported by the National Science Foundation under Grant No. SES-1659936, by the Research Center for Child Well-Being (NIGMS P20GM130420). and by a *Salvador de Madariaga* Grant No. PRX18/00297, and Grant No. PGC2018-093838-B-I00 from the Spanish Ministerio de Ciencia, Innovación y Universidades. We thank the Research Computing Center at the University of South Carolina for providing the computing resources that contributed to the results of this paper.

We have no known conflict of interest to disclose.

Abstract

Holding model misspecification constant, the behavior of fit indices depends on factors such as the number of variables being modeled (model size), and the average observed correlation (magnitude of factor loadings or measurement quality). We examine by simulation the interplay of these factors with sample size in CFA models. When a biased estimator of the fit index is used (CFI, TLI, or GFI), the behavior of the sample indices depends on sample size, rendering establishing cutoff values impossible. When an unbiased estimator is used (SRMR, or RMSEA) the behavior of the indices matches that of the population parameter and depends on the average R^2 of the observed variables (communality); and for the RMSEA, also on model size. The use of the unbiased SRMR with a cutoff value adjusted by R^2 is recommended as it enables assessing the degree of a model misspecification across model size, sample size, and measurement quality.

Keywords: Structural equation modeling (SEM), goodness-of-fit indices, magnitude of factor loadings, reliability paradox.

One of the main issues applied researchers have to deal with when applying structural equation modeling (SEM) is the assessment of the goodness of fit of the proposed model. Whether a SEM model fits the data exactly can be assessed using the chi-square test. In applications, however, the test often rejects the model of interest. At this point, applied researchers turn to goodness-of-fit indices to determine whether the model may be retained, or should be rejected. Goodness-of-fit indices are sample statistics that summarize the discrepancy between the observed and expected values for the means and covariance matrix under the SEM model. The model is retained or rejected if these sample statistics exceed some cutoff criteria provided in the sample (e.g., Hu & Bentler, 1999). Thus, sample goodness-of-fit indices are used as if they were goodness-of-fit statistics (Marsh et al., 2004) while ignoring their sampling variability (Maydeu-Olivares, 2017).

An alternative view, more consistent with current practices elsewhere in statistics, is to consider the effect size of the misfit after the model is rejected by the test of exact fit (Maydeu-Olivares, 2017; Saris, Satorra, & van der Veld, 2009). Effect sizes of model misfit are population parameters that capture the discrepancy between the fitted model and the data generating process, and the decision on whether to retain or reject a model is based not solely on the result on the test of exact fit but also on the magnitude of the population parameter and its confidence interval (Maydeu-Olivares, 2017). Various forms of effect sizes exist in the SEM literature, including measures which are unstandardized (e.g., the root mean squared error of approximation, or RMSEA: Browne & Cudeck, 1993, Steiger, 1990); standardized (e.g., the standardized root mean squared residual, or SRMR: Jöreskog & Sörbom, 1988); or relative (e.g., the goodness-of-fit index, or GFI, Jöreskog & Sörbom, 1988; Maiti & Mukherjee, 1990; the comparative fit index, or CFI: Bentler, 1990; and the Tucker-Lewis index or TLI: Tucker & Lewis, 1973). These approaches can be reconciled by considering sample goodness-of-fit indices as estimators (not necessarily optimal) of population parameters (effect sizes of model misfit or ‘population goodness-of-fit indices’), and in this article we will refer to them as population and sample goodness-of-fit indices, respectively.

Current goodness-of-fit indices standards (e.g. Hu & Bentler, 1999) involve the use of constant cutoff criteria derived using sample goodness-of-fit indices. In other words, the same cutoff is applied regardless of the model fitted (Marsh et al. 2004; Barrett, 2007), and the cutoffs were derived via simulation without taking into account sampling variability (Maydeu-Olivares, 2017). However, previous research has shown that holding model misfit constant, the behavior of population (Savalei, 2012; Shi, Lee, & Maydeu-Olivares, 2019; Shi, Maydeu-Olivares, & DiStefano, 2018), and sample goodness-of-fit indices (Anderson & Gerbing, 1984; Fan & Sivo, 2005; Marsh et al., 1988) is not constant. Rather, it depends on factors extraneous to the degree of model misspecification. Two such main factors have emerged in the literature: a) the number of observed variables being modeled, aka model size (Anderson & Gerbing, 1984; Herzog et al., 2007; Jackson, 2003; Kenny & McCoach, 2003; Moshagen, 2012; Shi, Lee, & Terry, 2015; Shi et al., 2019), and b) the magnitude of factor loadings, precision of measurement, and average correlation among the observed variables. Recent research has examined in detail the effect of model size on the validity of the cutoff values for close fit. However, to our knowledge, only Shi et al. (2018; 2019) have considered the effect of the magnitude of factor loadings on the validity of the cutoff values for close fit. Shi et al. (2018) examined the effect of the magnitude of the factor loadings only on SRMR population values, whereas Shi et al. (2019) examined the effect of factor loadings both at the population and sample levels, but only included the CFI, TLI, and RMSEA indices.

Shi et al. (2019) found that model size had an important impact on the population values of CFI, TLI, and RMSEA, and when fitting large SEM models, a disagreement between the sample goodness-of-fit indices and their population counterparts would likely be observed, particularly when the quality of measurement was poor (i.e., low factor loadings). They also noted that the phenomenon of the *reliability paradox* (i.e., poor measurement quality associated with better model fit) might have operated both for RMSEA population values and their sample estimates. Concerning the study by Shi et al. (2018), they found that the larger the average factor loading, the larger the model misspecification as measured by the population SRMR, and that the phenomenon of the *reliability paradox* might also operate for the SRMR population values. Furthermore, they found that the effect of factor loadings on the population SRMR was best captured by considering the average R^2 of the observed variables (communality in factor analysis models) and that a cutoff criterion in terms of $\text{SRMR} / \overline{R}^2$ effectively separates models with a substantially ignorable misspecification (SIM) from non-SIM models, and avoids the phenomenon of the *reliability paradox*. As noted above, Shi et al. (2019) found a sizable effect of the magnitude of factor loadings on the performance of fit indices. However, the emphasis of their study was on the effect of model size (i.e., the number of variables being modeled), and the effect of the magnitude of factor loadings was not analyzed in detail.

The present study aims at filling both gaps. We performed an extensive simulation study to investigate a) the effect of the magnitude of factor loadings on various goodness-of-fit indices both in the population and in finite samples, b) the adequacy in finite samples of using an SRMR cutoff criterion adjusted by the average R^2 of the observed variables (communality in factor analysis models), and the potential advantages of the SRMR correction over other fit indices, as it is the only index that addresses the phenomenon of the *reliability paradox*.

The present manuscript goes beyond Shi et al. (2019) in that we include two important goodness-of-fit indices that were not investigated in their study: the SRMR and the GFI. More specifically, our simulation includes two indices (RMSEA and SRMR) for which unbiased estimates are implemented in SEM software, and three indices in which biased estimators are currently in use (CFI, TLI, and GFI)[1], and we are interested in assessing whether sample size moderates the effect of factor loadings on the average behavior of sample indices if a biased estimator is used. Put differently, we conjecture that the behavior of the sample indices will closely match that of their population counterparts when an unbiased estimator is used (SRMR or RMSEA), whereas when a biased estimator is used (biased SRMR, CFI, TLI, and GFI), sample size will be a major driver of the behavior of the index in small samples rendering any cutoff established at the population level (or using large samples) useless. The present article also goes beyond Shi et al. (2018) in that it verifies whether a cutoff criterion established at the population level that depends on the strength of the associations among the observed variables (i.e., the magnitude of factor loadings in factor analysis models) may be effective in applications.

The remainder of this article is organized as follows. First, we briefly describe the goodness-of-fit indices used and we review the literature on the effect of factor loadings on goodness-of-fit indices. Next, we describe the design of our simulation study in which we manipulate model size, model misspecification, magnitude of factor loadings, and sample size. We then summarize the results, evaluate our sample size × factor loading effect × biased estimator hypothesis, and evaluate the adequacy of the SRMR adjusted by communality cut-off criterion proposed by Shi et al. (2018) in finite samples. We conclude with a general discussion of the results and their practical implications for applied researchers.

Effect of the magnitude of factor loadings on goodness-of-fit indices

[1] For good measure, we also include the biased SRMR in our study.

The present study explores the effect of the magnitude of the factor loadings on the estimated SRMR, RMSEA, CFI, TLI, and GFI goodness-of-fit indices. Table 1 provides a brief description of these indices and their formulas both at the population and at the sample levels.

----- INSERT TABLE 1 ABOUT HERE -----

The current estimators (see Table 1) for the SRMR, CFI, TLI, and GFI are consistent (i.e., they converge to the population parameter as sample size increases) but they are not asymptotically unbiased (in large samples the mean of the sampling distribution of the estimator does not coincide with the parameter). It is important to use unbiased estimators because the sample goodness-of-fit indices can be severely biased at small to moderate sample sizes (Steiger, 1990). Confidence intervals are also highly valuable in small samples because the point estimates may display a large sampling variability.

Arguably, the most widely used goodness-of-fit index is the RMSEA. The popularity of this index stems from being first defined at the population level (as an effect size of misfit) and defining the sample RMSEA as an approximately unbiased estimate of the population parameter (see Browne & Cudeck, 1993). Also, a confidence interval for the population parameter, and if desired, a statistical test of close fit ($H_0\text{: RMSEA} \leq c$) can be easily obtained. However, the use of the population RMSEA parameter itself is problematic, as its magnitude is impossible to interpret because it is in an unstandardized metric (adjusted by degrees of freedom). Researchers cannot really judge whether any given value of the RMSEA (say 0.05) is large or small (Savalei, 2012).

The interpretation problems of unstandardized effect sizes such as the RMSEA can be overcome using standardized effect such as the SRMR, which can be approximately interpreted as the average residual correlation. The sample SRMR is a consistent estimator of the population parameter but it is not asymptotically unbiased. Recently, Maydeu-Olivares (2017) derived an unbiased estimator of the population SRMR that is implemented in lavaan (Rosseel, 2012). We will include this unbiased estimator of the SRMR in our study to gauge the effect of using unbiased estimators of a population goodness-of-fit index parameter vs. simply a consistent estimator (e.g., the sample SRMR).

The interpretation problem of the population RMSEA can be alternatively overcome if a relative fit index is used. The fitted model can be compared to an independence model (CFI) or to a saturated model (GFI). Finally, the widely popular TLI index incorporates features of the CFI (fit relative to an independence model) and RMSEA (adjustment for model parsimony). The current estimators for these goodness-of-fit indices are consistent (see Bentler, 1990) but they are not unbiased. For instance, the sample CFI proposed by Bentler (1990) - also known as McDonald and Marsh's (1990) relative non-centrality index (RNI), see Table 1 - involves an unbiased estimator of its numerator and its denominator, but the estimate of the ratio that conforms the CFI is not itself unbiased. An unbiased estimator of the GFI exists (Maiti & Mukherjee, 1990: eq. 19) but it is not implemented in SEM software and will not be considered here. Recently, Lai (2019) has proposed two new consistent estimators for the CFI. However, they are not implemented in SEM software and they do not appear to outperform the current formula (Shi, DiStefano, Maydeu-Olivares, & Lee, 2021).

Confidence intervals can be obtained for the SRMR (Maydeu-Olivares, 2017), and the CFI (Lai, 2019) and in principle, confidence intervals for the GFI and TLI can also be obtained but we are not aware of any proposal. Nevertheless, the focus of the present study is on the average behavior of point estimates and the influence of measurement precision (magnitude of factor loadings).

In CFA models, larger factor loadings lead to larger correlations among the observed variables, and to increased precision of measurement. As a result, a review of the literature of the impact of the magnitude of factor loadings on the behavior of goodness-of-fit indices must also include literature on the effects of the size of the observed correlations and/or precision of measurement on goodness-of-fit indices. For instance, Steiger (2000) considered a simple path model and found that RMSEA was sensitive to the size of the correlation between the observed variables, controlling for the size of the misspecification. Saris et al. (2009) reported a similar finding in path analysis models (see also Cole & Preacher, 2014).

Hancock and Mueller (2011; see also McNeish et al., 2018) introduce the term *reliability paradox* to denote the finding that "models with poorer quality measurement appear to have better data–model fit, whereas models with better quality measurement appear to have worse data–model fit " (p. 306) illustrating how, holding other factors constant, population values of goodness-of-fit indices (GFI and CFI) decrease as standardized loadings increase, while population values of badness-of-fit[2] indices (SRMR and RMSEA) increase. Browne et al. (2002) gave a statistical explanation of why this issue takes place: the power of likelihood ratio test statistic depends on the eigenvalues of the model implied covariance matrix, which in turns depends on the variances of model errors. The smaller the variance of the model errors, the higher the power of the statistic. In factor analysis models, the errors' variances (uniquenesses) depend on the magnitude of the factor loadings. This explains the dependency of the RMSEA to factor loadings' magnitude. Heene et al. (2011) extended this work and showed that any test statistic (or fit index) based on the difference between the observed and implied covariance matrix (e.g., the SRMR) will also depend on the magnitude of the factor loadings. This could also explain the behavior of incremental fit indices

[2] Smaller values of the CFI and GFI indicate poorer fit, whereas larger values of the RMSEA and SRMR indicate poorer fit.

(e.g. CFI, TLI, GFI): decreasing (standardized) factor loadings lead to decreasing values on the indices "because the difference between a theoretical model and the null model becomes smaller (p. 330)".

Monte Carlo Simulation Study

The guidelines for Monte Carlo simulation designs in SEM recommended by Skrondal (2000) and Boomsma (2013) are used to present the design of the simulation study.

Step 1: Research question and theoretical framework

This study explores the effect of the magnitude of the factor loadings (λ) on the estimated SRMR, RMSEA, CFA, CFI, TLI, and GFI goodness-of-fit indices both in correctly specified and misspecified models. The design of the study includes varying conditions of model size (p) and sample size (N). The sample RMSEA is an approximately unbiased estimator of the population parameter (Browne & Cudeck, 1993). The sample estimators of the CFI, TLI, and GFI currently in use are consistent but not unbiased estimators, whereas for the SRMR we will use both the naïve (consistent but biased) sample estimator of the SRMR ($SRMR_b$) currently implemented in most SEM software packages, and its unbiased estimator ($SRMR_u$) to better illustrate the effect of using biased vs. unbiased estimators.

The study is a follow-up of previous research on the performance of the unbiased SRMR goodness-of-fit index and other commonly used SEM fit indices (Shi et al., 2018; 2019), but focuses specifically on how the magnitude of the factor loadings affects the validity of cutoff values for close fit. The research questions and hypotheses examined are: First, we expect all population fit indices to be impacted by the magnitude of the factor loadings and we aim to answer questions such as: If the factor loadings are high or low, what is the best goodness-of-fit index to be used? Do I need a specific sample size in my study? Are there other characteristics of the model (e.g., model size) that will affect the decision on the election of the goodness-of-fit index? etc. Second, in finite samples, we expect the effect of sample size to be smaller for the unbiased fit indices (SRMR$_u$ and RMSEA). Finally, we will evaluate whether the Shi et al.'s (2018) correction for the SRMR index based on the communality level works reasonably well at the sample level, and could be used with other fit indices.

Step 2: Experimental design

Population models

Following Boomsma's (2013) recommendations, the choice of the population models is based on previous research to increase the comparability of the experimental results and contribute to their external validity. The generating models follow Shi et al. (2018; 2019). More specifically, the population model is a CFA model with two correlated factors in which each item depends on a single factor.

Experimental factors and response variables

The independent variables are the number of variables in the model (*p*), the degree of model misspecification (ρ), the magnitude of factor loadings (λ), and sample size (*N*). Table 2 summarizes the variables used in our design.

----- INSERT TABLE 2 ABOUT HERE -----

The number of observed variables included in the model were $p = 12, 36, \text{and } 72$; representing small, medium, and large model sizes frequently encountered in psychological research. The same number of indicators per factor was used. Model misspecification was introduced by ignoring the multidimensionality of the population model and fitting a single-factor model to the simulated data. The level of model misspecification was manipulated by changing the magnitude of the correlation between the factors in the data generating model (ρ =.70, .80, .90, and 1.00). The smallest inter-factor correlation (ρ = .70) indicates the greatest level of misspecification considered; a correlation of ρ=.80 indicates a medium level of misspecification, a correlation of ρ=.90 a small or substantively ignorable misspecification (SIM: Shi et al. 2018), and a correlation of ρ=1.00 corresponds to no model misspecification.

The magnitude of factor loadings was specified using three levels: λ=.30, .60, and .90; representing weak, medium, and strong factor loadings, respectively (Briggs & Maccallum, 2003; Ximénez, 2006, 2009). When generating the data, the factor loading values used were the same for all variables across factors. The variances of the error terms were set as $1 - \lambda^2$.

Sample sizes included *N* = 100, 200, 500, and 1000 observations representing small, medium, large, and very large sample sizes. A wide range of sample sizes was used to determine the effect of the magnitude of factor loadings under different conditions of sample size, and to give practical recommendations to researchers on which sample sizes to use to warrant an adequate assessment of the goodness of fit of their models.

In summary, the number of conditions examined was 144 = 3 (model sizes) x 4 (degree of model misspecification) x 3 (factor loading levels) x 4 (sample size levels).

Step 3: Estimation and replication

For each condition, 1,000 replications were generated with the simsem package in R (Pornprasertmanit et al., 2013). Data were generated from a multivariate normal distribution. Maximum likelihood (ML) estimates of the model parameters and goodness-of-fit indices were computed with the *lavaan* package in R (R Development Core Team, 2019; Rosseel, 2012). Population values of the goodness-of-fit indices were obtained by fitting the estimated model (also by ML) using *lavaan* to the population covariance matrices implied by the different simulated conditions.

Step 4: Analyses of output

Analyses of variance (ANOVAs) were conducted using as dependent variable the population values as well as sample estimates of each goodness-of-fit index, and using the simulation conditions as the independent variables. Following Skrondal (2000), a simple meta-model was used to analyze the results, which included only the main effects and the double interaction effects of each independent variable on the dependent variable. The explained variance associated with each of the effects was measured using the partial eta-squared statistic (η^2). The magnitude of the main effects and double interactions was judged relevant if the partial eta-squared statistic was larger than .10. Multiple comparisons were also conducted for the effects that were shown to be statistically and practically relevant.

Results

The population values of the fit indices and the average sample estimates of the simulation under the study conditions are provided as supplementary material. We computed correlations among the different sample fit indices to investigate whether they might behave similarly across levels of model misspecification and factor loading. The estimated correlations are reported in Table 3. As we can see in this table, the patterns of correlations reveal two major clusters of fit indices. On the one hand, SRMR_u and RMSEA, the unbiased indices, behave similarly (their inter-correlations are high). The other cluster contains the SRMR_b, CFI, TLI, and GFI, which are the biased fit indices. However, these indices only appear highly correlated when the factor loadings are low. As the magnitude of the factor loadings increase, their inter-correlations depend on the degree of model misspecification.

----- INSERT TABLE 3 ABOUT HERE -----

Figures 1 and 2 show the average sample estimates of each fit index against factor loading levels (λ = .30, .60, and .90), sample size (N = 100, 200, 500, and 1000), model size (p = 12, 36, and 72), and model misspecification (ρ= 1.00 no misspecification, ρ= 0.90 substantially ignorable misspecification or SIM, ρ= 0.80 medium misspecification, and ρ = 0.70 large misspecification). These graphs also include the corresponding population value (solid line). A horizontal red line has been drawn in these figures to mark the recommended cutoff values for SRMR (.08), RMSEA (.05), CFI (.95), TLI (.95), and CFI (.95) suggested by Hu and Bentler (1999).

----- INSERT FIGURE 1 ABOUT HERE -----

----- INSERT FIGURE 2 ABOUT HERE -----

SRMR results

The results from the ANOVA showed that the magnitude of factor loadings, model misspecification, and their interaction have the largest effects on the population SRMR (η^2 = 1.00 in all three cases). The pattern of the solid lines in the graphs of Figure 1a indicate that the population SRMR value decreases as factor loading also decreases, and that this effect is more pronounced as the level of model misspecification increases. At the population level, the use of Hu and Bentler's cutoff (SRMR < .08) will lead us to conclude that all conditions provide a close fit to the estimated model except for models severely misspecified (with ρ = .70), and with very large factor loadings (λ = .90).

As seen in the graphs of Figure 1a, the behavior of the unbiased SRMR (SRMR$_u$) matches very well the behavior of the population SRMR. In fact, the largest discrepancy between the average behavior of the SRMR$_u$ and the population SRMR occurs when the model is exactly specified, the number of observed variables is small, and sample size is also small. In the SIM condition (a one-factor model fitted to two-factors model data with ρ = .90), as the factor loading increased from .30 to .60 and to .90 both the population and the sample SRMR_u increased from 0 to .02, and to .04. In contrast, when the two-factor data were generated using $\rho = .80$, both the population and sample SRMR_u increased from .01 to .04, and to .08. Finally, when the two-factor data were generated using $\rho = .70$, both the population and sample SRMR_u increased from 0 to .05, and to .13.

The pattern of results for the sample SRMR_b index is quite different (see the graphs in Figure 2a). The results from the ANOVA showed that sample size is a main driver of the behavior of this index ($\eta^2 = .99$), and as sample size increases, it provides an increasingly misleading impression of poor fit (average value of sample index is higher than population value). Interestingly, the effect of sample size is hampered by the magnitude of the factor loadings, as the smaller the factor loading, the larger the effect of sample size on the behavior of this index. Thus, the discrepancy between the average value of the SRMR_b and its population counterpart suggest that the fit is poorest when small sample sizes ($N = 100$ or 200) and weak factor loadings ($\lambda = .30$) are employed, but all conditions show acceptable goodness-of-fit values according to Hu and Bentler's cutoff (SRMR < .08). However, for the conditions of medium and large model misspecification the sample results are more similar to their population counterparts, indicating that the goodness of fit worsens as the magnitude of factor loadings increases and that sample size exerts an effect. For example, for small samples ($N = 100$), fit is unacceptable regardless of the level of the factor loading, and this effect is more pronounced for models including a larger number of indicators.

In summary, results indicate that the unbiased estimator of the SRMR (SRMR_u) behaves better than the naïve estimator (SRMR_b), and leads to different conclusions in terms of model fit. The population SRMR can be accurately approximated by the SRMR_u index and suggests an acceptable fit ($\mathrm{SRMR}_u < .08$) in all the study conditions except when the model misspecification is severe (i.e., when collapsing two factors with $\rho = .70$), and the magnitude of the factor loadings is very large (i.e., $\lambda = .90$). In contrast, the average behavior of the biased index (SRMR$_b$) differs from its population counterpart across all conditions. For example, under correctly specified models, when the factor loadings were $\lambda = .30$ and the sample size $N = 100$, the average estimates of SRMR$_b$ changed from .00 (perfect fit) in the population, to .07 (close fit) in the sample, and as p increased from 12 to 72 variables the average estimates of SRMR$_b$ changed from .07 (close fit) to .09 (poor fit). Additionally, as expected, sample size is a main driver of the behavior of the SRMR$_b$ index whereas the SRMR$_u$ index is not affected by the number of observations in the sample.

RMSEA results

At the population level, the largest effects found in the ANOVA for RMSEA are the magnitude of factor loadings (λ, $\eta^2 = .95$), model misspecification (ρ, $\eta^2 = .90$), and their interaction (ρ x λ, $\eta^2 = .88$). The population RMSEA value increases as the level of model misspecification also increases, but this effect depends on the magnitude of the factor loadings. At the population level, the use of Hu and Bentler's cutoff (RMSEA < .05) will lead us to conclude that all conditions provide a close fit to the estimated model when $\lambda = .30$. However, the RMSEA suggests that fit is much poorer (values between .09 and .17) when the model is misspecified and the magnitude of the factor loadings is very large ($\lambda = .90$).

At the sample level, the results from the ANOVA showed that the main drivers of the behavior of the RMSEA index are also the magnitude of the factor loadings (λ, $\eta^2 = .93$), model misspecification (ρ, $\eta^2 = .85$), and their interaction (ρxλ, $\eta^2 = .84$). The graphs of Figure 1b show that under the three conditions of model misspecification, the estimated RMSEA values suggest close fit (RMSEA < .05) when $\lambda = .30$ and .60, in all cases except for $N = 100$. However, for $\lambda =$

.90 fit is very poor for all sample sizes and worsens as model size decreases. Our finding of RMSEA decreasing as the number of indicators per factor increases is congruent with Kenny and McCoach's (2003), and Savalei's (2012) studies. The graphs of Figure 1b also make it clear that as p increases, the difference between the population RMSEA and the sample average values becomes larger. For example, under correctly specified models with $N = 100$ and $\lambda = .30$ the difference between the population RMSEA and the sample average RMSEA increased from .03 (close fit when $p = 12$) to .08 (poor fit when $p = 72$).

In summary, and as with the unbiased SRMR index, our results indicate that the population RMSEA can be accurately approximated by the sample RMSEA index across all study conditions and suggest that the fitted model provides a close approximation to all data generating models employed except when $\lambda = .90$ and there is model misspecification by altering the dimensionality of the model (i.e., when collapsing two factors with ρ values between .70 and .90). However, the RMSEA index is affected by model size whereas the SRMR goodness-of-fit indices are not. In particular, holding other factors constant, the RMSEA decreases as the number of variables in the model decreases (see also Savalei, 2012).

CFI results

At the population level, the largest effects found in the ANOVA for CFI are the magnitude of factor loadings, model misspecification ($\eta^2 = 1.00$ in both main effects), and their interaction (ρ x λ , η^2=.99). The population CFI value decreases as the level of model misspecification increases but this effect depends on the magnitude of the factor loadings. At the population level, the use of Hu and Bentler's cutoff (CFI > .95) will lead us to conclude that there is an adequate fit when the factor loadings are weak ($\lambda = .30$), regardless of model misspecification. However, the CFI will lead us to conclude that fit is much poorer (CFI values between .72 and .92) when the model is misspecified and the factor loadings are very large (λ = .90).

At the sample level, the largest effects on the CFI index are sample size (N, η^2 = .81), model misspecification (ρ, η^2 = .73), model size (p, η^2 = .71), and the $\lambda \times N$ interaction effect ($\lambda \times N$, η^2 = .75), but the magnitude of factor loadings and the $p \times N$ and $\rho \times \lambda$ interactions also had a large effect (λ, η^2 = .50; $p \times N$, η^2 = .68; and $\rho \times \lambda$, η^2 = .49). The graphs of Figure 2b show that the population CFI (solid line) is a constant of 1.00 independent of p and λ under the correct model. However, when the model is correctly specified, the estimated CFI values are indicative of close fit (CFI > .95) only for models with medium and very large factor loadings (i.e., λ = .60 and .90) and with samples of 200 or more observations. The sample CFI suggests an increasing level of misfit as the number of variables increases, and this effect is especially pronounced for $N = 100$. Under the three conditions of model misspecification, the estimated CFI values are indicative of good fit (CFI > .95) only in samples with 500 or more observations and this effect depends on model size. As p increased from 36 to 72 variables, the sample CFI decreased from .60 to .30 when $\lambda = .30$, from .90 to .70 when $\lambda = .60$, and from .90 to .85 when $\lambda = .90$. That is, CFI decreased as the number of indicators per factor increased. Figure 2b also makes it clear that as p increases, the difference between the population CFI and the sample CFI average values becomes larger. For example, under correctly specified models with $N = 100$ and $\lambda = .30$ the difference between the population CFI and the average sample CFI increased from .15 (when $p = 12$) to .70 (when $p = 72$). We conclude that because of the tendency of underestimation of population CFI values across all conditions, models with no specification error or with minor specification errors can be rejected if evaluated solely based on their sample CFI.

TLI results

The results for the TLI index are very similar to the ones already commented on for the CFI. At the population level, the largest effects found in the ANOVA for TLI are due to the magnitude of factor loadings and model misspecification ($\eta^2 = .99$ in both cases), and to their interaction (ρ x λ, $\eta^2 = .98$). The use of the TLI > .95 cutoff will correctly lead researchers to conclude that their model yields an adequate fit when $\lambda = .30$ and .60. However, when $\lambda = .90$ the use of this cutoff will lead to incorrectly rejecting well-fitting models.

At the sample level, the largest effects are due to sample size (N, $\eta^2 = .66$), the λ x N interaction (λ x N, $\eta^2 = .60$), model misspecification (ρ, $\eta^2 = .59$), model size (p, $\eta^2 = .45$), and the p x N interaction (p x N, $\eta^2 = .40$). However, the magnitude of factor loadings and its interaction with model misspecification also exerted an effect (λ, $\eta^2 = .28$; and ρ x λ, $\eta^2 = .26$). As seen in Figure 2c, the population values of the TLI tended to be underestimated when sample size decreased and model size increased. For example, a correctly specified model with 72 variables and $N = 100$ would be rejected for all λ values if the .95 cutoff value for TLI were applied: the average sample TLI values, in this case, range between .30 and .87. Under the SIM condition, consisting of collapsing two highly correlated factors ($\rho = .90$) into one factor, when sample size is 200 or less and the number of variables in the model is 36 or more, the population TLI would be substantially underestimated and therefore, the model would be rejected. In this case, these effects were more pronounced as the magnitude of the factor loadings was smaller.

GFI results

At the population level, the largest effects found in the ANOVA for GFI are due to the magnitude of factor loadings (λ, $\eta^2 = .98$), model misspecification (ρ, $\eta^2 = .93$), and their interaction (ρ x λ, $\eta^2 = .94$). The population GFI suggests perfect fit (GFI = 1.00) when $\lambda = .30$ regardless of model misspecification. However, it suggests a very poor fit (GFI values between

.20 and .52) when the model is misspecified and λ = .90. In the SIM condition (substantially ignorable misspecification) the GFI suggests that fit was adequate (GFI > .95) only for the small and medium magnitudes of factor loadings (λ = .30 and .60), whereas for large factor loadings (λ = .90) the population GFI failed to reach the cutoff for a good fit.

At the sample level, the largest effects for the GFI index were also due to the magnitude of factor loadings (λ, η^2 = .97) and model misspecification (ρ, η^2 = .91) main effects, and their interaction ($\rho \times \lambda$, η^2= .92), but model size and sample size also had a large effect (p, η^2= .91; N, η^2 = .77). The graphs of Figure 2d show that the values for GFI were underestimated as sample size decreased and model size increased. For instance, a correct model with 36 or more variables would be rejected using the conventional cutoff with samples of 500 or fewer observations.

Under model misspecification conditions, the discrepancy between the population and the estimated GFI increases as the number of observed variables in the model increases and sample size decreases, and the estimated GFI is particularly poor for models with large factor loadings (λ = .90). In summary, the effect of the magnitude of factor loadings on the sample GFI depends on the degree of model misspecification, model size, and sample size, and the sample GFI suggests the model fits adequately when the factor loadings have medium and small magnitudes (λ = .30 and .60), the number of variables is 36 or less, and sample size is 500 or larger.

Revisiting cutoffs of the SRMR fit index

As seen in the previous section, our results indicate that the effect of the magnitude of the factor loadings is important for the estimation of all the SEM fit indices considered in the study, and that model size and sample size are main drivers of the behavior of the majority of these fit indices. As expected, when an unbiased point estimator is used (SRMR_u and RMSEA) it is possible to accurately approximate the population parameters across all study conditions. However, and consistent with previous research, the RMSEA parameter itself is affected by model size and has the problem that it is difficult to interpret because it is defined in an unstandardized metric.

Based on our findings, our recommendation favor the use of the unbiased SRMR index, which is not affected by sample size or by model size, and it is the only fit index formulated in a standardized metric and having an associated statistical test of close fit. Given the influence of the magnitude of the factor loadings on the SRMR$_u$ index, we also suggest, following Shi et al. (2018), to use a cutoff value for this index that is a function of the average communality of the observed variables, $\bar{R}^2$. More specifically, they proposed using $\text{SRMR} / \bar{R}^2 \leq .05$ to identify 'close fitting' models, and $\text{SRMR} / \bar{R}^2 \leq .10$ to identify 'adequate fitting' models. However, when putting forth their proposal, Shi et al. examined solely the behavior of these cutoff values at the population level. It is therefore of interest to examine the validity of these cutoff values in finite samples.

The left-hand side of Table 4 shows the cutoffs given by Shi et al (2018) for the $\bar{R}^2$ values used in our study, and the right-hand side of the table offers the sample estimates for the unbiased SRMR index under the study conditions. As seen in Table 4, in the models with a small average estimated communality ($\bar{R}^2 = .09$), corresponding to the condition of weak factor loadings ($\lambda = .30$), the cutoffs proposed for the SRMR (.005 and .009) are hard to achieve, even under conditions of correctly specified models, where the SRMR$_u$ study values range from .022 to .102, and for the models with SIM (i.e., models collapsing two correlated factors with $\rho = .90$ into one factor), where the SRMR$_u$ values range from .044 to .139. These results suggest that the Shi et al. correction does not work well when the factor loadings are small and that a sample size of 500 or more observations is needed to warrant a close fit, particularly when the model includes a smaller number of indicators per factor. However, these results are congruent with the literature on the recovery of weak factor loadings (Briggs & Maccallum, 2003; Ximénez, 2006; 2009) that indicates that ML often fails to recover weak factors, particularly with small sample sizes (e.g., $N = 100$), as is the case here. Concerning the cutoffs proposed for the average estimated communalities of .36 and .81, corresponding to the conditions of $\lambda = .60$ and .90, they are quite accurate both for the correct models (SRMR$_u$ values from .004 to .026 for $\lambda = .60$, and from .0003 to .004 for $\lambda = .90$), and for the models with SIM (SRMR$_u$ values from .039 to .050 for $\lambda = .60$, and from .045 to .049 for $\lambda = .90$). Finally, as expected, in the conditions of medium and large misspecification (i.e., models collapsing two factors with correlations of $\rho = .80$ and .70 into one factor), the estimated cutoffs are inappropriate (i.e., $\text{SRMR} / \overline{R}^2 > .10$) across almost all conditions, indicating the lack of fit of the model.

----- INSERT TABLE 4 ABOUT HERE -----

Figure 3 offers a visual presentation of the correction of Shi et al. (2018), illustrating the behavior of the sample SRMR$_u$ under the simulation conditions, and the proposed cutoffs for $\text{SRMR} / \overline{R}^2$. This figure is similar to Figure 1a, but in this case the doted lines represent the sample SRMR$_u$ values, and the red solid lines, instead of representing Hu and Bentler's (1999) cutoff value (i.e., $\text{SRMR} < .08$), they mark Shi et al.'s (2018) cutoff values for close fit ($\text{SRMR} / \overline{R}^2 \leq .05$, thin line) and adequate fit ($\text{SRMR} / \overline{R}^2 \leq .10$, thick line). As can be seen, both in the correct models and in the models with SIM, the dotted lines fall below the cutoffs, indicating a close fit; whereas, they are above such cutoffs in the models with medium and large model misspecification, indicating unacceptable goodness of fit.

----- INSERT FIGURE 3 ABOUT HERE -----

Overall, these results indicate that the correction proposed by Shi et al. (2018) to determine the close fit of the SRMR index as a function of the communality works reasonably well at the sample level. Our results show that, after this correction, the interpretation of the goodness of fit as a function of the reliability is straightforward. That is, the lower the reliability, the lower the goodness of fit, and vice versa. Moreover, Shi et al.'s cutoffs detects misspecified models whereas the SRMR$_u$ index without the correction cannot detect and reject a misspecified model that includes weak or moderate factor loadings.

Discussion and Conclusion

This article addressed the problem of how the magnitude of factor loadings affects the goodness of fit of the model. Previous research has addressed the problem but has not specifically focused on how the magnitude of the factor loadings may affect the validity of the cutoffs typically recommended for the SEM indices to assess the goodness of fit of the model. We presented the results of a simulation study investigating the problem of how the magnitude of factor loadings affects the goodness of fit of the model under varying conditions of model size, model misspecification, and sample size. The study focuses on the behavior of two groups of indices: A group of unbiased indices: The unbiased SRMR index proposed by Maydeu-Olivares (2017), that is the preferred one because it is the only fit index formulated in a standardized metric that has an associated statistical test of close fit, and the RMSEA, the most widely used index; and a group of biased indices commonly used in practice: The (biased) SRMR, the CFI, the TLI, and the GFI. The purpose of the study was to examine the effect of the magnitude of factor loadings on the estimation of such indices both in correctly and incorrectly specified models, and both at population and at the sample level. We analyzed the correspondence between the sample goodness-of-fit indices and their population counterparts to evaluate the validity of the typically recommended cutoffs for the population SEM indices (Hu & Bentler, 1999). Conditions regarding the characteristics of the model, such as the number of observed variables or the sample size were also manipulated to better understand the behavior of the SEM fit indices.

Our results indicated that all fit indices were affected by the magnitude of the factor loadings. In the group of the unbiased indices ($SRMR_u$ and RMSEA), under model misspecification, the goodness of fit was poorer in the conditions with the largest factor loadings (i.e., $\lambda = .90$), and the largest level of model misspecification. However, congruent with previous research (Savalei, 2012), the RMSEA index suggests model fit is acceptable in misspecified models with a large number of variables and large sample size (e.g., $p = 72$ variables and $N > 500$). The results in the group of the biased indices ($SRMR_b$, CFI, TLI, and GFI) indicated that the magnitude of factor loadings also exerted a large effect. However, the interpretation for such indices was opposite to the one found for the unbiased indices. The CFI, TLI, and GFI indices rejected the misspecified models but fit was much poorer for models with weak factor loadings ($\lambda = .30$), and this effect was more pronounced when using small sample sizes ($N = 200$ or 500 observations) and large model sizes ($p = 36$ or 72 variables). In addition, we found that both the model size and the sample size were main drivers of the behavior of the biased indices. The CFI, TLI, and GFI indices do not yield values that suggest acceptable fit when the model is correct and includes 36 or more variables, unless the sample includes 500 or more observations. Thus, we recommend not using the CFI, TLI, and GFI indices in models containing a large number of variables and weak factor loadings, particularly if the sample size is medium or small.

The phenomenon of poor measurement quality being associated with better model fit has been named the *reliability paradox* (Hancock & Mueller, 2011) and has been studied mathematically at the population level (Heene et al., 2011), and at the sample level (McNeish, An, & Hancock, 2018). The findings of the present study show that the *reliability paradox* may have operated for sample $SRMR_u$ and RMSEA values across all conditions under study. However, for sample values of CFI, TLI, and GFI our findings show that the *reliability paradox* does not operate and that the effect of measurement quality depends on other factors such as the sample size and model size. These results are congruent with the ones found in the study by Shi et al. (2019), implying that sample estimates of CFI, TLI, and GFI, on average, tended to indicate worse fit under the condition of poorer measurement quality when a model of large size was fit to a sample of small to medium size.

Our findings also support the validity of the correction proposed by Shi et al. (2018) to determine close fit using the SRMR index. Overall, the $\text{SRMR} / \overline{R}^2$ correction was very accurate and, more importantly, could detect the model misspecification and thus reject the misspecified models. This result is important as, without the correction, the SRMR index could not distinguish between different levels of model misspecification. Moreover, the correction provides a straightforward interpretation of the SRMR and avoids the *reliability paradox*. However, future research should examine in more detail the appropriateness of the Shi et al. correction in models including small factor loadings. Our results showed that, for such models, the correction only does well in models with a large number of indicators per factor and when using large sample sizes. These results could be due to the particular conditions examined in our study. Thus, future studies should continue to examine these effects under other conditions. For instance, in models with non-identical factor loadings and in models including a mixture of weak, moderate, or high factor loadings. Future research can also be directed to develop corrections for other goodness-of-fit indices. To our knowledge, no similar correction for the RMSEA has ever been proposed but, given the importance of the magnitude of factor loadings and the effect of the *reliability paradox* under certain conditions (see Shi et al., 2019), further research on whether a similar correction could be applied to the RMSEA is needed.

Overall, researchers should be aware that sample values of fit indices depend on the characteristics of the model, such as the magnitude of the factor loadings, the number of observed variables or the sample size. For instance, our study has shown that for models with weak factor loadings (e.g., λ=.30), the SRMR$_u$ and RMSEA indices will suggest that the model fits well even in severely misspecified models. In the case of the RMSEA, this effect will be more pronounced when the model includes a large number of variables (e.g., p = 72) and the sample size is small (e.g. N = 100 or 200). In contrast, the CFI and the TLI indices will incorrectly suggest that fit is poor when the magnitude of factor loadings is small (e.g., λ = .30 or .60) and when the model includes a larger number of variables (p = 36 or more). In such cases, a sample size of N=500 observations or more is needed for these indices to provide an accurate assessment of the degree of misfit of the model. Something similar happens with the GFI but in this case the magnitude of the factor loadings only affects the goodness of fit when the model is misspecified, and when it includes a large number of variables. Given this, one needs to be cautious about deciding the goodness-of-fit index to be used to report the results of a study.

Based on our findings, we conclude that researchers should favor the use of the unbiased indices and, if using biased indices, they should be aware that a much larger sample size is needed. In addition, we recommend considering not only a single cutoff for assessing the degree of fit of a model but different cutoffs based on the average value of the factor loadings. Our recommendation is to favor the use of the unbiased SRMR index of Maydeu-Olivares (2017) with the correction proposed by Shi et al. (2018) as a function of the communality ($\text{SRMR} / \overline{R}^2$) for the correct interpretation of the SRMR.[3] Some years ago, Bentler (2007) also offered this recommendation of

[3] Presently the unbiased SRMR index and its confidence intervals and tests of close fit are available on the lavaan package version 0.6-7 in R (see function *lavResiduals*).

reporting the SRMR, arguing that SRMR needed little further research, as it is interpretable on its own. The present research has contributed to that aim by comparing the unbiased SRMR with other commonly used SEM indices.

As is the case with any Simulation study, our results will hold only in conditions similar to those considered herein. In the present study our aim was to investigate the drivers of the behavior of sample goodness-of-fit indices and relate them to those of population goodness-of-fit indices. Congruent with previous research (Maydeu-Olivares et al., 2017; Shi et al., 2018; 2019), we found that the main drivers of the behavior of sample goodness-of-fit indices were sample size, factor loading, and model size. On the basis of our findings, we suggest that model size can be controlled by using a standardized effect size; factor loading by using a correction as proposed by Shi et al. (2008); and sample size by using an unbiased estimator and a confidence interval. However, future research should continue to examine these effects under different study conditions. For instance, further study could be directed to examining more complex models (e.g., hierarchical models or structural equation models) and other forms of model misspecification (e.g., omitting cross-loadings or residual correlations).

In closing, we hope that this research provides additional information to SEM researchers to assist them when conducting the difficult task of assessing the goodness of fit of their models.

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

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. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*, 825–829. http://dx.doi.org/10.1016/j.paid.2006.09.024

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 518–540. https://doi.org/10.1080/10705511.2013.797839

Briggs, N. E., & Maccallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, *38*(1), 25–56. https://doi.org/10.1207/S15327906MBR3801_2

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

Browne, M. W., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403–421. https://doi.org/10.1037//1082-989X.7.4.403

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 343–367. https://doi.org/10.1207/s15328007sem1203_1

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

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: a cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

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

Jackson (2003) Revisiting sample size and number of parameter estimates: some support for the n:q hypothesis. *Structural Equation Modeling*, *10*, 128-141, https://doi.org/10.1207/S15328007SEM1001_6

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

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. https://doi.org/10.1207/S15328007SEM1003_1

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

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

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*(3), 391–410. https://doi.org/10.1037/0033-2909.103.3.391

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

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

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107*(2), 247–255. https://doi.org/10.1037/0033-2909.107.2.247

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

Miles, J. & Shevlin, M. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences*, *25*, 85-90. http://dx.doi.org/10.1016/S0191-8869(98)00055-5

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

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

R Development Core Team. (2019). *R: A lenguage and enviornment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org/index.html

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. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement*, *72*(6), 910–932. https://doi.org/10.1177/0013164412452564

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*, 935-43. http://dx.doi.org/10.1016/j.jbusres.2003.10.007

Shi, D., DiStefano, C., Maydeu-Olivares, A., & Lee, T. (2021). Evaluating SEM model fit with small degrees of freedom. *Multivariate Behavioral Research*, *0*(0), 1–36. https://doi.org/10.1080/00273171.2020.1868965

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). Abstract: 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., 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

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*(2), 137–167. https://doi.org/10.1207/S15327906MBR3502_1

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. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduck and Glaser. *Structural Equation Modeling*, *7*(2), 149–162. https://doi.org/10.1207/S15328007SEM0702_1

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*, 893-98. http://dx.doi.org/10.1016/j.paid.2006.09.017

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(4), 587–614. https://doi.org/10.1207/s15328007sem1304_5

Ximénez, C. (2009). Recovery of weak factor loadings in confirmatory factor analysis under conditions of model misspecification. *Behavior Research Methods*, *41*(4), 1038–1052. https://doi.org/10.3758/BRM.41.4.1038

Table 1

SEM fit indices used in the study

Population indices	Sample indices	Description
$SRMR = \sqrt{\frac{\boldsymbol{\varepsilon}_s{}'\boldsymbol{\varepsilon}_s}{t}} = \sqrt{\frac{1}{t}\sum_{i\le j}\frac{\left(\sigma_{ij}-\sigma_{ij}^{0}\right)^2}{\sqrt{\sigma_{ii}\sigma_{jj}}}}$ where $\boldsymbol{\varepsilon}_s$ is the vector of t population standardized residual covariances $t = p(p + 1)2$ σ_{ij} unknown population covariance σ_{ij}^{0} population covariance under the fitted model	$\widehat{\text{SRMR}}_b = \sqrt{\frac{\mathbf{e}_s{}'\mathbf{e}_s}{t}} = \sqrt{\frac{1}{t}\sum_{i\le j}\left(\frac{s_{ij}-\hat{\sigma}_{ij}}{\sqrt{s_{ii}s_{jj}}}\right)^2}$	*Standardized Root Mean Residual*, Cutoff: $\text{SRMR}_b \le .08$ (Hu & Bentler, 1999)
	$\widehat{\text{SRMR}}_u = \hat{k}_s^{-1}\sqrt{\frac{\max\left(\mathbf{e}_s{}'\mathbf{e}_s - \text{tr}\left(\hat{\boldsymbol{\Xi}}_s\right),0\right)}{t}}$ where $\hat{\boldsymbol{\Xi}}_s$ is the asymptotic covariance matrix of $\mathbf{e}_s$, the sample standardized residual covariances $\hat{k}_s^{-1} = 1 - \frac{\text{tr}\left(\hat{\boldsymbol{\Xi}}_s^2\right) + 2\mathbf{e}_s\hat{\boldsymbol{\Xi}}_s\mathbf{e}_s}{4(\mathbf{e}_s{}'\mathbf{e}_s)^2}$	*Unbiased Standardized Root Mean Residual*, Cutoff: $\text{SRMR}_u / \bar{R}^2 \le .05$ (Shi, Maydeu-Olivares & DiStefano, 2018) Test of close fit: $\text{H}_0 : \text{SRMR}_u \le c$*
$\text{RMSEA} = \sqrt{\frac{F_k}{df_k}}$	$\widehat{\text{RMSEA}} = \sqrt{\frac{\max\left(\chi_k^2 - df,0\right)}{df_k(N\text{-}1)}}$	*Root Mean Squared Residual*, Cutoff: $\text{RMSEA} \le .05$ (Steiger, 2007) Test of close fit: $\text{H}_0 : \text{RMSEA} \le c$
$\text{CFI} = 1 - \frac{F_k}{F_0}$ where F_0 and F_k are the minimum of the discrepancy function for the baseline and the fitted model	$\widehat{\text{CFI}} = \frac{\max\left(\chi_0^2 - df_0,0\right) - \max\left(\chi_k^2 - df_k,0\right)}{\max\left(\chi_0^2 - df_0,0\right)}$ where χ_0^2 and χ_k^2, and df_0 and df_k are the chi-square statistics and degrees of freedom for the baseline and the fitted model	*Comparative Fit Index*, Cutoff: $\text{CFI} \ge .95$ (Hu & Bentler, 1999)
$\text{TLI} = 1 - \frac{F_k/df_k}{F_0/df_0}$	$\widehat{\text{TLI}} = \frac{\chi_0^2/df_0 - \chi_k^2/df_k}{\chi_0^2/df_0 - 1}$	*Tucker-Lewis Index*, Cutoff: $\text{TLI} \ge .95$ (Sharma, Mukherjee, Kumar & Dillon, 2005)
$GFI = \frac{p}{p + 2F_k}$	$\widehat{\text{GFI}} = 1 - \frac{\text{tr}\left(\hat{\Sigma}^{-1}\mathbf{S} - \mathbf{I}\right)^2}{\text{tr}\left(\hat{\Sigma}^{-1}\mathbf{S}\right)^2}$ where $\mathbf{S}$ and $\hat{\Sigma}$ denote the sample covariances and estimated covariances under the fitted model	*Goodness-of-Fit-Index*, Cutoff: $\text{GFI} \ge .95$ (Miles & Shevlin, 1998)

Table 2

Variables considered in the Monte Carlo study

Code	Variable	Levels
p	Number of variables	12 (small model size) 36 (medium model size) 72 (large model size)
ρ	Degree of model misspecification	$\rho = 1.00$ (no misspecification) $\rho = .90$ (small or substantially ignorable misspecification, SIM) $\rho = .80$ (medium misspecification) $\rho = .70$ (large misspecification)
λ	Magnitude of factor loadings	.30 (small or weak) .60 (medium) .90 (large)
N	Sample size	100 200 500 1000

Table 3

Correlations between population and sample fit indices (along the diagonal) and between sample fit indices (off-diagonal) for various levels of factor loading and model misspecification

	$\lambda = .30$						$\lambda = .60$						$\lambda = .90$					
ρ = .70	1.	2.	3.	4.	5.	6.	1.	2.	3.	4.	5.	6.	1.	2.	3.	4.	5.	6.
1. SRMR$_u$	(.64)**						(.93)**						(.66)*					
2. SRMR$_b$	.64*	(.18)					.01	(.30)					.99**	(.56)*				
3. RMSEA	.64*	.85**	(-.37)				-.68*	.38	(.62)*				-.63*	-.53	(.99)**			
4. CFI	-.66*	-.82**	-.98**	(.51)			-.45	-.72**	-.23	(.62)*			-.76**	-.73**	.73**	(.77)**		
5. TLI	-.73**	-.86**	-.97**	.97**	(.46)		-.38	-.74**	-.30	1.00**	(.57)*		-.32	-.39	-.15	.56	(.17)	
6. GFI	-.70**	-.80**	-.94**	.98**	.95**	(.63)**	-.73**	-.63*	.27	.85**	.81**	(.92)**	-.64*	-.54	1.00**	.77**	-.11	(1.00)**
ρ = .80																		
1. SRMR$_u$	(-.10)						(.89)**						(1.00)**					
2. SRMR$_b$	.84**	(.17)					-.09	(.23)					.92**	(.81)**				
3. RMSEA	.63*	.87**	(-.37)				-.26	.68*	(.20)				-.95**	-.76**	(.99)**			
4. CFI	-.55**	-.83**	-.99**	(.48)			-.41	-.70*	-.67*	(.59)*			-.81**	-.85**	.70*	(.77)**		
5. TLI	-.84**	-.86**	-.81**	.76**	(.18)		-.38	-.71*	-.70*	1.00**	(.56)		-.49	-.66**	.32	.90**	(.29)	
6. GFI	-.48	-.81**	-.94**	.97**	.71**	(.62)**	-.58*	-.70*	-.36	.90**	.88**	(.82)**	-.97**	-.81**	.99**	.77**	.42	(1.00)**
ρ = .90																		
1. SRMR$_u$	(-.35)						(.94)**						(1.00)**					
2. SRMR$_b$	.83**	(.17)					.02	(.18)					.63*	(.63)*				
3. RMSEA	.60*	.89**	(-.35)				.22	.83**	(-.25)				-.96**	-.46	(.97)**			
4. CFI	-.49	-.83**	-.99**	(.45)			-.45	-.69*	-.94**	(.52)			-.71*	-.82**	.53	(.73)**		
5. TLI	-.27	-.67*	-.92**	.96**	(.56)		-.45	-.69*	-.94**	1.00*	(.52)		-.59*	-.80**	.40	.99**	(.64)**	
6. GFI	-.38	-.81**	-.94**	.97**	.96**	(.61)*	-.57	-.77**	-.85**	.91**	.91**	(.68)*	-.98**	-.68*	.94**	.78**	.68**	(1.00)**
ρ = 1.00																		
1. SRMR$_u$	-						-						-					
2. SRMR$_b$	.81**	-					.79**	-					.80**	-				
3. RMSEA	.58*	.89**	-				.57	.90**	-				.57	.90**	-			
4. CFI	-.45	-.82**	-.99**	-			-.31	-.69*	-.93**	-			-.29	-.66*	-.91**	-		
5. TLI	-.30	-.74**	-.95**	.98**	-		-.30	-.69*	-.93**	1.00**	-		-.28	-.65*	-.91**	1.00**	-	
6. GFI	-.34	-.81**	-.94**	.96**	.97**	-	-.31	-.80**	-.94**	.90**	.91**	-	-.31	-.80**	-.94**	.88**	.88**	-

Note: Each correlation is based on 1,000 replications * $p < .05$ ** $p <.01$.

Table 4

Comparison of the cutoff SRMR corrected values for the estimated SRMR$_u$ index

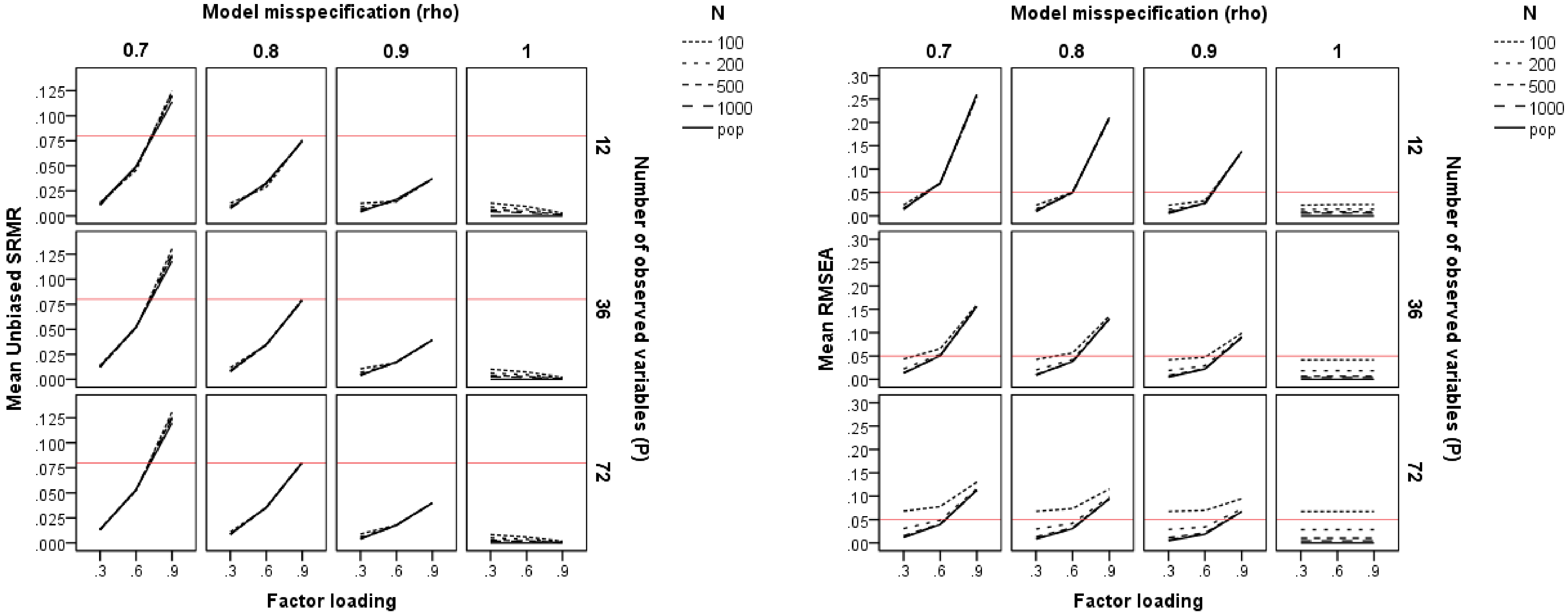
		SRMR reference values		SRMR$_u$ study values												
$\overline{R}^2$	ρ	Close fit	Adequate fit	Population SRMR	p = 12				p = 36				p = 72			
					N=100	N=200	N=500	N=1000	N=100	N=200	N=500	N=1000	N=100	N=200	N=500	N=1000
.09	1.00	.005	.009	.000	.102	.097	.066	.047	.094	.071	.044	.030	.083	.057	.033	.022
	.90	.005	.009	.048	.139	.099	.072	.054	.118	.076	.054	.044	.101	.067	.048	.044
	.80	.005	.009	.096	.142	.110	.092	.080	.131	.097	.086	.090	.123	.097	.095	.097
	.70	.005	.009	.144	.150	.128	.121	.122	.157	.133	.136	.142	.156	.143	.147	.147
.36	1.00	.018	.036	.000	.026	.018	.012	.009	.021	.013	.008	.005	.016	.010	.006	.004
	.90	.018	.036	.048	.041	.039	.042	.044	.047	.046	.048	.048	.050	.049	.049	.049
	.80	.018	.036	.096	.079	.086	.090	.091	.095	.095	.096	.097	.097	.098	.098	.099
	.70	.018	.036	.144	.127	.134	.136	.136	.143	.143	.145	.145	.145	.147	.148	.148
.81	1.00	.041	.081	.000	.004	.002	.002	.001	.003	.002	.001	.001	.002	.001	.001	.000
	.90	.041	.081	.048	.045	.046	.046	.046	.049	.048	.048	.049	.049	.049	.049	.049
	.80	.041	.081	.096	.094	.093	.092	.092	.099	.098	.097	.097	.100	.100	.099	.099
	.70	.041	.081	.145	.154	.152	.149	.147	.161	.156	.153	.151	.162	.159	.155	.153

Note. The cells for the Study values in SRMR$_u$ show the estimated value for each index under each misspecification condition (ρ) and communality value ($\overline{R}^2$) for each simulated model (with model sizes, p, from 12 to 72 variables and sample sizes, N, from 100 to 1000 observations).

Figure 1

Behavior of the sample unbiased indices (SRMR$_u$ and RMSEA) under the Simulation study conditions

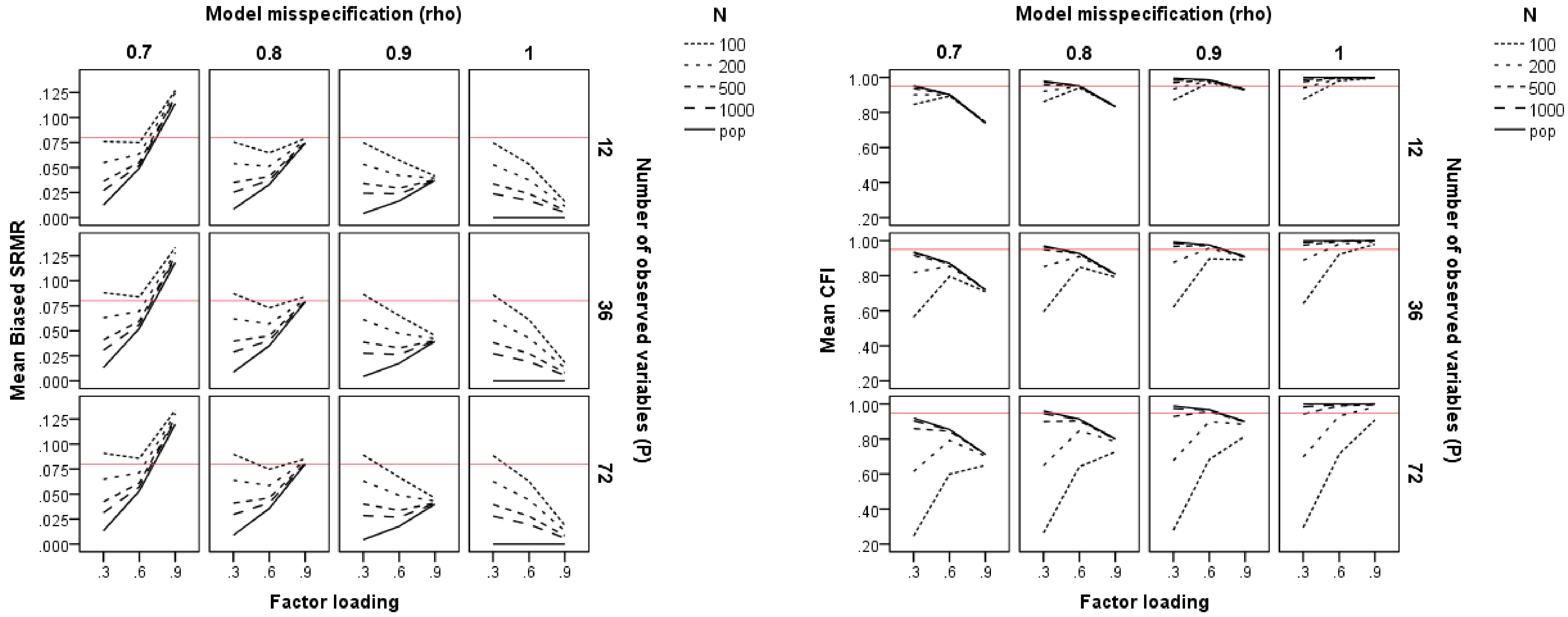
(a) Behavior of the unbiased SRMR index (*SRMR$_u$*) (b) Behavior of the RMSEA index



Note: rho is model misspecification (ρ = 1.00, .90, .80, and .70), N is the sample size (100, 200, 500, and 1000), P is the number of observed variables (12, 36, and 72), Factor loading is the magnitude of the factor loadings (.30, .60, and .90), pop is the population values for each index, and the red solid line corresponds to Hu and Bentler's (1999) cutoff for each index.

Figure 2

Behavior of the sample biased indices (SRMR$_b$, GFI, CFI, and TLI) under the Simulation study conditions

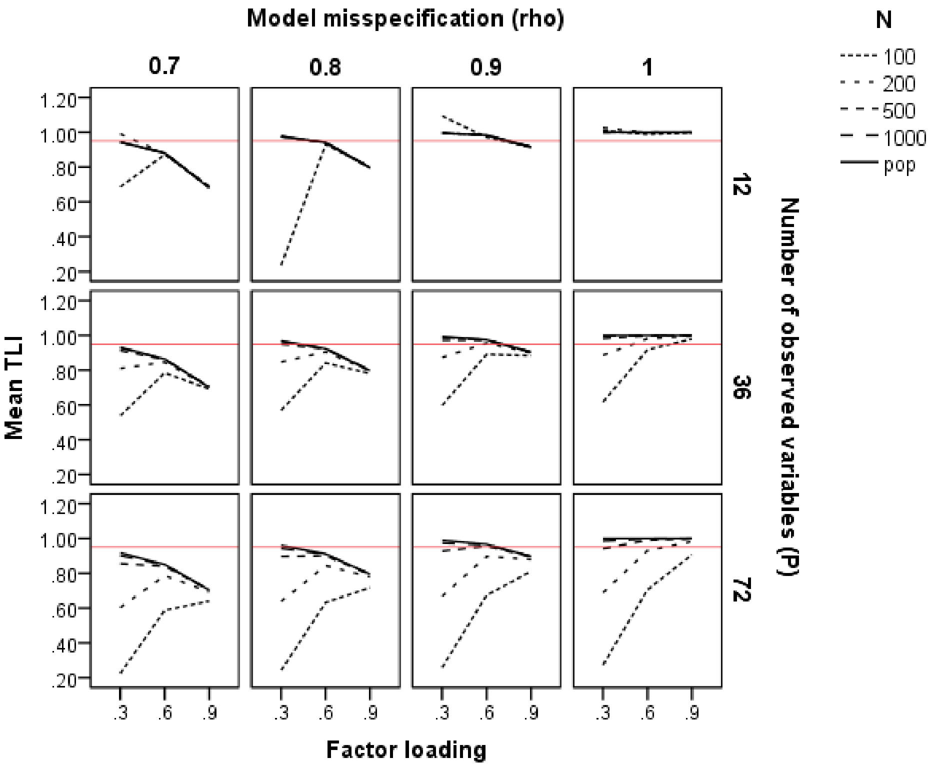


(a) Behavior of the biased SRMR index (*SRMR$_b$*)

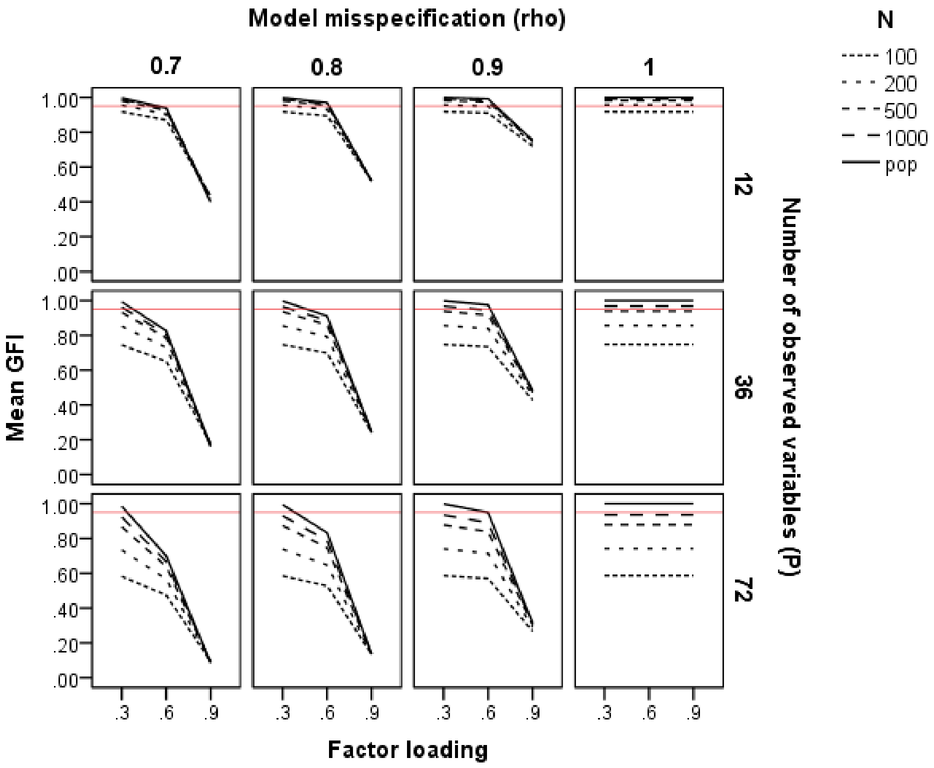
(b) Behavior of the CFI index

Figure 2 (continued)

(c) Behavior of the TLI index



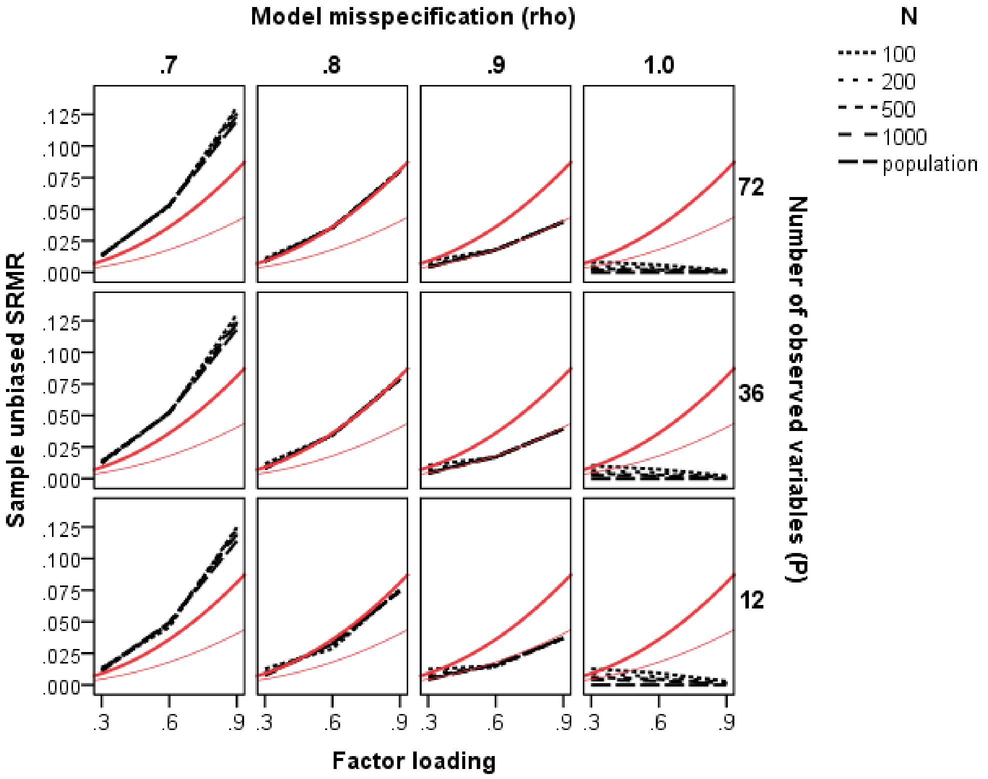
(d) Behavior of the GFI index



Note: rho is model misspecification (ρ = 1.00, .90, .80, and .70), N is the sample size (100, 200, 500, and 1000), P is the number of observed variables (12, 36, and 72), Factor loading is the magnitude of the factor loadings (.30, .60, and .90), pop is the population values for each index, and the red solid line corresponds to Hu and Bentler’s (1999) cutoff for each index.

Figure 3

Behavior of the SRMR$_u$ *cutoffs of Shi et al.* (*2018*)



Note: Dotted lines indicate behavior of sample SRMR$_u$ and population SRMR under the study conditions; and red solid lines indicate two cutoffs for $\text{SRMR}_u / \overline{R}^2 = .05$ (thin line, close fit) and .10 (thick line, adequate fit).