Psychological Methods 21.1 (2016): 93–111

DOI: http://dx.doi.org/10.1037/met0000064

Are Fit Indices Really Fit to Estimate the Number of Factors with Categorical Variables? Some Cautionary Findings Via Monte Carlo Simulation

Luis Eduardo Garrido[1] Francisco José Abad[2] Vicente Ponsoda[2]

[1]Universidad Iberoamericana en República Dominicana

[2]Universidad Autónoma de Madrid

Author Note

Luis E. Garrido, Decanato de Investigación Académica, Universidad Iberoamericana en República Dominicana.

Francisco J. Abad, Facultad de Psicología, Universidad Autónoma de Madrid.

Vicente Ponsoda, Facultad de Psicología, Universidad Autónoma de Madrid.

Francisco Abad was supported by Grant PSI2013-44300-P (Ministerio de Economía y Competitividad, Spain).

Vicente Ponsoda was supported by Grant PSI2012-33343 (Ministerio de Economía y Competitividad, Spain).

Correspondence concerning this article should be addressed to Luis Eduardo Garrido, Decanato de Investigación Académica, Universidad Iberoamericana, Ave. Francia No. 129, Gazcue, Santo Domingo, Dominican Republic. Contact: l.garrido@prof.unibe.edu.do

Abstract

An early step in the process of construct validation consists in establishing the fit of an unrestricted "exploratory" factorial model for a pre-specified number of common factors. For this initial unrestricted model, researchers have often recommended and used fit indices to estimate the number of factors to retain. Despite the logical appeal of this approach, little is known about the actual accuracy of fit indices in the estimation of data dimensionality. The present study aimed to reduce this gap by systematically evaluating the performance of four commonly used fit indices –CFI, TLI, RMSEA, and SRMR– in the estimation of the number of factors with categorical variables, and comparing it with what is arguably the current golden rule, Horn's parallel analysis. The results indicate that CFI and TLI provide nearly identical estimations and are the most accurate fit indices, followed at a step below by RMSEA, and then by SRMR, which gives notably poor dimensionality estimates. Difficulties in establishing optimal cutoff values for the fit indices and the general superiority of parallel analysis, however, suggest that applied researchers are better served by complementing their theoretical considerations regarding dimensionality with the estimates provided by the latter method.

Keywords: fit indices, number of factors, categorical variables, exploratory factor analysis, exploratory structural equation modeling, parallel analysis

¿Are Fit Indices Really Fit to Estimate the Number of Factors to Retain? Some Cautionary Findings Via Monte Carlo Simulation with Categorical Variables

Methodologists and applied researchers have recommended and used fit indices with increased frequency in recent years to estimate the number of factors to retain within the context of unrestricted factor analysis (e.g., Asparouhov & Muthén, 2009; Campbell-Sills, Liverant, & Brown, 2004; Ferrando & Lorenzo-Seva, 2000; Sanne, Torsheim, Heiervang, & Stormark, 2009; Tepper & Hoyle, 1996). This approach is advantageous because while assessing the fit of factor models researchers have access to important model diagnostic information, such as the presence of correlated residuals among factor indicators, which can be taken into consideration when making the dimensionality decision. In contrast, the classic retention methods that have been widely used or recommended in the factor analysis literature, such as the eigenvalue-greater-than-one rule (Kaiser, 1960), the minimum average partial method (Velicer, 1976), and Horn's parallel analysis (Horn, 1965), are based on principal component analysis, where such diagnostic information is not available. Furthermore, using fit indices to estimate the number of factors reduces the need for ad-hoc model manipulation in the more advanced stages of testing, such as the evaluation of a restricted "confirmatory" model or a full-blown structural equations "SEM" model, due to a poorly conceived unrestricted factor structure (Mulaik & Millsap, 2000; Patil, Singh, Mishra, & Donovan, 2008).

Despite the logical appeal of using fit indices to estimate the number of underlying factors, little is known about their actual *accuracy* in this area of research (Frazier & Youngstrom, 2007; Yang & Xia, 2014). This situation is disconcerting, as the critical dimensionality decision is oftentimes being made without any prior information regarding the level of performance that can be expected from the different fit indices. Moreover, there is also limited knowledge regarding the behavior of fit indices with categorical variables (Barendse, Oort, & Timmerman, 2015;

Beauducel & Herzberg, 2006), which are typically encountered in the social and behavioral sciences (Flora & Curran, 2004). This is also troublesome, as the measures of association, estimation methods, and fit functions that are recommended for the factor analysis of categorical variables are different than those for continuous variables (Savalei & Rhemtulla, 2013), and may impact their performance differentially (Nye & Drasgow, 2011).

As a result of the aforementioned issues in the literature, our motivating goal was to investigate the accuracy of fit indices in the estimation of the number of factors with ordered-categorical variables. In this regard, we aimed to systematically assess the performance of four commonly used fit indices –CFI, TLI, RMSEA, and SRMR– under a wide range of factorial models and sample conditions. There are, however, important issues regarding the use and interpretation of fit indices that must be taken into consideration first. To this end, the rest of this section will be organized according to the following areas of relevance: (1) EFA/ESEM vs. CFA to estimate the number of factors; (2) Categorical variable estimators; (3) Evaluation of model fit with fit indices; (4) Performance of fit indices with CFA and SEM models; and (5) Accuracy of fit indices in the estimation of the number of factors.

EFA/ESEM vs. CFA to Estimate the Number of Factors

The literature regarding when and how to use EFA-CFA appears to have strong roots in some historical limitations of the EFA procedure. For example, Floyd and Widaman (1995) remarked that CFA departed markedly from EFA in that it relied "on a different set of standards for evaluating the adequacy of factor solutions" (p. 293). Furthermore, Myers (2013) observed that typical implementations of the EFA procedure in software have been limited by the "absence of standard errors for parameter estimates, restrictions on the ability to incorporate a priori content knowledge into the measurement model, an inability to fully test factorial invariance, and an inability to simultaneously estimate the measurement model within a fuller structural model"

(p. 712). Because of these historical limitations, CFA has been preferred over EFA in some cases where there wasn't sufficient *a priori* measurement theory to warrant a confirmatory approach (Myers, 2013; Patil et al., 2008).

Recent advances in factor analysis have, however, eliminated the above-mentioned shortcomings of the EFA procedure. In this line, the development of exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009; Marsh et al., 2009) has provided researchers with a flexible factor modeling technique that offers the same fit information available in CFA and can be incorporated into broader model testing, such as full SEM models, multiple group EFA with measurement and structural invariance testing, longitudinal EFA with across-time invariance testing, EFA with covariates and direct effects, and EFA with correlated residuals (Asparouhov & Muthén, 2009). As a result, the choice between EFA/ESEM and CFA is, presently, one that need only be made on the basis of the hypotheses that are to be tested.

In order to better understand the similarities and differences between EFA/ESEM and CFA, it may be useful to frame the discussion in terms of the types of models that can be fitted by each technique. In EFA/ESEM, the observed variables are fitted to an *unrestricted* factor model, where the indicators are allowed to load freely on all the factors that are to be extracted. In addition, an unrestricted solution does not restrict the factor space, allowing for multiple factor solutions to be obtained by an arbitrary rotation or transformation of the estimated factor solution, with each solution yielding the same fit (Ferrando & Lorenzo-Seva, 2000). Because no restrictions are imposed on the factor structure, EFA/ESEM essentially tests whether a specified number of common factors are able to account for the covariation among the observed variables (Tepper & Hoyle, 1996).

In CFA, on the other hand, a *restricted* factor model is fitted to the data, where specific relationships are posited between factors and indicators, between different factors and between different indicators. Therefore, assuming that the distributional assumptions are met, CFA constitutes a test of dimensionality *and* the plausibility of the restrictions imposed through the specified model. It then follows that a CFA may not fit the data because the number of hypothesized factors is inappropriate, the relations among variables and factors are not correctly specified or both (Ferrando & Lorenzo-Seva, 2000). And because these model hypotheses are tested simultaneously, the researcher cannot determine which (if not both) might be the cause of a bad-fitting model, thus making CFA an unsuitable framework to estimate the number of factors to retain. Based on this logic, it is concluded that unrestricted factor analysis in the form of EFA/ESEM is the most appropriate modeling technique to estimate the underlying dimensionality of a set of observed variables.

Categorical Variable Estimators

Normal theory estimators, such as maximum likelihood (ML) and generalized least squares (GLS), are generally used for model estimation with continuous variables because of their desirable asymptotic properties (Lei, 2009). However, these estimators assume that the observed variables follow a multivariate normal distribution, an assumption that is violated when the observed variables are of categorical nature. Moreover, if categorical variables are treated as if they are continuous by employing ML or GLS, distorted parameters estimations, standard errors, and χ^2 statistics can be obtained (Beauducel & Herzberg, 2006; Morata-Ramírez & Holgado-Tello, 2013).

Two strategies that take into account the categorical nature of the observed variables have been proposed to estimate the factor analysis model (Jöreskog & Moustaki, 2001): the *underlying response variable* approach (URV) and the *response function* or *item response theory* approach (IRT). Because the URV approach is the one generally used in factor analysis, it will constitute the focus of this study. Nevertheless, for those interested in the details regarding its relationship to Samejima's (1969) graded response IRT model, see Forero, Maydeu-Olivares and Gallardo-Pujol (2009) and Takane and de Leeuw (1987).

Within the URV approach, the observed categorical variables are considered to be manifestations of underlying normally distributed continuous variables that are partially observed through their categorical counterparts (Olsson, 1979). An observed categorical variable x_i with m_i ordered response categories is linked to its respective underlying continuous response variable x_i^* via a threshold relationship:

$$x_i = c_i \Leftrightarrow \tau_{c_i-1}^{(x_i)} < x_i^* < \tau_{c_i}^{(x_i)} \tag{5}$$

where $\tau_{c_i}^{(x_i)}$ is the c_ith threshold of variable x_i and $-\infty = \tau_0^{(x_i)} < \tau_1^{(x_i)} < \cdots < \tau_{m_{i-1}}^{(x_i)} < \tau_{m_i}^{(x_i)} = +\infty$. That is, an individual will choose response alternative c_i when his latent response value x_i^* is between thresholds τ_{c_i-1} and τ_{c_i}. In addition, for a set of p observed variables, the factors are connected to the latent response variables $\mathbf{x}^*$ through the standard factor analytic model:

$$\mathbf{x}^* = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{6}$$

where $\boldsymbol{\eta}$ is a k x 1 vector of factors, $\boldsymbol{\Lambda}$ is a p x k matrix of factor loadings, and $\boldsymbol{\varepsilon}$ is an k x 1 vector of measurement errors.

This formulation of the common factor model assumes that the factors $\boldsymbol{\eta}$ and the measurement errors $\boldsymbol{\varepsilon}$ are both normally distributed, that the factors and measurement errors are uncorrelated, that the means of the factors and measurement errors are zero, and that the measurement errors are mutually uncorrelated.

The URV factor model is generally estimated in three stages: First, the thresholds are estimated separately for each variable by ML. Second, polychoric correlations (ρ; Olsson, 1979) are estimated independently for each pair of categorical variables, also using ML. Third, the parameters of the factor model are estimated by using the thresholds and polychoric correlations estimated in the previous two stages and minimizing the least squares function:

$$\mathbf{F} = (\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta}))'\widehat{\mathbf{W}}(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})) \quad (7)$$

where $\widehat{\boldsymbol{\rho}}$ is the sample polychoric correlation matrix, $\boldsymbol{\rho}(\boldsymbol{\theta})$ is the model-implied polychoric correlation matrix for the estimated $\boldsymbol{\theta}$ trait parameters, and $\widehat{\mathbf{W}}$ is a positive definite weight matrix (Forero et al., 2009).

The categorical variable estimation methods differ in their weight matrix $\mathbf{W}$. In the case of the unweighted least squares (ULS) estimator, $\mathbf{W}$ is an identity matrix (Muthén, 1978), thereby making **F** a simple sum of squared model residuals $(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta}))^2$. For the weighted least squares (WLS) estimator, on the other hand, $\mathbf{W}$ is the inverse of the asymptotic variance-covariance matrix of polychoric correlations (Muthén, 1978). The dimension of this square matrix $\mathbf{W}$ is $p(p-1)/2$, which can only be efficiently estimated with very large sample sizes (Yang-Wallentin, Jöreskog, & Luo, 2010). As a means to partially sort out this difficulty, the diagonally weighted least squares (DWLS) estimator uses as $\mathbf{W}$ a weight matrix that only contains the diagonal elements of the asymptotic variance-covariance matrix of polychoric correlations (Rhemtulla, Brosseau-Liard, & Savalei, 2012). This estimator is also referred to as robust WLS or weighted least squares with mean and variance-adjusted standard errors (WLSMV). Both ULS and DWLS require the full weight matrix to compute the standard errors and the χ^2 test, which is mean and variance adjusted in the WLSMV case. These robust adjustments are necessary because ULS and DWLS are less efficient than WLS as a consequence of not using the full weight matrix (Rhemtulla et al., 2012; Yang-Walentin et al., 2010).

According to the factor analytic literature, the robust DWLS and ULS estimators perform well in the estimation of CFA and SEM models with categorical variables across a wide range of sample sizes and data characteristics (Flora & Curran, 2004; Forero et al., 2009; Lei, 2009; Nestler, 2013; Yang-Walentin et al., 2010). In addition, it appears that DWLS generally outperforms ULS in convergence rates (Forero et al., 2009), but ULS slightly outperforms DWLS in estimation accuracy (Forero et al., 2009; Savalei & Rhemtulla, 2013; Yang-Walentin et al., 2010). On the other hand, neither estimator is appropriate when the data characteristics are especially adverse, such as the intersection of small samples, few response categories, and highly skewed categorical variables (Forero et al., 2009; Savalei & Rhemtulla, 2013). In contrast to the DWLS and ULS estimators, the full WLS estimator is of limited usefulness because it tends to produce inflated χ^2 model fit statistics and negatively biased standard error estimates with categorical data that is typically found in applied research settings (Flora & Curran, 2004; Yang-Walentin et al., 2010). This estimator is therefore only recommended for very large sample sizes and small models (Flora & Curran, 2004).

Evaluation of Model Fit with Fit Indices

Numerous fit indices have been proposed in the factor-analytic literature as measures of the degree of fit of factor models (Hu & Bentler, 1999). These descriptive indices are generally favored against the statistical chi-square test of exact fit because psychometric models are known *a priori* to be false to some degree, and therefore will always be rejected with large enough samples (Browne & Cudeck, 1992; Yu, 2002). Some of the most commonly used fit indices are the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), the Tucker-Lewis index (TLI), and the Standardized Root Mean Square Residual (SRMR). These fit indices, which will be the focus of the current study, have performed relatively well in previous confirmatory factor analysis (CFA) and SEM Monte Carlo studies (e.g., Hu & Bentler, 1999; Sharma, Mukherjee, Kumar, & Dillon, 2005; Yu, 2002), and are highly popular in applied research (e.g., Campbell-Sills et al., 2004; Sanne et al., 2009).

RMSEA Index

$$\text{RMSEA} = \max\left(\sqrt{\frac{\lambda_\text{M}}{\text{df}_\text{M}(N-1)}}, 0\right) \qquad (1)$$

where λ_M is the noncentrality parameter of the specified model, df_M are the degrees of freedom of the specified model, and N is the sample size. The noncentrality parameter λ_M is computed as $\chi^2_\text{M} - \text{df}_\text{M}$, where χ^2_M is the chi-square statistic that tests the equivalence of the population covariance matrix of observed variables and the model-implied covariance matrix[1].

The RMSEA index is a measure of *absolute* fit that assesses the discrepancy due to approximation in the population, estimated as $\lambda_\text{M}/(\text{N} - 1)$, and corrected for model complexity through the division by the degrees of freedom, df_M. This index is intended to recover the model that maximizes verisimilitude (a model's proximity to the objective truth in the population) (Preacher, Zhang, Kim & Wells, 2013). In addition, RMSEA is a function of χ^2 and can be considered as a measure of *misfit detectability* that depends not only on the type/size of misfit, but also on the data characteristics and the accuracy of measurements (Browne, McCallum, Kim, Andersen, & Glaser, 2002). The RMSEA index is bounded below by zero, with lower values indicating a better fit to the data or less error of approximation. The CFA/SEM literature suggests that RMSEA values less than .08 and .05 are indicative of reasonable and close fit to the data, respectively (Browne & Cudeck, 1992; Chen, Curran, Bollen, Kirby, & Paxton, 2008; Marsh, Hau, & Wen, 2004; Yu, 2002).

CFI and TLI Indices

[1] Note that with categorical variables a robust χ^2 statistic is used to compute the fit indices. In the case of the M*plus* software, the robust χ^2 statistic is mean- and variance-adjusted. For more information see the M*plus* Technical Appendices (Muthén, 1998-2004).

$$\mathrm{CFI} = 1 - \frac{\max(\lambda_{\mathrm{M}}, 0)}{\max(\lambda_{\mathrm{N}}, \lambda_{\mathrm{M}}, 0)} \tag{2}$$

$$\mathrm{TLI} = 1 - \frac{\frac{\lambda_{\mathrm{M}}}{\mathrm{df}_{\mathrm{M}}}}{\frac{\lambda_{\mathrm{N}}}{\mathrm{df}_{\mathrm{N}}}} = 1 - \left(\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{N}}}\right)\left(\frac{\mathrm{df}_{\mathrm{N}}}{\mathrm{df}_{\mathrm{M}}}\right) \tag{3}$$

where λ_{N} and df_{N} are the noncentrality parameter and degrees of freedom of the baseline model, respectively.

The CFI and TLI indices are measures of *incremental* fit that assess the degree to which the specified model is superior to an alternative “baseline” model in reproducing the observed covariance matrix. The baseline model is usually a null model in which all the observed variables are uncorrelated (Hu & Bentler, 1999). The CFI index has boundaries of 0 and 1, with higher values indicating greater gains in fit in comparison to the baseline model. Likewise, the TLI index generally ranges from 0 to 1, but, as the index is not normed, it can sometimes obtain values that fall outside of this range. The TLI index differs from the CFI index in that it informs of the relative reduction in misfit *per degree of freedom*, an additional adjustment that takes into account model parsimony (Mahler, 2011). In addition, the values of TLI are always lower than those of CFI because the term that is subtracted from 1 in the formula is multiplied by $\mathrm{df}_{\mathrm{N}}/\mathrm{df}_{\mathrm{M}}$, which is always greater than one (Kenny & McCoach, 2003). On the other hand, the values of CFI and TLI tend to become more similar as the number of observed variables, p, increases, because as p increases the ratio of $\mathrm{df}_{\mathrm{N}}/\mathrm{df}_{\mathrm{M}}$ tends toward unity. According to the CFA/SEM literature, CFI and TLI values greater than .90 and .95 can be considered to reflect acceptable and excellent fit to the data (Hu & Bentler, 1999; Marsh et al., 2004; Yu, 2002).

SRMR Index

$$\mathrm{SRMR} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left(\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}}\sqrt{\hat{\sigma}_{jj}}}\right)^2}{p(p+1)/2}} \quad (4)$$

where s_{ij} is the observed covariance, $\hat{\sigma}_{ij}$ is the model-implied covariance, s_{ii} and s_{jj} are the observed standard deviations, $\hat{\sigma}_{ii}$ and $\hat{\sigma}_{jj}$ are the model-implied standard deviations, and p is the number of observed variables. In the case of categorical variable estimators, the covariances in the formula are substituted by the polychoric correlations and the standard deviations are replaced by their standardized value of unity.

The SRMR index is a measure of *absolute* fit that computes the standardized difference between the observed and model-implied covariance/correlation matrices. This index has a lower bound of zero, with smaller values indicating a better fit or less residual error. Because SRMR evaluates raw sample misfit and does not take into account the sample variability of the residuals, its values depend on the sample size and the characteristics of the model being estimated (Hu & Bentler, 1998). Values of SRMR lower than .08 have been found to suggest a good fit to the data (Hu & Bentler, 1999).

Performance of Fit Indices with CFA and SEM Models

Although this study is concerned with the accuracy of fit indices in the assessment of data dimensionality with unrestricted factor models, most of what is known about their empirical properties has come from CFA and SEM studies. Because this information could aid in understanding and anticipating how fit indices might perform in the estimation of the number of factors to retain with EFA/ESEM models, we will briefly summarize next the major findings from this literature.

The size of the *factor loadings* has been found to strongly impact the power of fit indices to detect model misfit (Browne et al., 2002; Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011; Mahler, 2011; Savalei, 2012). The fit indices that appear to be most affected by this variable are RMSEA and SRMR, sometimes indicating a close fit to the data for models that have gross misspecifications when the factor loadings are low, and other times suggesting a poor fit to the data for models that have only minor misspecifications when the factor loadings are high (Browne et al., 2002; Heene et al., 2011; Mahler, 2011; Saris, Satorra, & van der Veld, 2009; Savalei, 2012). In contrast to the behavior of RMSEA and SRMR, the CFI and TLI indices tend to exhibit poorer fit for models that have lower factor loadings (Heene et al., 2011; Mahler, 2011; Sharma et al., 2005). Part of the reason for this behavior of CFI and TLI appears to be that lower factor loadings entail lower covariances between the observed variables, which reduce the distance between the specified model and the baseline null model.

Sample size has also been shown to have a considerable impact on the performance of fit indices (Chen et al., 2008; Hu & Bentler, 1998, 1999; Nye & Drasgow, 2011; Yu, 2002), and its effects appear to interact with the number of *manifest variables* (Kenny & McCoach, 2003; Marsh, Hau, Balla, & Grayson, 1998; Sharma et al., 2005). The effects of sample size on the performance of fit indices are partly due to the behavior of the χ^2 statistic, which has a tendency to overestimate its theoretically expected values with small samples, leading, in turn, to overly high rejection rates (Curran et al., 2002; Marsh et al., 1998). Moreover, this upward bias in the χ^2 statistic can remain considerable even in larger samples, if the size of the model to be estimated is also large (Herzog, Boomsma, & Reinecke, 2007). This problem is further exacerbated with categorical variables that have few response options and high levels of skewness (Forero et al., 2009; Savalei & Rhemtulla, 2013). The incremental fit indices, even though they compare against a baseline model, are also affected because this upward bias in the χ^2 statistic is less pronounced for misspecified models, such as the baseline null model used in their computation (Curran et al., 2002). SRMR, although not χ^2 based, is even more dependent on the size of the sample, with fit values that decrease markedly with increasing sample sizes as a result of more precise estimations of the population covariances/correlations (Nye & Drasgow, 2011; Yu, 2002).

Accuracy of Fit Indices in the Estimation of the Number of Factors

There is limited information available regarding the accuracy of fit indices in the estimation of the number of factors. We are aware of only three studies that have systematically evaluated their performance with unrestricted factor models: Preacher et al. (2013) with continuous variables, Barendse et al. (2015) with continuous and categorical variables, and Yang and Xia (2014) with categorical variables. The major findings from this literature are summarized below.

First, RMSEA seems to select the number of major factors in the population more often when the sample sizes are larger, the factor loadings are higher, the factor structures are less complex, there are more response options, the factor correlations are smaller, or there are more variables per factor (Barendse et al., 2015; Preacher et al., 2013; Yang & Xia, 2014). With conventional cutoff values of .05 or .06, this index will tend to underfactor with 2-point scales or factor correlations of .50 (Yang & Xia, 2014), but may overfactor with small samples of 100 to 200 observations (Barendse et al., 2015; Preacher et al., 2013).

Second, the SRMR-based dimensionality decisions appear to be affected similarly to those of RMSEA by the levels of factor loadings, number of response options, and complexity of the factor structures (Barendse et al., 2015). However, SRMR has displayed the undesirable property of becoming less accurate with larger samples, where it appears to systematically select fewer major factors than those present in the population. These results may be attributed to the lower

SRMR values that are obtained in these conditions as a consequence of more precise correlation estimates (Barendse et al., 2015).

Third, little is known about the accuracy of incremental fit indices such as CFI and TLI. Only Yang and Xia (2014) evaluated an incremental fit index, CFI, and they reported that it performed similarly to or not as well as RMSEA and did not provide any further results for it.

Fourth, the WLSMV estimator seems to lead to more accurate estimations with categorical variables. When compared to other estimators, such as ML of covariances, ML of polychoric correlations, robust ML, and WLS of polychoric correlations, the WLSMV categorical variable estimator had the highest convergence rates and led to the best dimensionality estimates from various fit indices (Barendse et al., 2015).

Fifth, not much is known about the accuracy of fit indices in comparison to Horn's parallel analysis (PA; Horn, 1965). The PA method, which posits that factors should be retained as long as their eigenvalues are greater than the corresponding ones from samples of random variables that are uncorrelated at the population level, is arguably the most accurate retention method available at the moment (Henson & Roberts, 2006). Even though Yang and Xia (2014) included PA in their study, they used different criteria to evaluate its accuracy and those of the fit indices (mean values for the fit indices vs. percentage of selected models for PA), making any comparisons difficult to undertake.

Goals of the Current Study

Although previous studies with fit indices have provided valuable information regarding their performance in the estimation of the number of factors to retain, they contain several limitations that make it difficult to generalize their findings. For example, Preacher et al. (2013) and Yang and Xia (2014) only simulated variables with population loadings of .70 or greater, values that are notably high and which may not be representative of most research situations.

Also, the available studies have evaluated only a limited number of conditions (32 to 72), which means that relevant independent variables have either not been manipulated (e.g., the number of major factors was kept constant at 3 in both Barendse et al. and Yang and Xia) or have contained too few levels (e.g., only samples of 200 or 1,000 observations were evaluated in Barendse et al. and only variables with 2 or 4 response options were simulated in Yang and Xia). Further, only Barendse et al. (2015) evaluated the impact of choosing different cutoff values, and as Marsh et al. (2009) pointed out, the optimal cutoff values in EFA/ESEM may be different from those established in CFA, where the number of estimated parameters is usually much smaller. Thus, the main goal of this study was to address some of these limitations in the factor analytic literature by carrying out an in-depth analysis of the accuracy of four frequently used and recommended fit indices –CFI, TLI, RMSEA, and SRMR– in the estimation of the number of factors with *categorical variables*.

At the moment we are not aware of studies that have *compared* these four fit indices directly in the dimensionality assessment of the same data, a necessary step in order to determine their relative accuracy. In addition, whereas previous studies analyzed only a relatively small number of conditions, and in some cases only with continuous variables, this study considered a more comprehensive set of *factors* and *factor levels*, which produced a total 2,268 categorical variable conditions that enabled a deeper evaluation of these fit indices. Also, the fit indices were examined in this study across a larger than usual range of *cutoff values* in order to better understand their performance. Finally, the accuracy of the fit indices was assessed with the underlying continuous variables (prior to categorization) so as to establish a *baseline* for their accuracy with the categorical variables, and their estimations were compared against those of Horn's parallel analysis so as to better ascertain their practical usefulness.

Method

Study Design

Monte Carlo methods were employed to systematically assess the accuracy of the retention methods. In accordance with numerous simulation studies in the factor analytic literature (e.g., Forero et al., 2009; Nestler, 2013; Velicer, Eaton, & Fava, 2000), the simulation procedure involved the generation of factor models that had a simple structure design at the population level, with factor indicators only loading on one factor, variables possessing homogeneous properties (e.g., same factor loading, absolute skewness, response categories, and factor correlations), and without minor factors. Although this strategy does not take into consideration model error at the population level, or the empirical variability in the properties of the observed and latent variables, it allows for valuable insight to be gained by utilizing models that have known and unambiguous dimensionalities at the population level and by isolating the impact of precise values of the manipulated variables.

The factorial design included the manipulation of four "*structure*" factors –factor loading, number of variables per factor, number of factors, and factor correlation– and three "*sample*" factors –sample size, number of response categories, and skewness– for a total of seven independent variables. Altogether, these seven variables have been shown to affect the performance of factor retention methods with categorical variables (Barendse et al., 2015; Garrido, Abad, & Ponsoda, 2011, 2013; Timmerman & Lorenzo-Seva, 2011; Yang & Xia, 2014).

The levels for the independent variables were chosen so that they were representative of the range of values that are encountered in applied settings. In each case, an attempt was made to include a small/weak, medium/moderate, and large/strong level. A brief description of the rationale that was followed in the selection of the factor levels is presented next.

Factor loading (FLOAD): with levels of .40, .55, and .70, which can be considered as low, medium, and high, respectively (Cho, Li, & Bandalos, 2009). Similar factor loadings have also been generated in previous factor analytic studies with categorical variables (e.g., Forero et al., 2009; Nestler, 2013; Savalei & Rhemtulla, 2013).

Variables per factor (VARFAC): with levels of 4, 8, and 12, which include a value that is just over the minimum of 3 that is required for identification, another that denotes a moderately strong factor, and one for a highly overidentified factor (Velicer et al., 2000; Widaman, 1993).

Number of factors (FAC): with levels of 1, 2, and 4, which include the unidimensional condition as well as common number of traits for modern behavioral inventories (Henson & Roberts, 2006).

Factor correlation (FCORR): with levels of .00, .30, and .50, which include the orthogonal condition, plus moderate and strong correlation levels (Cohen, 1988).

Sample size (N): with levels of 100, 300, and 1,000, which may be considered as small, medium, and large, respectively, for the factor analysis of categorical variables (Forero et al., 2009; Muthén & Kaplan, 1985; Savalei & Rhemtulla, 2013).

Number of response categories (RESCAT): with levels of 2, 3, 4, 5, and continuous, which include all the possible numbers of response options below 6, where the results for categorical and continuous variable estimators tend to become highly similar (Rhemtulla et al., 2012).

Skewness (SKEW): with levels of 0, ±1, and ±2, which include the symmetrical condition as well as values that can be regarded as a meaningful departure from normality and a high level of skewness (Meyers, Gamst, & Guarino, 2006, p. 50; Muthén & Kaplan, 1985). The smaller levels of skewness are more typical of attitude tests and personality inventories, while larger levels of oppositely skewed categorical variables can be found on aptitude tests, where the items are designed to have difficulty levels that range from very easy to very difficult (Geranpayeh & Taylor, 2013, p.249; Rhemtulla et al., 2012).

Because some levels of the independent variables cannot cross with others (e.g., there are no factor correlations for the 1-factor condition), the 2,457 factor combinations derived from the factorial design are better broken up into these four completely crossed conditions:

(1) The *continuous unidimensional* conditions: with a 3 x 3 x 1 x 3 (FLOAD x VARFAC x FAC x N) factorial design, totaling 27 conditions.

(2) The *continuous multidimensional* conditions: with a 3 x 3 x 2 x 3 x 3 (FLOAD x VARFAC x FAC x FCORR x N) factorial design, totaling 162 conditions.

(3) The *categorical unidimensional* conditions: with a 3 x 3 x 1 x 3 x 3 x 4 (FLOAD x VARFAC x FAC x N x SKEW x RESCAT) factorial design, totaling 324 conditions.

(4) The *categorical multidimensional* conditions: with a 3 x 3 x 2 x 3 x 3 x 3 x 4 (FLOAD x VARFAC x FAC x FCORR x N x SKEW x RESCAT) factorial design, totaling 1,944 conditions.

Data Generation

For each of the 2,457 simulated conditions, 100 sample data matrices were generated according to the following common factor model procedure: first, the reproduced population correlation matrix (with communalities in the diagonal) was computed as:

$$\mathbf{R}_{\mathrm{R}} = \boldsymbol{\Lambda}\boldsymbol{\phi}\boldsymbol{\Lambda}^{\mathrm{T}} \tag{8}$$

where $\mathbf{R}_{\mathrm{R}}$ is the reproduced population correlation matrix, $\boldsymbol{\Lambda}$ is the population factor loading matrix, and $\boldsymbol{\phi}$ is the population factor correlation matrix.

The population correlation matrix $\mathbf{R}_{\mathrm{P}}$ was then obtained by inserting unities in the diagonal of $\mathbf{R}_{\mathrm{R}}$, thereby raising the matrix to full rank. The next step was performing a Cholesky decomposition of $\mathbf{R}_{\mathrm{P}}$, such that:

$$\mathbf{R}_{\mathrm{P}} = \mathbf{U}^{\mathrm{T}}\mathbf{U} \tag{9}$$

where $\mathbf{U}$ is an upper triangular matrix.

The sample matrix of continuous variables $\mathbf{X}$ was subsequently computed as:

$$\mathbf{X} = \mathbf{Z}\mathbf{U} \tag{10}$$

where $\mathbf{Z}$ is a matrix of random standard normal deviates with rows equal to the sample size and columns equal to the number of variables.

The sample matrix of categorical variables was obtained by applying a set of thresholds to $\mathbf{X}$ according to the specified levels response categories and skewness. The thresholds (τ) for the symmetric conditions were computed by partitioning the continuum from z = -3 to z = 3 at equal intervals. Thresholds for the asymmetric conditions were created so that as the skewness level increased, the observations were "piled up" in one of the extreme categories (see Garrido et al., 2011; Muthén & Kaplan, 1985). In addition, half of the variables of each factor were categorized with the same positive skewness and the other half with the same negative skewness. All threshold values used for this study are included in the Appendix.

All the sample data matrices were generated under the MATLAB programming environment (version R2010a; The MathWorks, Inc., 1984-2010). These sample matrices were subsequently inputted into the M*plus* program (version 6.11; Muthén & Muthén, 1998-2010), where the factor models were estimated and the fit values obtained. In order to obtain the fit values of the factor models, the normally distributed continuous variables were factorized using the ML estimator over Pearson correlations. In the case of the categorical variables, the WLSMV estimator over polychoric correlations was employed. The WLSMV estimator was selected as it has been shown to perform well with categorical data, and because among the categorical variable estimators, it is the most common method of analysis among practitioners (Savalei & Rhemtulla, 2013). As far as the PA method, it was programmed directly into MATLAB with code developed by the authors. In all cases, the polychoric correlations were computed using the maximum likelihood two-stage algorithms provided by Olsson (1979).

Estimation of the Number of Factors

The procedure used to estimate the number of factors with fit indices consisted of fitting sequential unrestricted factor models to the sample data. The process started by fitting a 1-factor model and comparing its fit to the pre-specified cutoff value of the fit index; if the model fit acceptably, the index suggested a 1-factor solution, if not, the number of factors was sequentially increased by 1 until a model with an acceptable fit was obtained. If no fit information was available due to non-convergence or lack of degrees of freedom, the extraction procedure was stopped and the number of factors was fixed at the last estimated value. For example, if a 1-factor model obtained an inadequate fit to the data but the subsequent 2-factor solution did not converge, the number of factors was fixed at 2. In other words, a factor model was not accepted if its level of fit did not reach the pre-specified cutoff value of the fit index, even if the subsequent model could not be tested. For each fit index considered in this study, 20 cutoff values were evaluated. In the case of CFI and TLI, 19 cutoff values were examined from .05 to .95 in increments of .05, while the 20^{th} cutoff value was .99. Regarding RMSEA and SRMR, the 20 cutoff values went from .20 to .01 in decrements of .01.

The estimation of the number of factors with PA, on the other hand, was carried out by comparing the eigenvalues from the sample matrices with underlying factors to those obtained from sample matrices of random variables that were uncorrelated at the population-level, but that otherwise had the same sample characteristics as the former (i.e., sample size, number of variables, skewness, and response categories). Additionally, the procedure was computed in accordance to the recommendations and simulation procedures described in Garrido et al. (2013), which included factorizing the full matrices of polychoric correlations and computing the mean eigenvalues from 100 sample matrices of independent variates.

Accuracy Criteria

The accuracy of the fit indices was evaluated according to three complementary criteria: the proportion of correct estimates (PC), the mean bias error (MBE), and the mean absolute error (MAE). The formulas for these criterion variables are presented in Equations 11-13:

$$\mathrm{PC} = \frac{\sum \mathrm{C}}{\mathrm{N_S}}, \quad \text{for} \quad \mathrm{C} = \begin{cases} 1 & \text{if } \hat{\theta} = \theta \\ 0 & \text{if } \hat{\theta} \neq \theta \end{cases} \tag{11}$$

$$\mathrm{MBE} = \frac{\sum(\hat{\theta} - \theta)}{\mathrm{N_S}} \tag{12}$$

$$\mathrm{MAE} = \frac{\sum |\hat{\theta} - \theta|}{\mathrm{N_S}} \tag{13}$$

where $\mathrm{N_S}$ is the number of sample data matrices generated for each condition (100), $\hat{\theta}$ is the estimated number of factors, and θ is the population number of factors.

The PC criterion has boundaries of 0 and 1, with 0 indicating a total lack of accuracy and 1 reflecting perfect accuracy. In contrast, a 0 on the MBE criterion shows a complete lack of bias, with negative and positive values indicating underfactoring and overfactoring, respectively. It is important to note that MBE cannot be used alone as a measure of method precision, because errors of under- and overfactoring can compensate each other (something that cannot happen with the PC or MAE criterion), creating a false illusion of accuracy. In terms of the MAE criterion, higher values signal larger absolute deviations from the population number of factors, while a value of 0 indicates perfect accuracy.

Results

Convergence Rates

The convergence rates given in this section indicate the proportion of cases that produced fit statistics for the *final model* estimated in the sequential factor extraction process. That is, it indicates the proportion of cases where the criterion cutoff value(s) was satisfied. Non-convergence was coded, on the other hand, when the iterative estimation process failed to converge (using the M*plus* default values) before the criterion cutoff value(s) had been satisfied, or when there were zero or negative degrees of freedom for a factor model that was to be tested.

With conventional cutoff value criteria (CFI > .95; TLI > .95; RMSEA < .05; SRMR < .08), the convergence rates for CFI, TLI, RMSEA, and SRMR, were 94.9%, 92.9%, 96.5%, and 92.6%, respectively. On the other hand, with the most stringent cutoff values evaluated for CFI (.99), TLI (.99), RMSEA (.01), and SRMR (.01), the convergence were 90.3%, 88.6%, 87.1%, and 15.4%, respectively. The substantial drop in the SRMR convergence rate suggests that it was very difficult to achieve a sample SRMR of .01 under the sample sizes that were considered (remember that the population SRMR was .00 for all structures). In contrast, the dimensionality estimates suggested by PA lead to an especially high convergence rate of 99.3%. It is important to note that of the non-converged models, .6% specified fewer factors than those in the population, 1.7% had the same number of factors, and 97.7% attempted to extract more factors. Thus, as in Barendse et al. (2015), overfactoring appears to have been the main reason for non-convergence in this study.

Agreement Between the Dimensionality Estimates

Lin's concordance correlation coefficient (Cc; Lin, 1989) was used to assess the level of agreement between the numbers of factors estimated by the retention methods. The Cc is a measure of absolute agreement for continuous variables that ranges from -1 to 1, with 1 indicating perfect agreement, -1 perfect disagreement, and 0 no agreement. In the specific case where two variables have the same means and standard deviations, Cc will be equal to Pearson's correlation coefficient; in all other instances, Cc will be lower in absolute value. The values of Cc were interpreted as follows: Cc < .20 was considered as *poor* agreement; .20 ≤ Cc < .40 *fair*; .40 ≤ Cc < .60 *moderate*; .60 ≤ Cc < .80 *good*; and .80 ≤ Cc ≤ 1.00 *very good*.

The levels of agreement for the categorical variables across cutoff values (cv) and methods are shown in Figure 1. In addition, Figure 1 includes the levels of agreement with the numbers of factors simulated at the population level. The commentary of these results will be organized in the following manner: first, the within agreement across cutoff values for each fit index; second, the between agreement across fit indices and cutoff values; and third, the agreement between the fit indices, parallel analysis, and the simulated/population factors.

PLEASE INSERT FIGURE 1 ABOUT HERE

According to the Cc heat maps shown in Figure 1, RMSEA only maintained a *very good* level of agreement across successive cutoff values, while SRMR achieved *very good* agreement across two cutoff values for the majority of the range that was evaluated. As far as CFI and TLI, although there was only *good* to *poor* agreement across successive cutoff values in the most liberal range (.05 to .25), there was *very good* agreement across two cutoff values for most of the range between the .30 and .99 cutoff values. In general, these results indicate that changes in cutoff value of more than .01 for RMSEA, more than .02 for SRMR, and more than .05 or .10 for CFI and TLI, produced notable changes in the number of factors that were estimated.

In terms of the levels of agreement across fit indices, CFI and TLI showed a similar pattern of agreement between them as they did within. The pattern, however, was slightly shifted, meaning that there was more agreement for CFI that had equal or higher cutoff values than TLI, than in the reverse case. This result was expected, as TLI will always be lower than CFI in the normed range between 0 and 1 (see Equations 2 and 3). For example, for CFI always one cutoff value lower than TLI, the mean Cc was .61; for CFI and TLI with equal cutoff values, the mean

Cc was .71; and for CFI always one cutoff value above TLI the mean Cc was .81. Also, the agreement became stronger with more stringent cutoff values, to the point where the estimations between these two indices became practically redundant at the higher end of cutoff values (e.g., Cc = .96 for CFI and TLI with .90 cv; Cc = .97 for CFI with .95 cv and TLI with .90 cv). Regarding their level of agreement with RMSEA, both obtained *very good* agreement for a portion of the intersection between the .90 to .99 cv for CFI/TLI and .01 to .02 cv for RMSEA (Cc_{max} = .96 for CFI/TLI with .99 cv and RMSEA with .01 cv). As far as the agreement between CFI/TLI and SRMR, a maximum agreement of *good* was achieved, and it occurred for parts of the crossing between CFI/TLI with .80 to .99 cv and SRMR with .05 to .11 cv (Cc_{max} = .74 for CFI/TLI with .99 cv and SRMR with .07 cv). Similarly, RMSEA and SRMR had a maximum agreement of *good*, which occurred at parts of the intersection of .01 to .03 cv for RMSEA and .06 to .11 cv for SRMR (Cc_{max} = .72 for RMSEA with .01 cv and SRMR with .07 cv).

Regarding the agreement of the fit indices with PA, both CFI (Cc_{max} = .72 for the .90 cv) and TLI (Cc_{max} = .72 for .90 cv) reached a maximum agreement of *good* with PA, while RMSEA and PA obtained a maximum agreement of *moderate* (Cc_{max} = .58 for the .02 cv), and SRMR and PA only achieved a level of *fair* agreement (Cc_{max} = .37 for the .08 and .09 cv). On the other hand, the method that had the highest agreement with the simulated factors was PA (Cc = .79), followed by CFI (Cc_{max} = .63 for .90 cv), TLI (Cc_{max} = .63 for .90 cv), RMSEA (Cc_{max} = .53 for .02 cv), and finally SRMR, which achieved an agreement of just *fair* (Cc_{max} = .34 for .08 and .09 cv). These latter results are particularly relevant as they assess the level of agreement with the number of factors in the population, thus making it also a measure of estimation accuracy.

Overall Accuracy Across Cutoff Values

A look at the overall accuracy of the fit indices across the different cutoff values is presented in Figures 2, 3, and 4. These figures summarize the performance of the fit indices according to each of the three dependent criterion variables, PC, MBE, and MAE. In order to make the results for the normal continuous variables (NCV) more directly comparable to those for the categorical variables, the latter were split into two groups: the unskewed (UOV) and the skewed (SOV) ordered-categorical variables. This way, the results for the NCV could be weighted against those obtained for the categorical variables with symmetric distributions. Furthermore, each graph includes a box plot for the parallel analysis method, so as to give proper context to the performance of the fit indices.

The results shown in Figures 2, 3, and 4, reveal that the behavior of the fit indices with NCV and UOV was highly congruent. As can be seen in these figures, the shapes of the box plots across the range of cutoff values are analogous for these two types of variables. Also, with the exception of SRMR, the peak levels of overall accuracy (highest mean PC, lowest mean MAE) were roughly equivalent for the NCV and the UOV. These results indicate that there was not a relevant loss in accuracy in the estimation of the number of factors when the NCV were categorized with symmetrical thresholds and subsequently factor analyzed with categorical variable estimators. In terms of the results for the SOV, the performance of all the fit indices tended to be less accurate (lower PC, higher MAE), and more variable at the ranges of peak accuracy (larger box plots, more extreme values), than for the UOV, signaling greater error in the estimation of the number of factors with skewed categorical variables. In this line, Figure 3 reveals that the MBE was higher for SOV than for UOV, with the former producing greater levels of overfactoring at the more stringent cutoff values (cv).

PLEASE INSERT FIGURE 2 ABOUT HERE

PLEASE INSERT FIGURE 3 ABOUT HERE

PLEASE INSERT FIGURE 4 ABOUT HERE

A comparison across fit indices and cutoff values in Figures 2 to 4 reveals that the three χ^2 based fit indices performed very similarly across the range of cutoff values that were evaluated, with RMSEA producing moderately larger variability across conditions and poorer *mean* accuracy levels ($\overline{\text{PC}}$, $\overline{\text{MBE}}$, $\overline{\text{MAE}}$) than CFI/TLI. The SRMR index, on the other hand, showed a notably worse performance, with extreme levels of overfactoring across the most stringent cutoff values (see Figure 3). Parallel analysis, on the other hand, was the most accurate out of all the methods, showing less variability across conditions, higher PCs, lower MAEs, and minimum levels bias for both continuous and categorical variables.

As far as the actual mean levels of accuracy obtained by the fit indices across the categorical variable conditions, the maximum $\overline{\text{PC}}$ in the UOV conditions was the .80 achieved by CFI (.95 cv), followed by .79 for TLI (.95 cv), .70 for RMSEA (.02 and .03 cv), and .57 for SRMR (.06 cv). Similarly, the lowest $\overline{\text{MAE}}$ for the fit indices in these conditions was the .28 obtained by CFI (.95 cv), tailed by .32 for TLI (.95 cv), .43 for RMSEA (.02 cv), and .84 for SRMR (.08). In the case of the SOV conditions, the maximum $\overline{\text{PC}}$ was the .69 produced by CFI (.95 cv), which was closely followed by the .67 of TLI (.95 cv), and then by the .59 of RMSEA (.02 cv), and the .45 of SRMR (.07 and .08 cv). The MAE criterion produced a similar ordering of the fit indices, with the minimum $\overline{\text{MAE}}$ of .60 obtained by CFI (.90 cv), and values of .64, .80, and 1.21, for TLI (.90 cv), RMSEA (.02 and .03 cv), and SRMR (.11 cv), respectively. These levels of accuracy were all inferior to the ones achieved by parallel analysis, which obtained a $\overline{\text{PC}}$ of .86, a $\overline{\text{MAE}}$ of .21, and a $\overline{\text{MBE}}$ of -.05 for the UOV conditions, and a $\overline{\text{PC}}$ of .78, a $\overline{\text{MAE}}$ of .36, and a $\overline{\text{MBE}}$ of .00, for the SOV conditions.

Accuracy Across Factor Levels and Cutoff Values

Due to the great similarity in the performance of the CFI and TLI indices, in particular for the most accurate ranges of cutoff values, only those results pertaining to CFI will be presented in this and the following sections. Also, and in order to limit the length of the manuscript, the MAE criterion will be the only one analyzed in an in-depth manner from this point forward. Although all three dependent variables considered in this study are highly informative and complement each other, the MAE statistic informs of the actual distance between the population and the estimated number of factors, which is especially relevant for applied research. The line plots corresponding to the MAE statistic across factor levels and cutoff values for the categorical variable conditions are presented in Figure 5.

Overall, the behavior of CFI and RMSEA across the levels of the independent variables was remarkably similar, while SRMR exhibited a markedly different pattern of performance. The general performance of CFI and RMSEA consisted of a gradual reduction in MAE with more stringent cutoff values until the next-to-last or second-to-last cutoff value, at which juncture the MAE started to increase (due to overfactoring). The accuracy of CFI and RMSEA, however, differed considerably across *factor loadings* and *factor correlations.* In the case of the factor loadings, while for CFI the MAEs were fairly similar across cutoff values for the different factor loadings, for RMSEA the MAEs varied considerably across a large portion of the range of cutoff values examined (≈ from .10 cv to .03 cv). In this regard, RMSEA needed increasingly more stringent cutoff values for a reduction in MAE as the factor loadings became weaker. Regarding the factor correlation variable, the aforementioned pattern was exactly reversed. Whereas RMSEA displayed similar MAEs across cutoff values for the different factor correlations, CFI needed increasingly more stringent cutoff values for a reduction in MAE as the factor correlations became stronger. In all, CFI produced accuracy levels that were slightly/moderately higher than those of RMSEA.

PLEASE INSERT FIGURE 5 ABOUT HERE

The most notable differences in the performance of SRMR were the extremely high MAEs that it produced at the most stringent cutoff values (cv ≤ .05), which reached magnitudes far greater than the ones obtained by the other fit indices. These results imply that much larger samples than those considered here are required for SRMR to approximate its population value (which was .00 for all the simulated structures). Another noteworthy result for SRMR was that for several variables a cutoff value that produced one of the lowest MAE for one level also produced one of the largest MAE for another level of the same variable. For example, with 1,000 cases SRMR achieved its lowest MAE of .37 with a cutoff value of .05, which, conversely, also produced an especially large MAE of 3.96 with 100 cases. Overall, SRMR produced the highest MAEs of all the fit indices at each factor level that was evaluated.

Regarding how the accuracy of the fit indices fared in comparison to PA, the latter produced the lowest MAE for 21 of the 22 factor levels that were evaluated. The one exception came with 4 variables per factor, where CFI obtained a MAE of .37 that was slightly lower than the .41 of PA. On the other hand, PA outperformed the fit indices by the greatest margin with 12 variables per factor ($MAE[PA] = .24 < MAE_{min}[CFI] = .60$), with 4 factors ($MAE[PA] = .50 < MAE_{min}[CFI] = .83$), and with skewness of ±2 ($MAE[PA] = .47 < MAE_{min}[CFI] = .79$).

Higher-Order Factor Interactions

The final series of analyses aimed to uncover potential patterns of performance that differed from the general ones presented in Figure 5. In order to carry out this goal, mixed Analyses of Variance (ANOVAs) were performed for each fit index, with cutoff value as the repeated measures *within-subjects* independent variable, the structure and sample factors as the *between-subjects* independent variables, and MAE as the *dependent* variable. Due to the especially poor performance of SRMR already evidenced in Figures 1 to 5, and in order to limit the length of the manuscript, no higher-order interactions affecting this index will be represented visually or commented on in this section. Similarly, only those higher-order interactions with large or near-large effect sizes will be presented. According to Cohen (1988), partial eta squared (η_p^2) effect sizes of .14 or greater can be considered as large effects. Because the repeated measures variable (CV) contained 20 levels, contrasts from order 1 up to order 19 could be tested. However, the results revealed that the highest effect sizes were consistently found for contrasts of order 1 "linear contrasts", of order 2 "quadratic contrasts", and of order 3 "cubic contrasts", so those will be the only ones presented here. It should be noted that the 1-factor condition was excluded from the ANOVAs because it did not cross with the factor correlation variable. The mixed ANOVA effect sizes for the CFI, RMSEA, and SRMR indices are shown next in Table 1.

PLEASE INSERT TABLE 1 ABOUT HERE

There were 3 three-way interactions that reached a large effect size for the CFI index: CV x VARFAC x N, CV x N x SKEW, and CV x FAC x FCORR. In addition, the four-way CV x VARFAC x N x SKEW interaction obtained a near-large effect size (η_p^2[linear] = .13). Similar to the CFI index, RMSEA also produced 3 three-way interactions that reached a large effect size, CV x VARFAC x N, CV x N x SKEW, and CV x FLOAD x FAC, which was the most salient (η_p^2[linear] = .31; η_p^2[cubic] = .24). Also, the same four-way CV x VARFAC x N x SKEW interaction obtained a notable effect size for RMSEA as well (η_p^2[linear] = .10). This four-way interaction, which contains 2 of the 3 salient three-way interactions, and the remaining three-way interactions (CV x FAC x FCORR for CFI and CV x FLOAD x FAC for RMSEA), are shown in Figure 6. Because the four-way interactions for CFI and RMSEA were nearly identical, only the one for CFI is represented in the Figure.

PLEASE INSERT FIGURE 6 ABOUT HERE

The three-way CV x FAC x FCORR interaction for CFI consists of the following patterns: (1) for each level of factor correlation the MAEs for 2 and 4 factors were separated by the largest magnitude with very liberal cutoff values (due to maximum underfactoring), but as the cutoff values become more stringent, the MAEs became gradually closer (due to a convergence towards the correct solution); and (2) with stronger factor correlations, more stringent cutoff values were needed for the MAEs to show a reduction and ultimately reach its minimum values, leading to a notable difference in the optimal cutoff values for the different levels of factor correlation. For example, with 2 factors the optimal cutoff values were .80, .85, and .95, for factor correlations of .00, .30, and .50, respectively. Similarly, with 4 factors the optimal cutoff values were .85, .95, and .95, for these same corresponding factor correlations.

In terms of the three-way CV x FLOAD x FAC interaction for RMSEA, the pattern was as follows: (1) for each level of factor loading the MAEs for 2 and 4 factors were separated by the largest magnitude with very liberal cutoff values, but as the cutoff values become more stringent, the MAEs became gradually closer; and (2) with weaker factor loadings, more stringent cutoff values were needed for the MAEs to show a reduction and ultimately reach its minimum values, leading (similarly to CFI) to a notable difference in the optimal cutoff values for the different levels of factor loading. In this regard, with 2 factors the optimal cutoff values were .03, .05, and .07, for factor loadings of .40, .55, and .70, respectively, whereas with 4 factors the optimal cutoff values were .01, .02, and .03, for these respective factor loadings.

The four-way CV x VARFAC x N x SKEW interaction for CFI is also shown in Figure 6. Because the factor structures that were simulated had no population error, the normal pattern for the MAEs with a “large-enough” sample would be to gradually decrease across the range of cutoff values. This pattern of results can generally be seen, for example, in the conditions with the largest sample size (1,000) or with the smallest number of variables per factor (4). However, when the ratio of sample size to variables became smaller, a notable increase in MAE was produced across the most stringent cutoff values (e.g., with N = 100 and VARFAC ≥ 8; with N = 300 and VARFAC = 12). In addition, this increase in MAE was *greater* with larger absolute skewness and also with smaller samples, which is the reason why the four-way interaction arose. These results are especially relevant because earlier it was seen that the most stringent cutoff values generally produced the lowest MAEs, but as can be seen in Figure 6, this finding does not apply to certain data conditions. Further, the distance in optimal cutoff values was sometimes quite large depending on the combination of the factor levels of these variables. For example, with 12 variables per factor and skewness of ±2, the optimal cutoff values for CFI were .65, .90, and .95, for samples of 100, 300, and 1,000 observations, respectively.

In terms of the comparison with PA, both CFI and RMSEA generally produced minimum MAEs with 2 factors that were approximately equal to the MAEs of PA (albeit for varying cutoff values across some factor levels), but PA was moderately more accurate with structures of 4 factors. Also, when the ratio of sample size to variables was larger, CFI/RMSEA obtained minimum MAEs that were generally similar to those of PA. However, when the ratio became smaller (and in particular with skewness of ±2), PA outperformed these fit indices by a considerable margin.

Discussion

Researchers in the social and behavioral sciences have been using fit indices to estimate the number of factors underlying sets of observed variables as part of a coherent validation strategy in which the fit assessment of the measurement model is not divorced from the dimensionality decision (e.g., Campbell-Sills et al., 2004; Tepper & Hoyle, 1996). This synergy between dimensionality and model fit assessment has been further propelled by the advent of exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009). Within the ESEM framework, researchers can explore unrestricted factor structures with all the measures of fit and model diagnostics that were available decades earlier for confirmatory factor analysis (CFA) and structural equation modeling (SEM). However, despite this increased use of fit indices to estimate data dimensionality, the systematic evaluation of their accuracy in this area has so far been scarce (Frazier & Youngstrom, 2007), with only a few recent studies attempting to address this issue (e.g., Barendse et al., 2015; Preacher et al., 2013; Yang and Xia, 2014). The current study, subsequently, sought to further reduce this gap in the literature by examining the accuracy of four commonly used fit indices –CFI, TLI, RMSEA, and SRMR– in the estimation of the number of factors with categorical variables, which are typically encountered in the human sciences (Flora & Curran, 2004).

A unique feature of this study was the examination of the fit indices across wide ranges of cutoff values which allowed to capture the majority of their practical range, going from maximum underfactoring to maximum overfactoring, and including their maximum estimation accuracy somewhere in between. This approach, in combination with the manipulation of a large number of independent variables and factor levels, as well as the evaluation of estimation accuracy from the perspective of different complementary criteria, enabled a broader look into the performance of fit indices as dimensionality assessment methods.

Main Findings

An initial set of analyses intended to compare the accuracy of fit indices with *continuous* versus *categorical* variables. Because much less is known about the performance of fit indices with categorical variables and estimators, it was important to establish whether the results obtained in this study were particular to the methods related to this level of measurement or if they could be generalizable across types of variables and estimators. In this regard, the chi-square based fit indices –CFI, TLI, and RMSEA– produced remarkably similar levels of accuracy for unskewed categorical variables (WLSMV estimator) and the "pre-categorization" normal continuous variables (ML estimator). These findings extend previous CFA/SEM research, which have shown the robust categorical variable estimators perform well across a variety of sample sizes and data characteristics (e.g., Flora & Curran, 2004; Forero et al., 2009; Lei, 2009; Nestler, 2013; Yang-Walentin et al., 2010). In contrast, the accuracy of SRMR was notably lower for categorical variables, in particular across the most stringent cutoff values, where it tended to overfactor at much larger rates than with continuous variables. On the other hand, all the fit indices produced substantially poorer dimensionality estimates for skewed categorical variables, with a notable bias toward overfactoring across the cutoff values that produced the best estimates for the unskewed conditions. These are not unexpected findings, as the categorical estimators tend to produce inflated model fit statistics with skewed variables, and the polychoric correlations have larger sampling errors when the indicators differ in skew (Forero et al., 2009; Timmerman & Lorenzo-Seva, 2011; Savalei & Rhemtulla, 2013).

In terms of the *differential* accuracy of the fit indices in the estimation of the number of factors with categorical variables, CFI and TLI produced the highest levels of accuracy, followed at a step below by RMSEA, and then by SRMR, which provided notably poor dimensionality estimates. These results are in line with Mahler (2011), who found CFI/TLI to be superior to RMSEA and SRMR in the detection of latent misspecification for CFA population models. Also, and in line with Yu (2002), the decisions based on these two indices were extremely similar, making them redundant for practical purposes. It should be noted that, as derived from their formulas, TLI always produces lower values than CFI, leading to slightly higher number of factor estimates for the same cutoff values. In general, changes in cutoff value greater than .05 or .10 for CFI/TLI, .01 for RMSEA, and .01 or .02 for SRMR, resulted in meaningfully different dimensionality estimates.

A controversial issue regarding the usefulness of fit indices for the evaluation of latent variable models is the appropriateness of applying *fixed* cutoff values (Chen et al., 2008; Heene et al., 2011; Marsh et al., 2004; Saris, Satorra, & van der Veld, 2009). Unfortunately, the findings from this study appear to further fuel these concerns by evidencing substantial problems in the performance of cutoff values across factor models and measurement conditions. In this respect, all the fit indices showed notable interactions between their estimation accuracy across cutoff values and the population and sample properties of the data. For all four fit indices examined, although more markedly for SRMR, the pattern of performance across cutoff values interacted strongly with the number of variables per factor, the sample size, and the skewness of the categorical variables. That is, the same cutoff values yielded more factors –for the same number of factors in the population– when small samples were combined with many variables per factor and high levels of skewness. This led to important fluctuations in the optimal cutoff values for the fit indices across conditions, in particular for SRMR. These findings are consistent with the CFA/SEM literature, which has shown that under these data conditions the chi-square statistic of the WLSMV estimator tends to be upwardly biased, over-rejecting correctly specified models (Forero et al., 2009; Savalei & Rhemtulla, 2013). In the case of SRMR, it is important to consider that it is an index that evaluates raw sample misfit and does not take into account the sample variability of the residuals, a characteristic that may make it more susceptible to the large sampling errors of the polychoric correlations (see also Yu, 2002).

In addition to the aforementioned results, CFI and TLI also displayed strong interactions between their accuracy across cutoff values and the magnitude of the factor correlations (the same cutoff values tended to estimate fewer factors –for the same number of factors in the population– with stronger factor correlations), while for RMSEA the performance across cutoff values interacted with the factor loadings (the same cutoff values tended to estimate fewer factors

–for the same number of factors in the population– with weaker factor loadings). Further, these patterns became more pronounced with structures that had higher population dimensionalities. This latter finding further extends previous CFA/SEM research where RMSEA has displayed a tendency to accept highly misspecified models when the observed variables have large unique variances (Heene et al., 2011; Mahler, 2011; Savalei, 2012). A theoretical explanation for this behavior of RMSEA has been given in Heene et al. (2011), who showed that increasing uniquenesses leads to a considerable loss of statistical power of the chi-square test and sensitivity of the chi-square based fit indices, which subsequently fail to reject models with even strong model misspecification. Although this characteristic should apply to all chi-square based fit indices, it is not observed for the incremental fit indices because the improvement of a given model over the null model becomes smaller with weaker factor loadings, thus flagging misspecified models as increasingly misfitting (Heene et al., 2011).

The current study also evaluated the usefulness of the fit indices by comparing them to what is arguably the most accurate factor retention method available at the moment, Horn's *parallel analysis*. In this regard, the findings were generally consistent: parallel analysis was more accurate than the fit indices across the different factor models and criterion variables that were considered, showing higher mean accuracy levels and less variability across conditions. This superiority of parallel analysis was especially evident in conditions where the ratio of variables to sample size was small and the variables were skewed. It thus appears that larger samples are needed for the fit indices to provide useful information about the fit of a given model than what is needed to assess the dimensionality of set of categorical variables with parallel analysis.

Limitations

The current study has some limitations that need to be considered. As noted in the Method section, all of the structures that were simulated had a simple structure design at the population level, with homogeneous indicator and factor properties and without minor factors. Although this strategy has some important benefits, such as the generation of structures with unambiguous dimensionalities, it limits the generalizability of the findings. For example, it is likely that more liberal cutoff values than those found here would be needed with empirical data, where the factor structures generally contain non-negligible levels of population error. In addition, future studies are required to determine the impact of including minor factors and heterogeneous data properties in the relative or comparative accuracy of the fit indices and parallel analysis.

Another limitation of this study, despite its large number of simulated conditions and in-depth evaluation of several commonly used fit indices, is that it only included one categorical variable estimator and may have excluded other relevant fit indices. In this line, future studies could examine estimators such as robust ULS or the polychoric instrumental variable estimator (PIV), which have been shown to work well in the estimation of factor models with categorical variables (Nestler, 2013). Furthermore, the accuracy of some fit indices might be enhanced by using complementary information, such as the confidence intervals associated with RMSEA (Preacher et al., 2013), or by applying the Hull method, which examines the plots of the fit indices' values against the degrees of freedom corresponding to the series of factor solutions (Lorenzo-Seva, Timmerman, & Kiers, 2011).

Practical Implications

The title of this manuscript posited the question: are fit indices really fit to estimate the number of factors with categorical variables? Given the findings from this study, as well as the current factor-analytic literature, the answer would have to be a less than favorable one. On one hand, the estimations by the fit indices display substantial interactions between the cutoff values chosen and the population and sample the properties of the data. This is particularly detrimental in terms of their applied usefulness, as researchers generally do not know the population properties of the data their analyzing and will have a hard time determining the optimal cutoff values for their particular datasets. On the other hand, even if the optimal cutoff values were somehow known in advance, the findings from this study indicate that parallel analysis would still be a better dimensionality estimator for the overwhelming majority of factor models. Consequently, we have to recommend that for the moment applied researchers lean primarily on the dimensionality estimates provided by *parallel analysis*. In the scenario that fit indices were used, CFI/TLI and RMSEA are clearly better choices than SRMR, which we believe should not be interpreted with categorical variables (see also Yu, 2002). In either case, we encourage researchers to perform Monte Carlo simulation studies in order to estimate the sample size required to produce "good-enough" dimensionality estimates for the type of models and retention methods they wish to evaluate and employ (see Muthén & Muthén, 2002, for more information).

It is important to emphasize that whatever factor retention methods or cutoff values researchers may wish to use, they should not be treated as inviolable or infallible rules that trump all other considerations. In this line, we strongly echo the message of other researchers (e.g., Chen et al., 2008; Marsh et al., 2004) that the appropriateness of factor models should not be based solely on statistical information, but also on substantive and theoretical considerations that require human judgment. Thus, all statistical methods ought to be employed as *aids* and not rules in the determination of the number of factors to retain.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438. doi:10.1080/10705510903008204

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling*, *22*, 87-101. doi:10.1080/10705511.2014.934850

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186-203. doi:10.1207/s15328007sem1302_2

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230-258. doi: 10.1177/0049124192021002005

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,* 403–421. doi:10.1037/1082-989X.7.4.403

Campbell-Sills, L., Liverant, G. I., & Brown, T. A. (2004). Psychometric evaluation of the Behavioral Inhibition/Behavioral Activation scales in large sample of outpatients with anxiety and mood disorders. *Psychological Assessment*, *16*(3), 244-254. doi:10.1037/1040-3590.16.3.244

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research, 36*, 462-494. doi:10.1177/0049124108314720

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement*, *69*(5), 748-759. doi: 10.1177/0013164409332229

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, Hillsdale, NJ.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research*, *37*(1), 1-36. doi:10.1207/S15327906MBR3701_01

Ferrando, P. J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica, 21*(3), 301-323.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491. doi:10.1037/1082-989X.9.4.466

Floyd, F J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286-299. doi:10.1037/1040-3590.7.3.286

Forero, C., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*, 625-641. doi:10.1080/10705510903203573

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*, *35*, 169-182. doi:10.1016/j.intell.2006.07.002

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement, 71*(3), 551-570. doi:10.1177/0013164410389489

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*(4), 454-474. doi: 10.1037/a0030005

Geranpayeh, A., & Taylor, L. (Eds) (2013) *Examining Listening: Research and practice in assessing second language listening, Studies in Language Testing 35*. Cambridge: UCLES/Cambridge University Press.

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. doi:10.1037/a0024917

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416. doi:10.1177/0013164405282485

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. doi:10.1007/BF02289447

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

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

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347-387. doi:10.1207/S15327906347-387

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and *Psychological Measurement*, *20*, 141-151. doi:10.1177/001316446002000116

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

Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality & Quantity, 43*, 495-507. doi:10.1007/s11135-007-9133-z

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255-268. doi:10.2307/2532051

Lorenzo-Seva, U., Timmerman, M. E., Kiers, H. A. L. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340-364. doi: 10.1080/00273171.2011.564527

Mahler, C. (2011). The effects of misspecification type and nuisance variables on the behaviors of population fit indices used in structural equation modeling. Unpublished doctoral dissertation, University of British Columbia.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*(2), 181-220. doi:10.1207/s15327906mbr3302_1

Marsh, H. W., Hau, K. T., & 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, 11*, 320-341. doi:10.1207/s15328007sem1103_2

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*(3), 439-476. doi:10.1080/10705510903008220

Meyers, L.S., Gamst, G., & Guarino, A. (2006). *Applied multivariate research: Design and interpretation.* Thousand Oaks, CA: Sage Publishers.

Morata-Ramírez, M. A., & Holgado-Tello, F. P. (2013). Construct validity of Likert scales through confirmatory factor analysis: A simulation study comparing different methods of estimation based on Pearson and polychoric correlations. *International Journal of Social Science Studies*, *1*(1), 54-61. doi:10.11114/ijsss.v1i1.27

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling, 7*(1), 36-73. doi:10.1207/S15328007SEM0701_02

Muthén, B. (1998-2004). *Mplus technical appendices*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551-560. doi:10.1007/BF02293813

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. doi:10.1111/j.2044-8317.1985.tb00832.x

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599-620. doi:10.1207/S15328007SEM0904_8

Myers, N. D. (2013). Coaching competency and (exploratory) structural equation modeling: A substantive-methodological synergy. *Psychology of Sport and Exercise*, *14*, 709-718. doi:10.1016/j.psychsport.201

Nestler, S. (2013). A Monte Carlo study comparing PIV, ULS and DWLS in the estimation of dichotomous confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *66*, 127-143. doi:10.1111/j.2044-8317.2012.02044.x

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3), 548-570. doi:10.1177/1094428110368562

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*(4), 443-460. doi:10.1007/BF02296207

Patil, V., Singh, S., Mishra, S., & Donavan, D. (2008). Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one' criterion. *Journal of Business Research, 61*(2), 162-170. doi:10.1016/j.jbusres.2007.05.008

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*, 28-56. doi:10.1080/00273171.2012.710386

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4).

Sanne, B., Torsheim, T., Heiervang, E., & Stormark, K. M. (2009). The strengths and difficulties questionnaire in the Bergen child study: A conceptually and methodically motivated structural analysis. *Psychological Assessment, 21*(3), 352-364. doi:10.1037/a0016317

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

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

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. doi:10.1111/j.2044-8317.2012.02049.x

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*, 935-943. doi:10.1016/j.jbusres.2003.10.007

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. doi:10.1007/BF02294363

Tepper, K. and Hoyle, R. H. (1996). Latent variable models of need for uniqueness. *Multivariate Behavioral Research*, *31*, 467-493. doi:10.1207/s15327906mbr3104_4

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. doi:10.1037/a0023353

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41(3), 321-327. doi:10.1007/BF02293557

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin, & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy.* (pp. 41-71). New York, NY, US: Kluwer Academic/Plenum Publishers. doi:10.1007/978-1-4615-4397-8_3

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28*(3), 263-311. doi:10.1207/s15327906mbr2803_1

Yang, Y., & Xia, Y. (2014, June 20). On the number of factors in exploratory factor analysis for ordered categorical data. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-014-0499-2

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*(3), 392-423.

Yu, C.Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Unpublished doctoral dissertation, University of California, Los Angeles.

Appendix

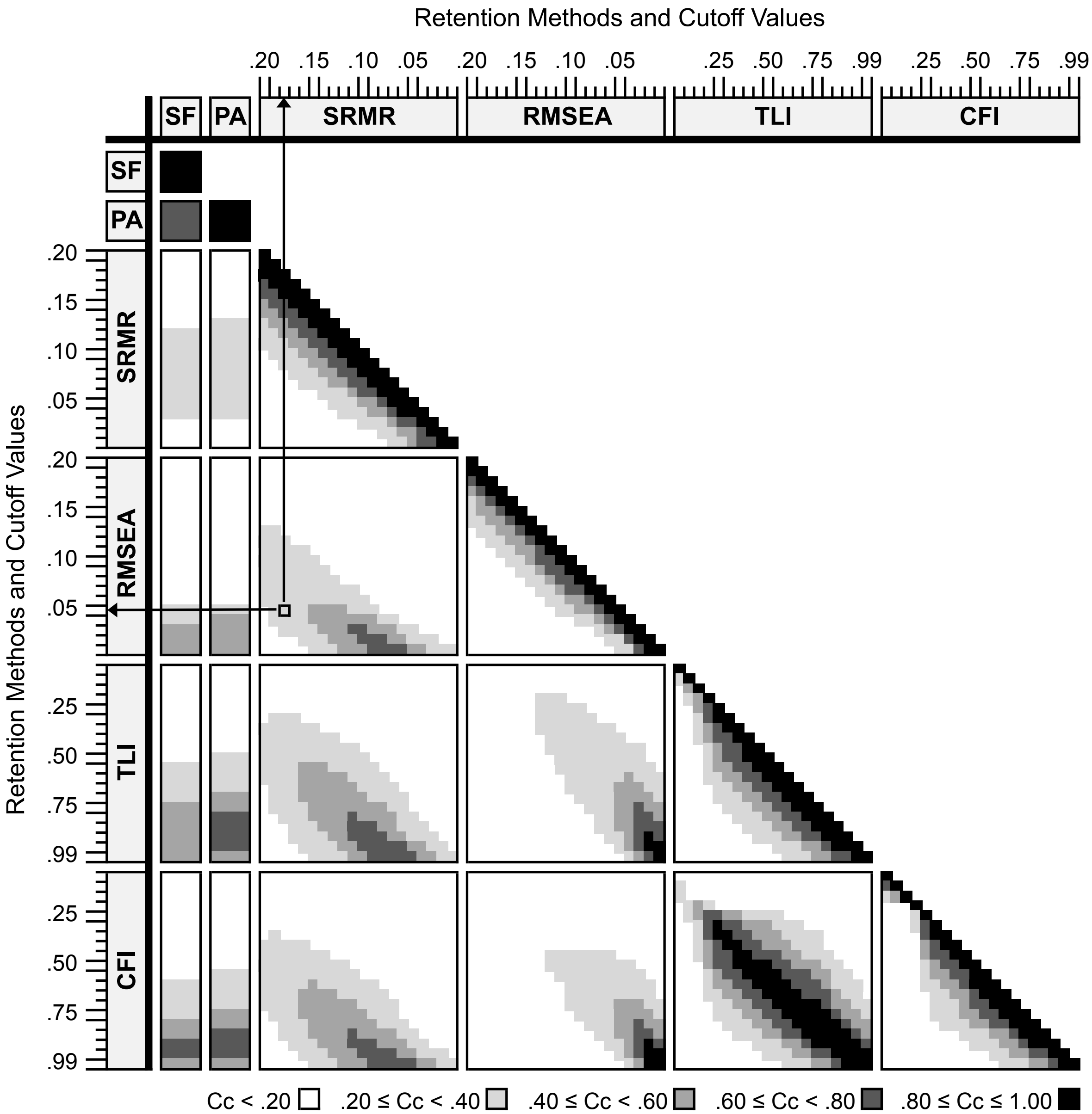
The thresholds (τ) for the symmetric conditions were: for 2 categories, $\tau_1 = 0.00$; for 3 categories, $\tau_1 = -1.00$, $\tau_2 = 1.00$; for 4 categories, $\tau_1 = -1.50$, $\tau_2 = 0.00$, $\tau_3 = 1.50$; for 5 categories, $\tau_1 = -1.80$, $\tau_2 = -0.60$, $\tau_3 = 0.60$, $\tau_4 = 1.80$. Thresholds for the asymmetric conditions with skewness level of +1 were: for 2 categories, $\tau_1 = 0.59$; for 3 categories, $\tau_1 = 0.32$, $\tau_2 = 0.99$; for 4 categories, $\tau_1 = 0.17$, $\tau_2 = 0.69$, $\tau_3 = 1.25$; for 5 categories, $\tau_1 = 0.05$, $\tau_2 = 0.51$, $\tau_3 = 0.94$, $\tau_4 = 1.45$. Thresholds for the asymmetric conditions with skewness level of +2 were: for 2 categories, $\tau_1 = 1.05$; for 3 categories, $\tau_1 = 0.85$, $\tau_2 = 1.38$; for 4 categories, $\tau_1 = 0.75$, $\tau_2 = 1.13$, $\tau_3 = 1.60$; for 5 categories, $\tau_1 = 0.68$, $\tau_2 = 1.00$, $\tau_3 = 1.34$, $\tau_4 = 1.77$. The thresholds for the negative skewness levels were obtained by changing the signs of the thresholds used to generate positively skewed categorical variables.

Table 1

Mixed Analysis of Variance Effect Sizes for the Fit Indices

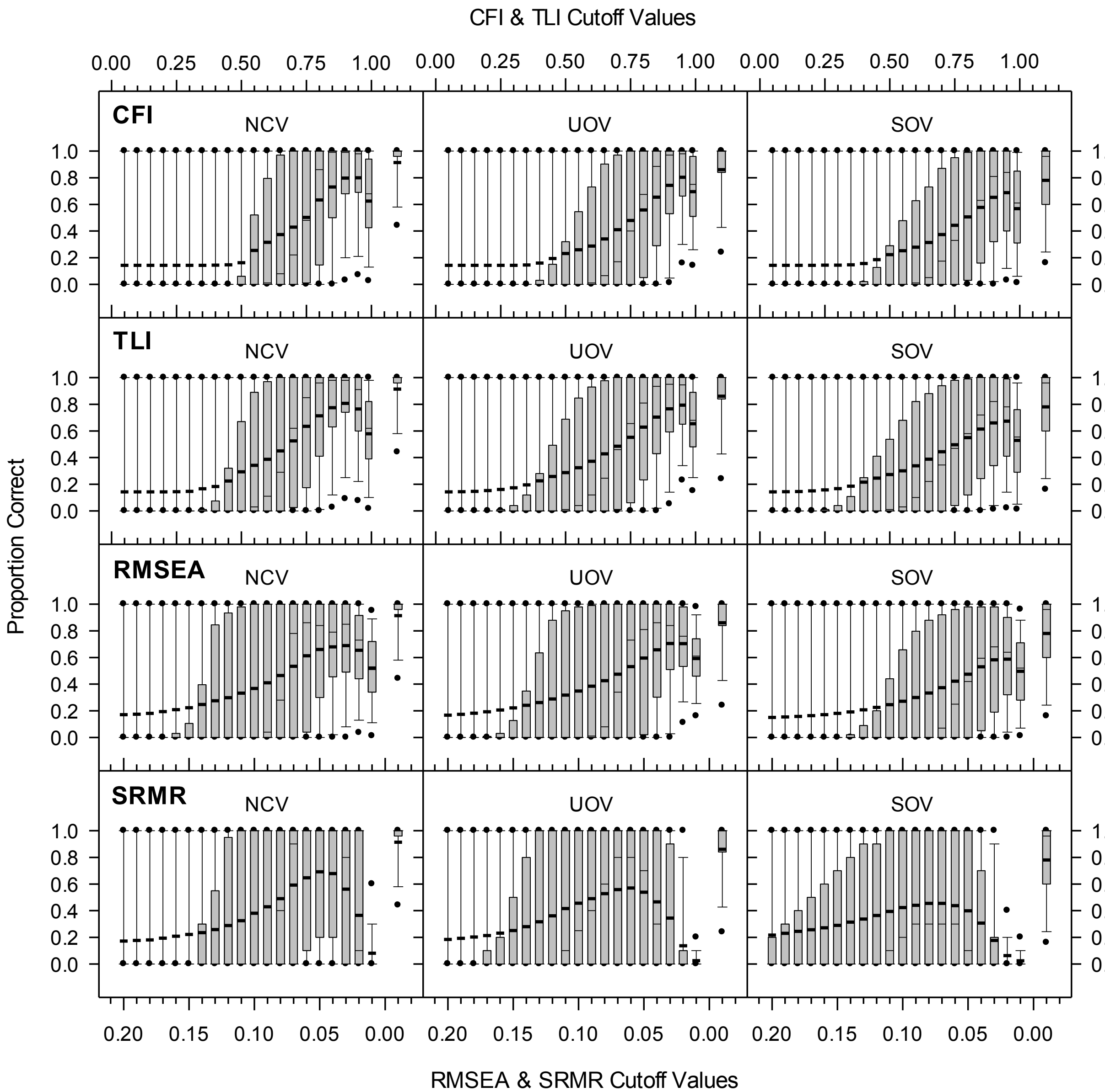
Effect Type		CFI			RMSEA			SRMR		
Variables		Lc	Qc	CUc	Lc	Qc	CUc	Lc	Qc	CUc
Main Effects										
CV	(Cutoff Value)	**.88**	**.22**	**.20**	**.84**	**.32**	.10	**.73**	**.85**	**.75**
Two-Way Interactions										
CV * FLOAD	(Factor Loading)	**.16**	.07	.01	**.38**	**.24**	**.37**	.06	.02	.03
CV * VARFAC	(Variables per Factor)	.06	**.14**	.12	.09	.01	.02	**.75**	**.66**	**.30**
CV * FAC	(Number of Factors)	**.61**	.12	.01	**.62**	**.47**	.06	.13	**.41**	**.38**
CV * FCORR	(Factor Correlation)	**.27**	**.45**	**.42**	.05	.06	.07	.01	.03	.00
CV * N	(Sample Size)	**.33**	**.30**	**.23**	**.36**	**.27**	**.18**	**.55**	**.14**	**.49**
CV * RESCAT	(Response Categories)	.01	.03	.03	.01	.06	.05	**.14**	.01	.12
CV * SKEW	(Skewness)	**.18**	.12	.04	**.23**	.04	.01	.08	.12	**.37**
Three-Way Interactions										
CV * FLOAD * FAC		.03	.01	.00	**.31**	.05	**.24**	.05	.00	.01
CV * VARFAC * FAC		.01	.03	.02	.04	.00	.01	**.29**	**.27**	.12
CV * VARFAC * N		**.17**	**.18**	.11	**.14**	.13	.10	**.31**	.03	**.24**
CV * VARFAC * SKEW		.08	.07	.02	.09	.08	.02	.01	.13	**.18**
CV * FAC * FCORR		.08	**.14**	.07	.06	.00	.04	.00	.01	.00
CV * FAC * N		.06	.07	.07	.02	.02	.06	**.14**	.09	.05
CV * N * SKEW		.11	**.18**	.07	.10	**.15**	.07	.02	**.37**	**.29**
Four-Way Interactions										
CV * VARFAC * N * SKEW		.13	.10	.02	.10	.07	.03	.04	**.21**	.07
CV * N * RESCAT * SKEW		.08	.07	.02	.08	.07	.02	.02	.04	**.15**

Note. Tabled values are partial eta squared (η_p^2) estimates of variance explained by each of the effects shown. The dependent variable was the mean absolute error in the estimation of the number of factors. Large effect sizes ($\eta_p^2 \geq .14$) are bolded and underlined. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; Lc = Linear Contrast; Qc = Quadratic Contrast; CUc = Cubic Contrast. $p < .01$ for all the effects shown in the table.



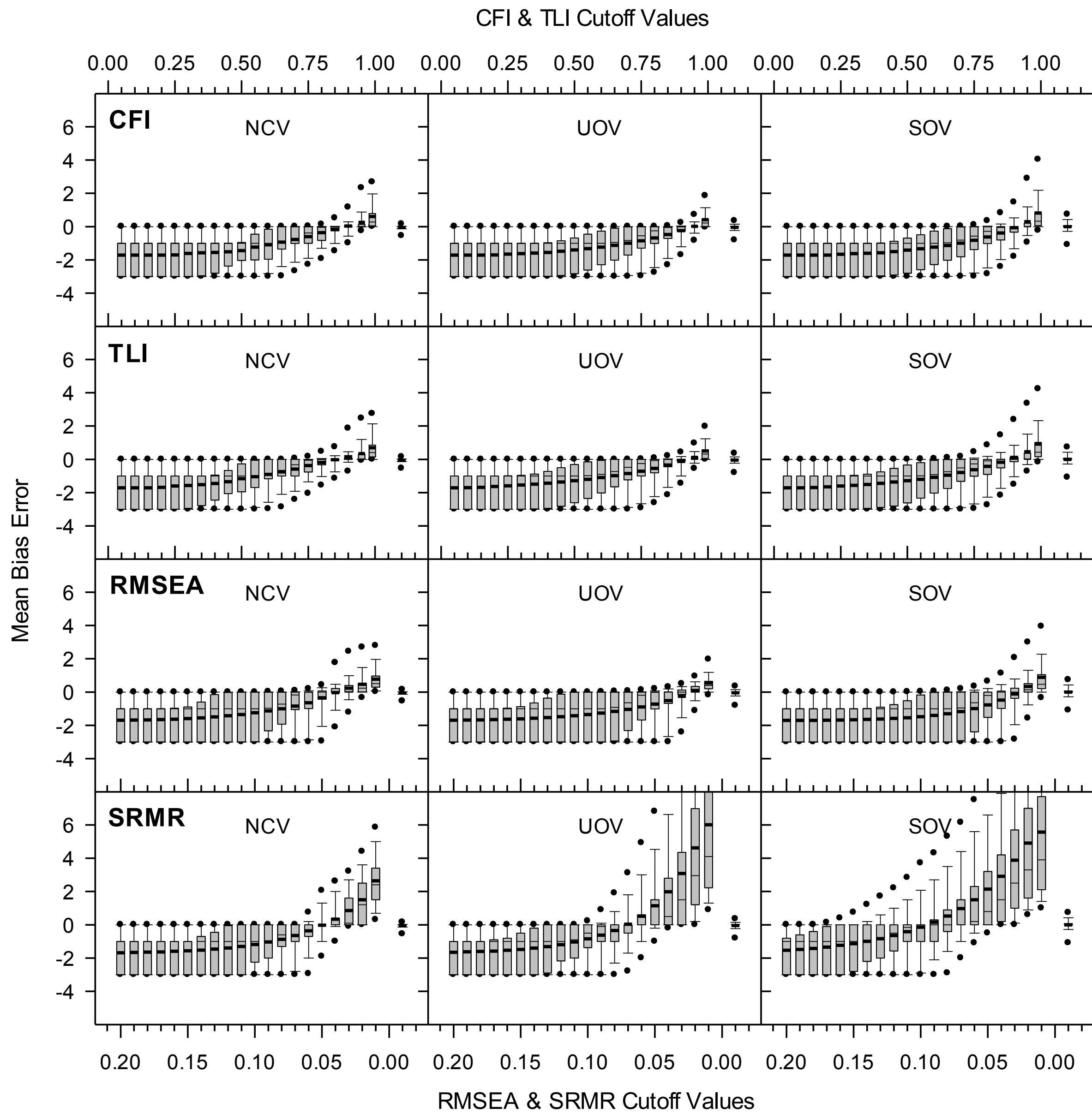
Note. SF = Simulated Factors (factors present in the population); PA = Parallel Analysis; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; Cc = Lin's Concordance Correlation Coefficient. The square highlighted in the figure shows the agreement between SRMR with a .18 cutoff value and RMSEA with a .05 cutoff value.

Figure 1: *Retention Method Agreement in the Estimation of the Number of Factors*



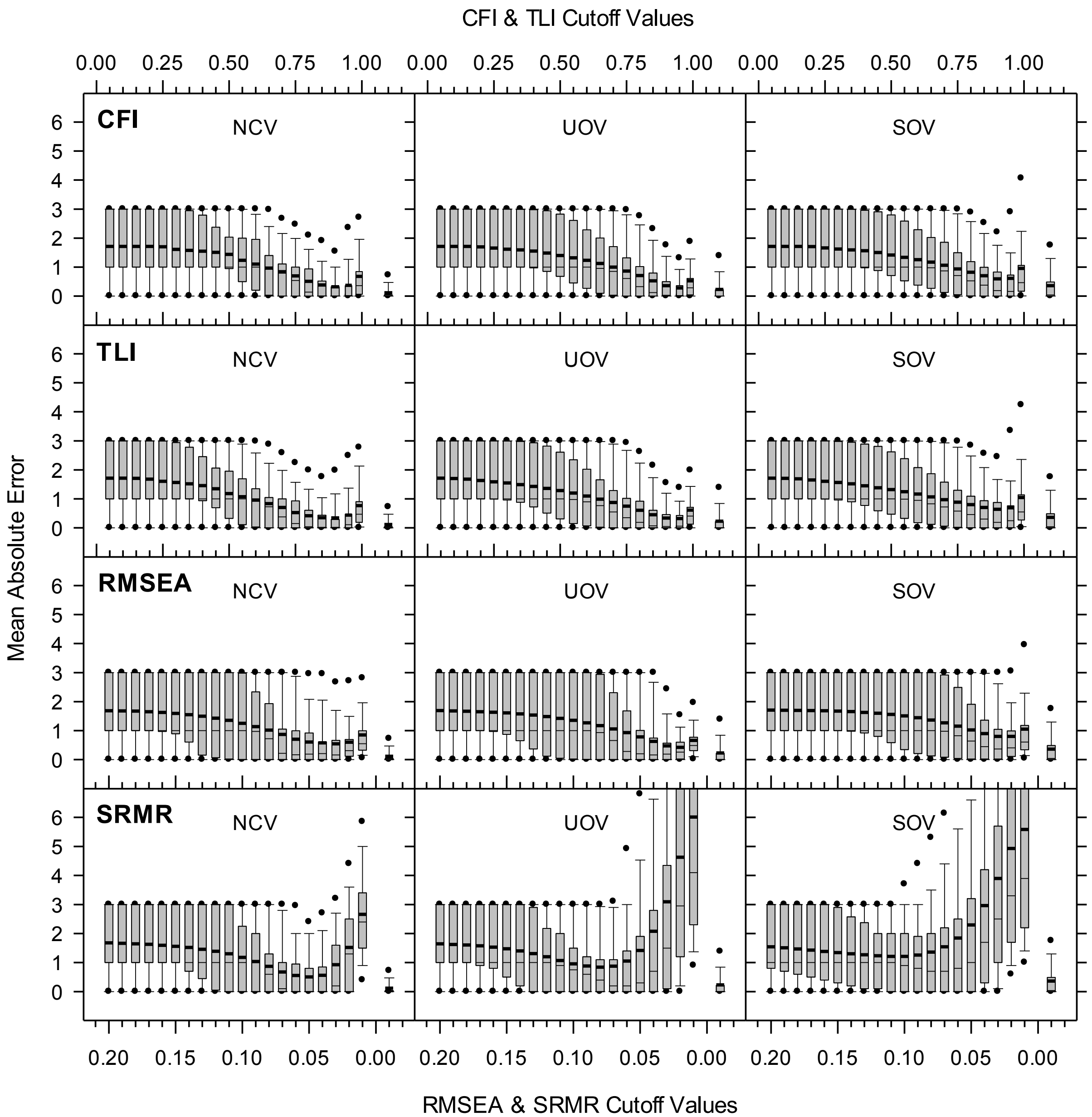
Note. NCV = normal continuous variables; UOV = unskewed ordered-categorical variables; SOV = skewed ordered-categorical variables; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual. The thick horizontal lines represent the mean proportion of correct estimates for each cutoff value, while the thin horizontal lines represent the median values. The top and bottom black circles indicate the 95^{th} and 5^{th} percentiles, respectively. The input values for the box plots are the mean proportion of correct estimates across 100 replications for each simulated condition. The rightmost box in each plot corresponds to the Parallel Analysis method. The last cutoff value plotted for the CFI and TLI indices is .99 (as opposed to 1.00).

Figure 2: *Box Plots for the Proportion of Correct Estimates Across Successive Cutoff Values*



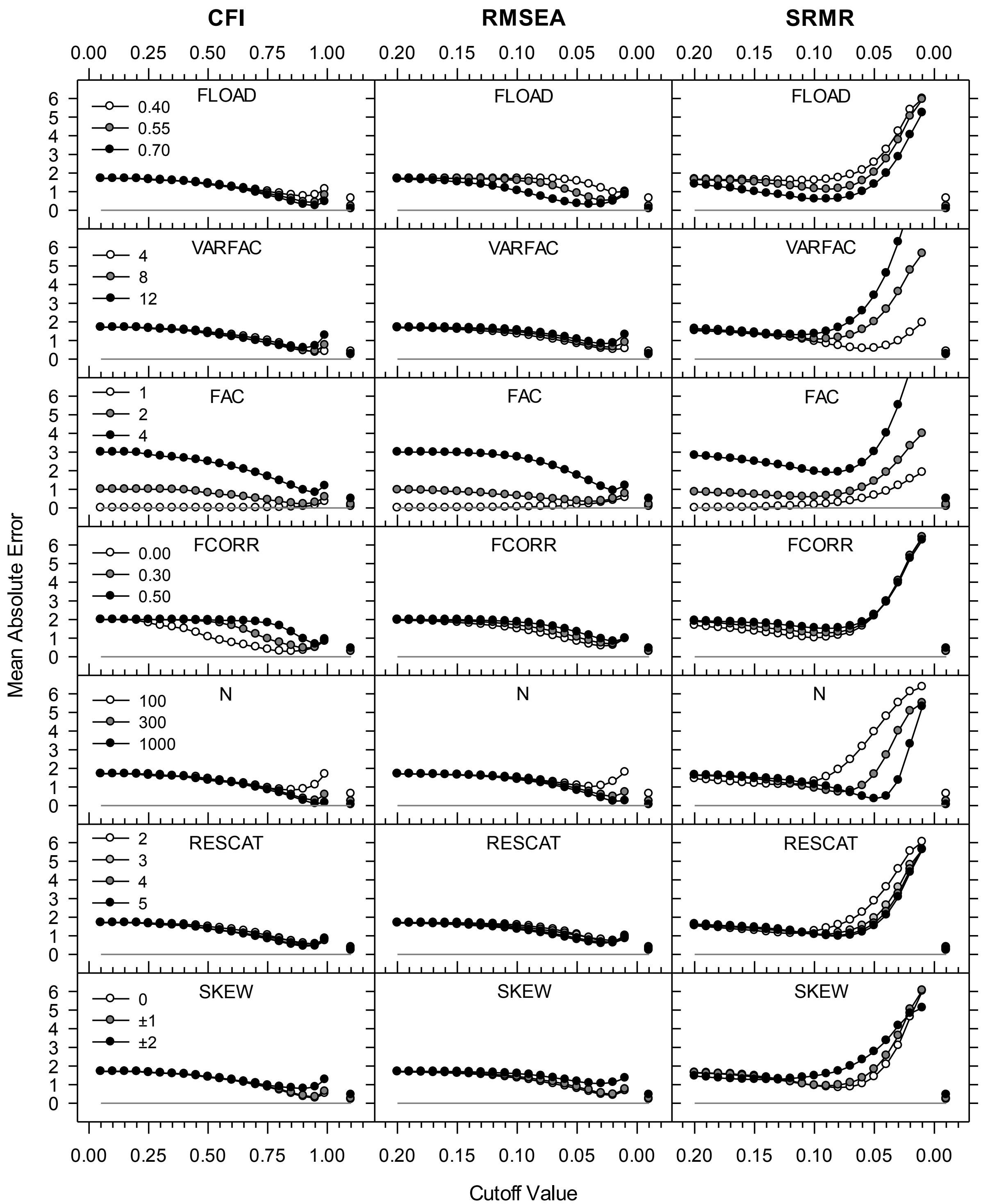
Note. NCV = normal continuous variables; UOV = unskewed ordered-categorical variables; SOV = skewed ordered-categorical variables; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual. The thick horizontal lines represent the mean bias error of estimations for each cutoff value, while the thin horizontal lines represent the median values. The top and bottom black circles indicate the 95th and 5th percentiles, respectively. The input values for the box plots are the mean bias error of estimation across 100 replications for each simulated condition. The rightmost box in each plot corresponds to the Parallel Analysis method. The last cutoff value plotted for the CFI and TLI indices is .99 (as opposed to 1.00). In order to facilitate the visual comparison of the methods, the range of the mean bias error was restricted between -4 and 6; this resulted in some truncated boxes for SRMR.

Figure 3: *Box Plots for the Mean Bias Error of Estimation Across Successive Cutoff Values*



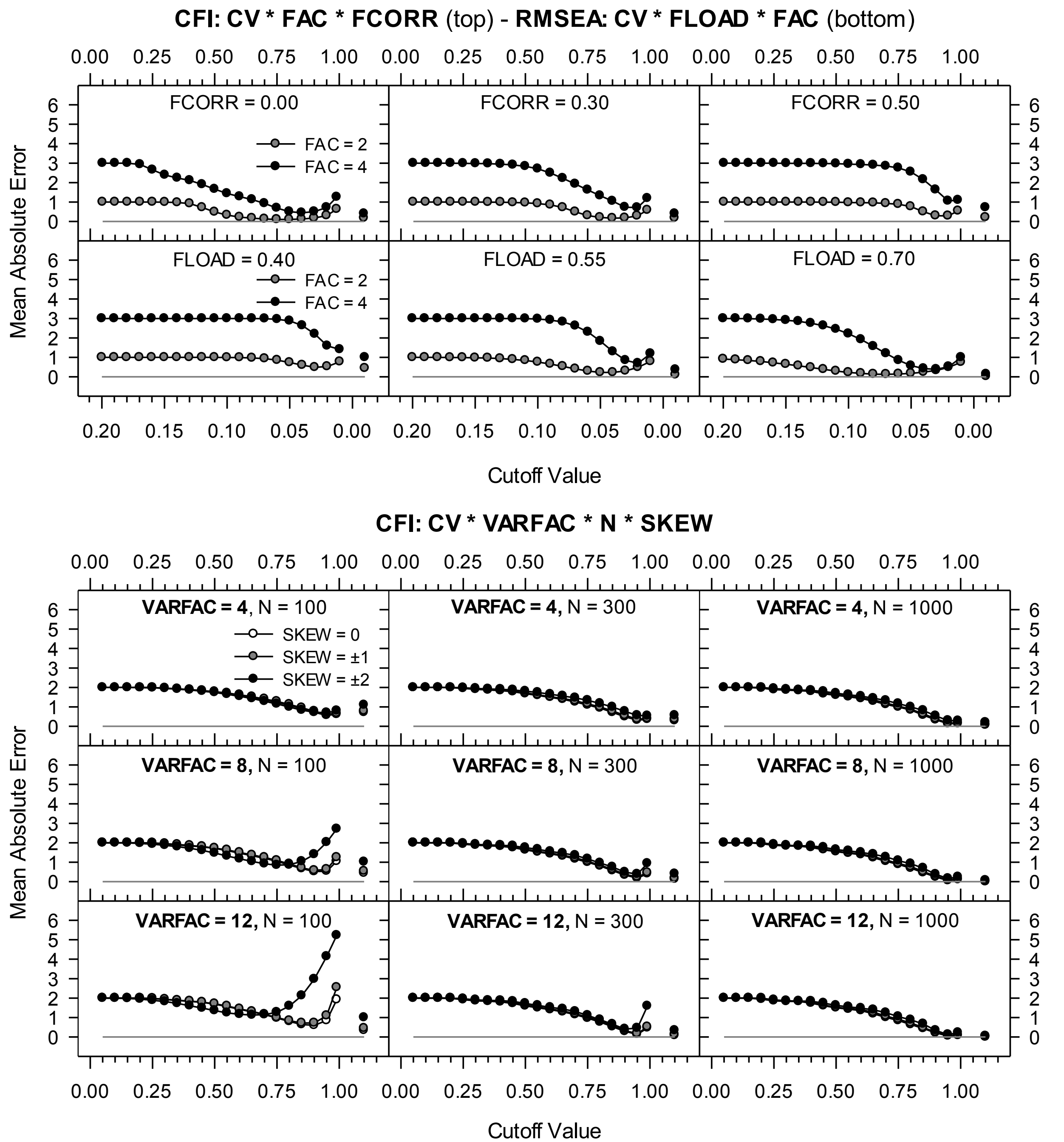
Note. NCV = normal continuous variables; UOV = unskewed ordered-categorical variables; SOV = skewed ordered-categorical variables; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual. The thick horizontal lines represent the mean absolute error of estimations for each cutoff value, while the thin horizontal lines represent the median values. The top and bottom black circles indicate the 95th and 5th percentiles, respectively. The input values for the box plots are the mean bias error of estimation across 100 replications for each simulated condition. The rightmost box in each plot corresponds to the Parallel Analysis method. The last cutoff value plotted for the CFI and TLI indices is .99 (as opposed to 1.00). In order to facilitate the visual comparison of the methods, the range of the mean absolute error was restricted between 0 and 6; this resulted in some truncated boxes for SRMR.

Figure 4: *Box Plots for the Mean Absolute Error of Estimation Across Successive Cutoff Values*



Note. FLOAD = Factor Loading; VARFAC = Variables per Factor; FAC = Number of Factors; FCORR = Factor Correlation; N = Sample Size; RESCAT = Response Categories; SKEW = Skewness. The 1-factor condition was not averaged across the levels of factor correlations. The rightmost circles in each plot correspond to the Parallel Analysis method. The last cutoff value plotted for the CFI index is .99 (as opposed to 1.00). The horizontal gray lines denote perfect accuracy. Some SRMR plots had to be truncated to facilitate the visual comparisons of the methods.

Figure 5: *Mean Absolute Error of Estimation Across the Levels of the Independent Variables*



Note. FLOAD = Factor Loading; VARFAC = Variables per Factor; FAC = Number of Factors; FCORR = Factor Correlation; N = Sample Size; SKEW = Skewness. The dependent variable was the mean absolute error of estimation. The 1-factor condition was not included in the ANOVAs. The rightmost circles in each plot correspond to the Parallel Analysis method. The last cutoff value plotted for the CFI index is .99 (as opposed to 1.00). The horizontal gray lines denote perfect accuracy.

Figure 6: *Mixed ANOVA Salient Higher-Order Interactions for the CFI and RMSEA Indices*