International Journal of Humanities and Social Science Review; Vol. 11 No. 1; March 2026
ISSN 2415-1157 (Online) ISSN 2415-1335 (Print)
Published by Research Institute for Progression of Knowledge

Beyond Conventional Thresholds: A Monte Carlo Reassessment of SEM Fit Indices Across Power Conditions

Stephen A. Sivo
Gordon E. Taub

University of Central Florida
United States of America

Abstract

This Monte Carlo study reexamines the performance of 13 commonly used fit indices in structural equation modeling (SEM) across a range of model conditions. Simulated data were generated from both confirmatory factor analysis (CFA) and latent curve autoregressive moving average (LCARMA) models under correct and misspecified conditions, with sample sizes of 125, 250, 500, 1000, 2000, and 4000, and varying degrees of model misfit. The results reveal that widely cited cutoff values—such as .95 for CFI and .08 for SRMR—often failed to distinguish correct from incorrect models, particularly in small samples or under mild misfit conditions. RMSEA demonstrated the most consistent performance, generally supporting a .05 threshold across scenarios. In contrast, incremental indices such as CFI and Gamma required values approaching .99 to achieve acceptable specificity. Parsimony indices (PGFI and PNFI) exhibited a systematic bias favoring complex models, even when model fit was held constant. The severity of misspecification and the complexity of the model structure both substantially influenced the effectiveness of fit indices. These findings challenge the application of universal cutoff values and underscore the need for context-sensitive interpretation of fit in SEM practice.

Keywords: Structural equation modeling, Fit indices, Monte Carlo simulation, Model misspecification, Cutoff criteria

Beyond Conventional Thresholds: A Monte Carlo Reassessment of SEM Fit Indices Across Power Conditions

A myriad of approximate fit indices have been developed for structural equation modeling over the last few decades to overcome a fundamental problem with using the original fit index, the chi-square test of exact fit, in practice: The models we specify are rarely (if ever) exactly true, even if accurate and useful. The consequence of using an exact fit test when practically no model ever exactly fits is that even accurate and useful models may be marked for elimination despite nothing more than trivial deviations from the data. Trivial deviations of a model from the data are magnified by sample size so any accurate, useful, yet imperfect model fitted to a large enough dataset can be pronounced poorly fitting with the chi-square test of exact fit. Approximate fit indices have been designed to overcome this limitation.

Evaluating how well a model fits sample data is required for the application of structural equation modeling (SEM). Although many approximate fit indices have been developed, the indices are by no means equivalent. Indeed, the plethora of available indices can leave an analyst in a quandary regarding which indices should be prioritized for interpretation. Clearly, indices can provide conflicting results.

In many ways, the interpretation of model fit has become somewhat of an art just as many SEM researchers acknowledge that interpretation will always involve some subjectivity (Cudeck & Henly, 1991; McDonald, 1989; McDonald & Marsh, 1990; Marsh, Hau, & Wen, 2004).

Indeed, personal preference regarding which fit results are consulted introduces more subjectivity into an analysis than one may recognize. Complicating matters further, several recent Monte Carlo studies find that cutoff criteria vary by sample size (Hu & Bentler, 1998;1999; Marsh, Hau, & Wen, 2004, Fan & Sivo, 2005; Sivo, Fan, Witta, & Willse, 2006), model type (Fan & Sivo, 2005), model size (Breivik & Olsson, 2001; Kenny & McCoach, 2003), and data distribution type (Hu & Bentler, 1998;1999), further consolidating that notion that universal cutoff values cannot be recommended for any one index.

A step towards reducing the degree of subjectivity in reporting fit indices in applied research might include (1) affirming that after repeated studies certain indices consistently emerge as preferable for interpretation and (2) identifying some of the conditions under which such indices optimally perform so that analysts can have a better sense of which index to consult in a given situation.

A summary of the findings of Monte Carlo studies evaluating fit index cutoff criteria

Hu and Bentler (1998; 1999), motivated by the reported arbitrariness of fit index cut off standards in SEM research (e.g., Browne & Cudeck, 1993; Macallum et al., 1996; McDonald & Marsh, 1990; Steiger, 1989), conducted simulation studies to understand the performance of fit index standards under simulated conditions. Their research was designed to wrestle empirically with the question of what fit index standard(s) *should* be used in evaluating CFA and SEM models. Their research plan essentially involved (a) generating data for two types of CFA models (namely, the Simple and Complex CFA model), (b) fitting nested, misspecified CFA models to their respective data (Simple and Complex data), and (c) examining the rejection rates across certain cut off criteria: .90 to .96 for fit indices scaled .00 to 1.00; .06 to .11 for the SRMR; and .05 to .08 for the RMSEA. The Simple and Complex Models for which the simulation data were generated were alike in that both were 15 variable, three factor CFA models; they were different in that factor to variable *cross loadings* (lambdas) were added in the latter CFA model type (thus assuming ambiguous variables). Hu and Bentler evaluated the utility of the *a priori* cutoffs by considering the degree to which rejection rates for the misspecified model variants generalized across various sample sizes, distributions, estimators, and, indeed, CFA model type (Simple versus Complex).

Any summary of Hu and Bentler's (1998; 1999) findings would include at least:

(1) Optimal cutoff criteria for various indices *within the limits of their study design* (a clearly expressed, but largely ignored, caveat) include: .90 for a ML based McDonald's (Mc) index; .95 for the CFI, RNI, Gamma, BL89, and TLI (a.k.a., NNFI); .08 for SRMR; .06 for RMSEA.
(2) A combination of two kinds of fit indices should be considered in any empirical investigation - particularly, SRMR and any one of the following indices: the RNI, CFI, Gamma, or BL89 (Hu & Bentler, 1999). SRMR, unlike all other indices evaluated previously, appeared to be much more sensitive to misspecified latent variable covariances, so it followed that SRMR should be coupled with any other index sensitive to latent to manifest relationships. Fan and Sivo (2005) later corrected this conclusion, discovering that the difference between SRMR and other indices was an artifact of Hu and Bentler's simulation design (1999).
(3) Combinational rules with the Mc, TLI, and RMSEA should be avoided when sample size is less than 250 (Hu & Bentler, 1999).
(4) Notably large differences exist between fit indices calculated using ML and GLS estimators, thus confirming the earlier results of Sugawara and McCallum (1993) (Hu & Bentler, 1998).

Hu and Bentler (1999), in summary, advise that a wider array of models in terms of model complexity and model specification conditions needs to be studied in future research. Yet, despite their call for a wider array of models, Hu and Bentler's (1998; 1999) conclusions have attained profound notoriety among applied SEM researchers. As of today, Hu and Bentler's 1999 article is the most cited article in the Structural Equation Modeling, a journal with a published impact factor higher than 3.000.

Several partial replications of Hu and Bentler's (1999) research have been published to evaluate their conclusions. With the influence of Hu and Bentler's (1999) conclusions in mind, Marsh, Hau, and Wen (2004) re-examined their work through a partial replication (focusing on the multivariate normal condition). Their investigation lead them to a number of different conclusions, perhaps the most significant of which was that "…the reason why Hu and Bentler's (1998, 1999) decision rules behave as they do is that they constructed misspecified models that should have been judged as acceptable—even according to their new, more stringent criteria" (p 328).

Marsh, Hau, and Wen (2004) illustrate this by presenting the probability density functions of the correctly specified Simple Model, and its two nested misspecified counterparts. They further presented the pdfs of the Complex Models (correct and incorrect). They conclude that the misspecified 15 variable, three factor CFA models in Hu and Bentler's (1998; 1999) studies, while misspecified, are nonetheless, acceptable. Indeed, they observed that "rejection of these acceptable-misspecified models should have constituted a Type 1 error (incorrect rejection of an "acceptable" model), leading to the seemingly paradoxical results whereby the probability of correctly rejecting misspecified models decreased substantially with increasing *N*." (p 320). This was as well a finding reported in Sivo, Fan, Willse, and Witta (2006). So, in the view of Marsh et al. (2004), Hu and Bentler's (1998; 1999) more stringent standards of fit are recommended because their design wrongly stages for comparison a correctly specified model and models that should be correct enough. In fact, Marsh, Hau, and Wen (2004) argue that fit index criteria should instead be relaxed rather than made more stringent. Their sentiment finds the company of many other SEM researchers who view models as useful simplifications. Indeed, the premise behind approximate fit indices is that our models are almost never exactly correct, so it stands to reason that a simulations study comparing correct and incorrect models in an effort to suggest fit criteria would miss an essential point when in fact the misspecified models qualify as useful approximations to the true underlying process. In a recent special issue of the journal *Personality and Individual Differences* focusing on the interpretation of fit indices, several notable contributors to the theme reinforced this idea that approximate fit indices find their purpose in that our models are never perfect (e.g., Millsap, 2007; Mulaik, 2007; Steiger, 2007). For a simulation study to consider recommending fit values, it stands to reason that researcher would contrive a study wherein the design compares an *approximately* correct model to other misspecified models.

Sivo, Fan, Witta, and Willse (2006) conducted another partial replication of Hu and Bentler's (1998;1999) research to examine whether their findings were robust to model type. In this study, data were generated for all six of Hu and Bentler's (1999) models (Hu and Bentler only generated data for one Simplex and one Complex model). Moreover, Simple and Complex models were allowed to fit one another's dataset as rival hypotheses. Although unintended, this strategy addressed a concern expressed in Marsh, Hau, and Wen (2004): That earlier simulation studies on fit indices had only tested restricted nested models against the correct model's data as rival hypotheses. Thirteen fit indices were compared. Key findings in Sivo, Fan, Witta, and Willse (2006) include:

(1) When correct models were fitted to their own data, True model type (3 Simple and 3 Complex models) did not affect the values of any fit index except SRMR, the parsimonious goodness-of-fit index (PGFI) and parsimonious normed fit index (PNFI). PGFI and PNFI fit values were the only indices greatly influenced by model type, accounting for 36% of the variation in PGFI values and 27% of the variation in PNFI values. True Model Type accounted for 4% of the differences in SRMR values.

(2) Certain indexes have the attribute of allowing correct models to be retained under very stringent cutoff values even when sample size is small (N = 150): CFI, non-normed fit index (NNFI), Mc, *RMSEA*, and Bollen's delta (Delta2).

(3) Cutoff values of the RMSEA and Mc actually *decreases* for misspecified models as sample size increases unlike eleven other fit indices considered including the CFI and SRMR- a desirable outcome. A review of discrepancies between the correct and misspecified model fit results suggests that the most observable distinctions occurred between these indices implying that these indices making it easier for analyst to compare rival models. This was not true for the goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), CFI, NNFI, normed fit index (NFI), Bollen's Rho (Rho1), and Delta2 as they changed very little or in some cases increased as sample size increased.

(4) The recommendation of .95 for any class of indexes may be inappropriate, ignoring the issue of sample size.

To further investigate the findings of Sivo, Fan, Witta, and Willse (2006) that fit indices are not impervious to model type, Fan and Sivo (2007) compared Hu and Bentler's CFA model to a four factor SEM while controlling for the severity of misspecification as quantified in terms of power. They reasoned that in order to examine whether fit values and corresponding cutoffs change as a function of model type, it is first necessary to equate different models to be compared in terms degree of misspecification. They reasoned that since Saris and Satorra (1993) and Satorra and Saris (1983) described statistical power for rejecting a misspecified model as a function of both the noncentrality parameter and its degrees of freedom, it follows that power calculations could be used to equate very dissimilar models in terms of misspecification severity.

In this way, dissimilar models might be contrasted in terms of fit index values and fit criteria. In addition, Fan and Sivo (2007), prompted by a reviewer, added a condition for each CFA and SEM model wherein the number of variables differed for each type. Small models and large models (operationalized in terms of number of variables) were thus compared. Fan and Sivo (2007) found that not only do fit index criteria vary by model type but also by model size (as demonstrated earlier in Breivik & Olsson, 2001; Kenny & McCoach, 2003). Fan and Sivo (2007) furthermore found that while Gamma, RMSEA, and Mc perform well under large model conditions; however, only Gamma performs well for small models.

Scientific Importance of the Study

The goal of this Monte Carlo investigation is to study the changing behavior of certain approximate fit indices and potential fit criteria under various conditions to understand better when they will optimally perform. The fit indices considered in this investigation are fit indices reported in most SEM software packages.

The specific aim of this investigation is to build upon prior Monte Carlo research by implementing a design that simultaneously takes into account salient factors never considered in a single study: Severity of misspecification quantified by power calculations (Fan & Sivo, 2007), model size (Breivik & Olsson, 2001; Kenny & McCoach, 2003), model specification complexity, sample size, and distribution type (Hu, Bentler, & Kano, 1992). By comparing the robustness of the fit indices across the conditions considered, we begin the task of identifying the relative strengths of indices that have previously prevailed as optimally performing.

Unlike Hu and Bentler (1998;1999) and Marsh, Hau and Wen (2004), this study considers an SEM model very different from the now very familiar 15 variable, three factor CFA used in their research: A multiple indicator version of the Latent Curve Autoregressive Moving Average model (LCARMA; Sivo, Fan, & Witta, 2005; Sivo & Fan, 2008). This model, while more constrained than the three factor CFA, incorporates an intricate system of structural paths. In the end, fit results for the LCARMA will be contrasted with Hu and Bentler's CFA model. Moreover, both the CFA and LCARMA models are specified under two size conditions for comparison: Small (6 variables) and large (15 variables). By doing this, our study takes a step towards meeting Hu and Bentler's (1999) call for a wider range of models in terms of model complexity and model specification conditions, a call yet to be fulfilled in a partial replication of their work other than Fan and Sivo (2007; 2009).

Another distinction of this study is that we do not choose, *a priori*, which cutoff values will be tested. Instead, we allow the simulation results to inform us about which cutoff values are optimal so that we may better understand the performance of fit index criteria. This was the procedure undertaken in Sivo, Fan, Witta, and Willse (2006). We underscore that our purpose is not to recommend criteria, but to understand better the conditions that affect such criteria. As such, noting the utility of cutoff values for the models under study is a convenient strategy for understanding the changing behavior of fit indices and potential criteria for evaluating them.

If our interest is to retain all correct models (i.e., no Type I error) while maximizing the chances of rejecting misspecified models (i.e., minimizing Type II error), then what cutoff values for these fives indexes are optimal for the conditions considered?

For this study, three primary research questions are asked:

Question 1: What are the highest possible fit index values (the lowest for SRMR, and RMSEA) not resulting in the rejection of any correctly specified models (i.e., 100% of the correct models are retained)? (The values identified are empirically derived cutoffs for the conditions at hand compared to Hu and Bentler's (1999) recommended cutoff criteria).

Question 2: What are the lowest possible fit index values (the highest for SRMR and RMSEA) resulting in the rejection of all misspecified (False) CFA and LCARMA models (i.e., 100% of the misspecified models are rejected) regardless of Model Size (Small or Large), and under various sample size and data distribution conditions? (The values identified are potential cutoffs for the conditions at compared to Hu and Bentler's (1999) recommended cutoff criteria).

Question 3: How do the empirically derived cutoffs for correctly specified and misspecified models compare? (Under what study conditions are 100% of the correct models retained and 100% of the misspecified models rejected?)

Method

Types of Models

This study will follow the procedures used in Sivo, Fan, Witta and Willse (2006). Fan and Sivo (2007). Twelve different models were used: six were Small CFA and LCARMA models (6 variables); six, Large CFA and LCARMA models (15 variables). The three small CFA models were subsumed under their larger counterparts. Both models were fitted to the same data in each iteration, with the exception that the small model included only 6 of the 15 available variables. Only one of the small CFA models and one of the large CFA models were so-called "True" models perfectly suiting the data generated. The remaining two small CFA models and large CFA models were designed to be nested within the "True" model counterparts. As restricted models they were misspecified by omitting 1 or 2 paths included in the True Models. This made it so that each model has three levels of specification error: true model (no specification error), and two levels of specification errors (Level 1 and Level 2), with Level 2 representing more severe misspecification than Level 1. At each level of misspecification, the severity of misspecification is comparable across the small and large CFA models. The severity of misspecification was *operationally* quantified by the estimated statistical power (Satorra & Saris, 1983; Saris & Satorra, 1993) for rejecting the misspecified model under a given sample size condition. In other words, the statistical power for rejecting the Level 1 (Power=.56) misspecified model is comparable across the small and large CFA models. Likewise, the statistical power for rejecting Level 2 (Power=.80) misspecified models was comparable across the small and large CFA models. This design makes it possible to investigate if the fit indices are sensitive to different types of models, while controlling for the severity of model misspecification.

Figure 1 presents the large confirmatory factor analysis model (CFA large), with the model parameters and the conditions of two levels of misspecification. This model is adapted from a model used in Hu and Bentler (1999), with some model parameters adjusted for the purpose of achieving the targeted specification error.

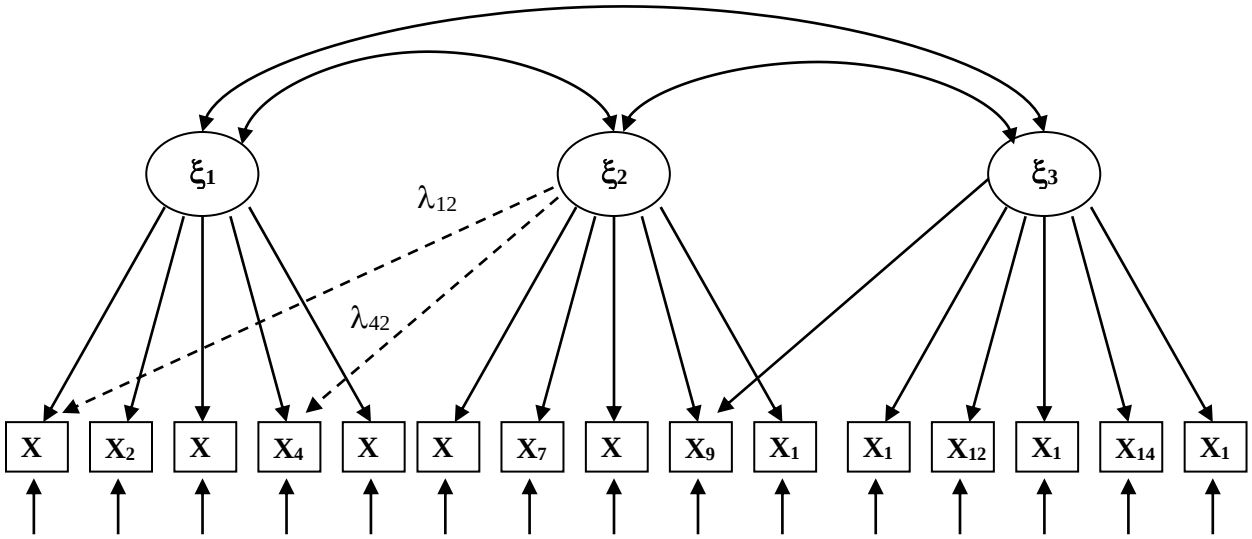
Figure 2 presents the small confirmatory factor analysis model (CFA small), with the model parameters and the conditions of two levels of misspecification.

The LCARMA models considered (both the small and large versions), like the CFA models, included a small and large correctly specified model contrasted with misspecified, nested LCARMA counterparts. The large model is found in Figure 3,

and the smaller version is presented in Figure 4. Not only does the LCARMA incorporate a more intricate system of paths, the lambda values were constrained to .60 so that the degrees of freedom for the model were much higher than the CFA model equivalents.

Model Misspecification Conditions

Using Fan and Sivo's (2005) strategy, the severity of misspecification was made comparable across the two models sizes (small and large). An operating assumption of this study is that if the power to reject misspecified models is comparable (either .56 or .80 at a sample size of 125), regardless of the shape or size of the models compared, the degree of misspecification is comparable. A Saris and Satorra based power program found in Sivo, Saunders, Chang, and Jiang (2006) was used to calculate analytical power.



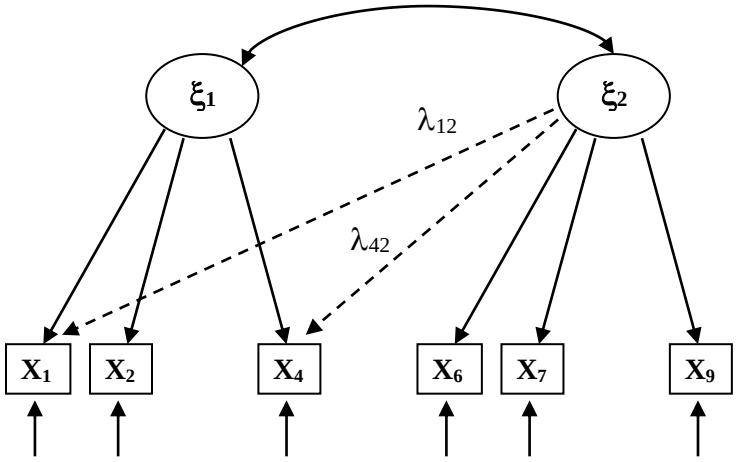
Level 1 Misspecified Model paths: λ_{42}
Level 2 Misspecified Model paths: λ_{42}; λ_{12}

$$\Phi = \begin{bmatrix} 1.00 & & \\ .50 & 1.00 & \\ .40 & .30 & 1.00 \end{bmatrix}$$

$$\Lambda_X\,(transposed) = \begin{bmatrix} .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ \underline{.70} & .00 & .00 & \underline{.70} & .00 & .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & \underline{.70} & .00 & .70 & .70 & .75 & .80 & .80 \end{bmatrix}$$

$$\theta_\delta\,(diagonal) = \begin{bmatrix} .51 & .51 & .4375 & .36 & .36 & .51 & .51 & .4375 & .36 & .36 & .51 & .51 & .4375 & .36 & .36 \end{bmatrix}$$

Figure 1: Large CFA with an indication of misspecified paths at Levels 1 and 2.



Level 1 Misspecified Model paths: λ_{42}
Level 2 Misspecified Model paths: λ_{42}; λ_{13}

$$\Phi = \begin{bmatrix} 1.00 & \\ .50 & 1.00 \end{bmatrix}$$

$$\Lambda_X(transposed) = \begin{bmatrix} .70 & .70 & .80 & .00 & .00 & .00 \\ \underline{.70} & .00 & \underline{.70} & .70 & .70 & .80 \end{bmatrix}$$

$$\theta_\delta(diagonal) = [.51 \quad .51 \quad .36 \quad .51 \quad .51 \quad .36]$$

Figure 2: Small CFA with an indication of misspecified paths at Levels 1 and 2 (nested within Large CFA).

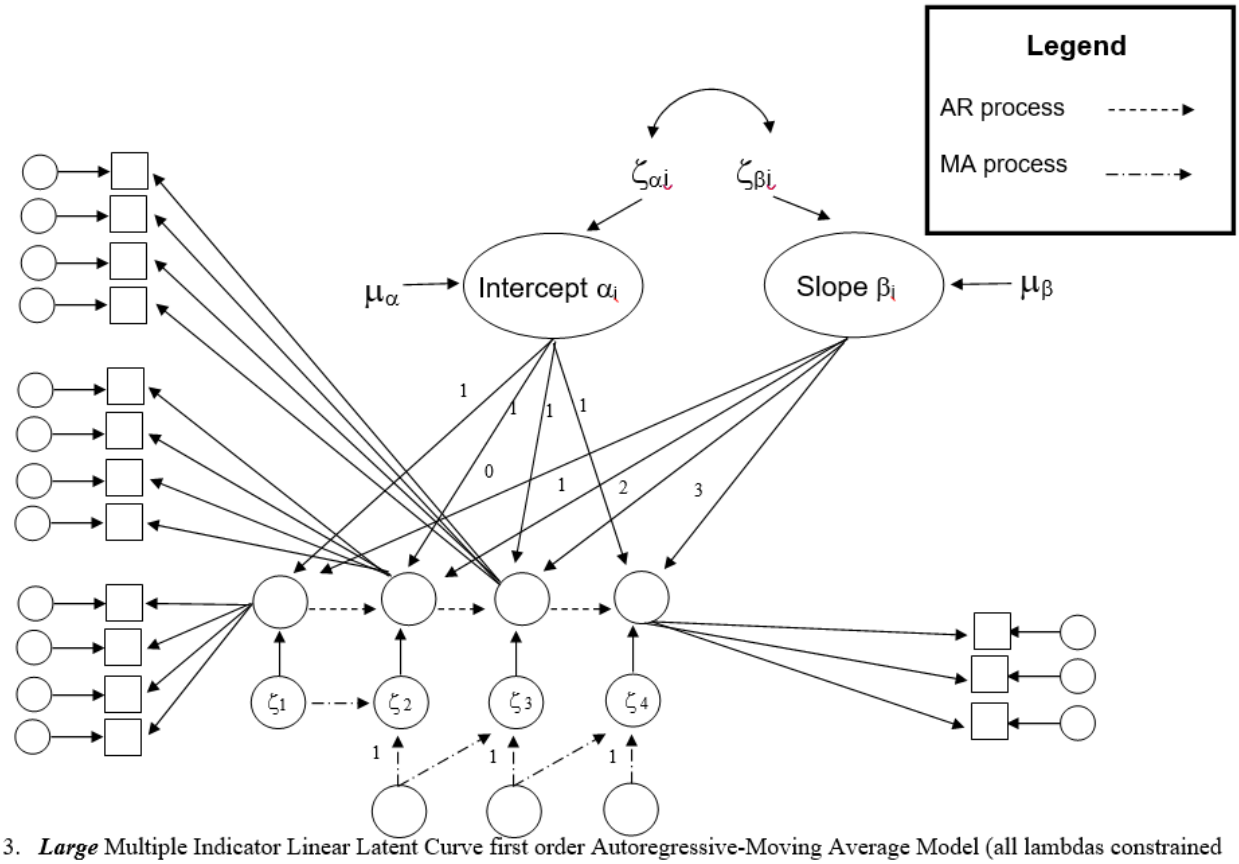


Figure 3. ***Large*** Multiple Indicator Linear Latent Curve first order Autoregressive-Moving Average Model (all lambdas constrained to .60).

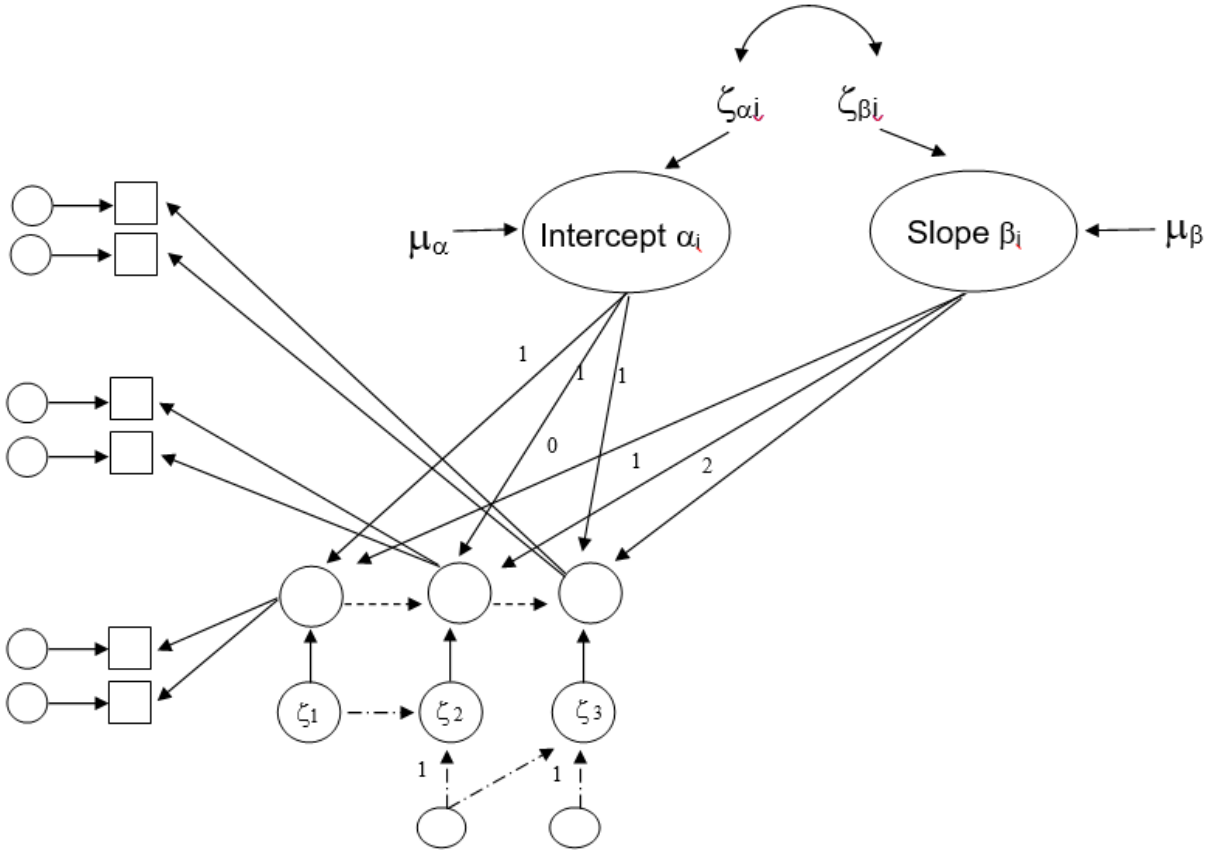


Figure 4. ***Small*** Multiple Indicator Linear Latent Curve first order Autoregressive-Moving Average Model (nested within Large LCARMA; (all lambdas constrained to .60).

Data Sources: Simulation Design and Data Conditions

In this study, maximum likelihood was the only estimation method used; however, Hu and Bentler's (1999) various data conditions were considered (multivariate normal and otherwise).

Each condition in this study was replicated 500 times and under 6 sample sizes (125, 250, 500, 1000, 2000, and 4000). Additionally, data were generated under seven conditions of data distribution discussed in Hu, Bentler, and Kano (1992) and in Sivo, Fan, Witta, and Willse (2006). Condition 1 consisted of multivariate normal latent scores and errors; Conditions 2 through 7 represented various departures from normality in terms of both/either the latent scores and/or the errors (see Hu, Bentler, & Kano, 1992). For a better understanding of the procedures used for data generation, see Fan, Felsovalyi, Sivo, and Keenan (2002).

Data generated for the Large three-factor CFA model were organized around the same parameters used in Hu and Bentler's (1999) so-called "Complex 1 model" with a minor adjustment to a crossloading specification to ensure the small CFA model is nested within the large CFA model. We use a similar CFA model intentionally to accomplish some level standardization or comparability across Monte Carlo study findings. The values for the Large three-factor CFA model data followed this factor-pattern matrix:

$$\begin{bmatrix} .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .70 & .00 & .00 & .70 & .00 & .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .70 & .00 & .70 & .70 & .75 & .80 & .80 \end{bmatrix}$$

with covariances among the factors as:

$$\begin{bmatrix} 1 & & \\ .50 & 1 & \\ .40 & .30 & 1 \end{bmatrix}$$

The values for the Small two-factor CFA model data followed this factor-pattern matrix:

$$\begin{bmatrix} .70 & .70 & .80 & .00 & .00 & .00 \\ .70 & .00 & .70 & .70 & .70 & .80 \end{bmatrix}$$

with a covariance among the factors as:

$$\begin{bmatrix} 1.00 & \\ .50 & 1.00 \end{bmatrix}$$

Variables analyzed with the Small CFA model were the same variables generated for the Large CFA condition as a control condition. This assured that differences in model size were uniquely attributable to model size and in no way a function of differences in the data.

Data generated for the Large Latent Curve Autoregressive Moving Average (LCARMA) Model possessed the following Lambda paths depicted in this factor-pattern matrix:

$$\begin{bmatrix} .60 & .60 & .60 & .60 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .60 & .60 & .60 & .60 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .60 & .60 & .60 & .60 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .60 & .60 & .60 \end{bmatrix};$$

and corresponding manifest errors of .40. Furthermore, the large LCARMA was generated to have first order latent means of

$$\begin{bmatrix} 0 & 1 & 2 & 3 \end{bmatrix};$$

a second order intercept mean and variance of .70 and .20, respectively; a slope mean and variance of .30 and .20, respectively; an Intercept-Slope covariance of .10, an AR coefficient of .40; and an MA coefficient of .30.

As with the CFA data, variables analyzed with the Small LCARMA model were the very same variables generated for the LCARMA CFA model. Again, this assured that differences in model size were uniquely attributable to model size and in no way a function of differences in the data generated. The values for the Small LCARMA model data followed this factor-pattern matrix:

$$\begin{bmatrix} .60 & .60 & .00 & .00 & .00 & .00 \\ .00 & .00 & .60 & .60 & .00 & .00 \\ .00 & .00 & .00 & .00 & .60 & .60 \end{bmatrix}$$

and corresponding manifest errors of .40. Furthermore, the Small LCARMA was generated to have first order latent means of

$$\begin{bmatrix} 0 & 1 & 2 \end{bmatrix};$$

a second order intercept mean and variance of .70 and .20, respectively; a slope mean and variance of .30 and .20, respectively; an Intercept-Slope covariance of .10, an AR coefficient of .40; and an MA coefficient of .30.

Lastly, data were generated for an Approximately True model. In this case, we returned to the Hu and Bentler Large 3 factor, CFA model with crossloadings and generated data wherein the first crossloading (with manifest variable 1) is specified as coming from Factor 2, when in reality the generating process relates Factor 3 to variable 1. The factor-pattern matrix is:

$$\begin{bmatrix} .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .70 & .00 & .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 \\ .70 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .70 & .00 & .70 & .70 & .75 & .80 & .80 \end{bmatrix}$$

Note that the misspecified models are the original models which makes the Approximately True model, a model that *is not nested* within the correct model for these new data.

Data Source and Analyses

Data were simulated using a combination of SAS macro, SAS BASE, SAS PROC IML (Interactive Matrix Language) following procedures detailed in Fan, Felsovalyi, Sivo, and Keenan (2002). Model fitting and estimation were implemented through SAS PROC CALIS. For each replication, the relevant fit indices were saved for later analyses.

Results and Discussion

The results of this study are organized around the three research questions. It is important to point out that the optimal result for any one index in this study is when, jointly, (a) 100% of all Correctly Specified (True) models are retained with a stringent cutoff criterion (very high values for most fit indices; very low values for SRMR and RMSEA); and (b) 100% of all Misspecified (False) models are rejected with a non-stringent cutoff criterion (lower values for most fit indices; higher values for SRMR and RMSEA). We want to be clear at the outset that we do not believe that retaining (or rejecting) 95% of the models instead of 100% is conducive to our purpose of evaluating the behavior of fit indices via their cutoffs.

Question 1: What are the highest possible fit index values (the lowest for SRMR, and RMSEA) not resulting in the rejection of any correctly specified models (i.e., 100% of the correct models are retained)?

To evaluate variability in fit criteria by True model condition, the highest GFI, AGFI, NNFI(TLI), NFI, Bollen rho, Bollen Delta, CFI, Mc, or Gamma cutoff point at which 100% of the True models are retained is tabulated in Table 4 for three factor CFA models and in Table 7 for LCARMA models (the lowest point for SRMR and RMSEA). Table 1 values suggest that at a sample size of 125, CFI, Mc, GFI, AGFI, NNFI(TLI), NFI, Bollen rho, Bollen Delta, Gamma, and SRMR values for the Small CFA model are right at the cut off points set by Hu and Bentler for each index. Table 2 presents results for the Large CFA model

Cutoff values for the Small True CFA model condition were directly comparable to those for the Small True LCARMA for all indices but SRMR (see Table 3). Moreover, the GFI, AGFI, NNFI(TLI), NFI, Bollen rho, Bollen Delta, CFI, Mc, and Gamma values in each

Small Model condition exceeded Hu and Bentler's cutoffs at n=125 and only improved as sample size approached 4000. High GFI, AGFI, NNFI(TLI), NFI, Bollen rho, Bollen Delta, CFI, Mc, and Gamma values for Correctly Specified Models are desirable because the opportunity of simultaneously rejecting misspecified models is more likely. SRMR performed well compared to Hu and Bentler's cutoff of .08 with the Small True CFA Model; however, for the Large True CFA, sample size needed to reach 250 before reaching the prescribed standard. The SRMR cutoffs for the True LCARMA models were never satisfactory for the Large model, but reached an acceptable level of the prescribed .08 at sample size of 500 for the Small model. Please note that the very same variables used in the Large LCARMA condition were used in the Small LCARMA condition. This suggests that SRMR has more difficulty with Large models more intricate than the now familiar three factor CFA.

The RMSEA cutoff values under the Small True CFA Model condition do not reach the Hu and Bentler cutoff until n=4000 and is more than twice that value at n=125 (RMSEA cutoff=.15). By contrast, RMSEA cutoff values for the Large True CFA condition are less than .08 at n=125 and only get lower as n approaches 4000. Similar results are found for the True LCARMA models. RMSEA cutoff values for the Large True LCARMA model at n=250 reach .06 while Small model cutoffs reach .06 at n=500.

Table 1

The Optimal [1,2] Index Value for Not Rejecting **Small True Models** Fitted to **CFA** Data (decimals omitted).

Small True CFA (6 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	94	96	98	99	99	99	
AGFI	80	88	92	94	96	97	
PGFI	<.40	<.40	<.40	<.40	<.40	<.40	
CFI	95	97	98	98	99	99	.95
NNFI/TLI	88	92	95	96	97	98	
NFI	93	96	97	98	99	99	
PNFI	<.40	<.40	<.40	<.40	<.40	<.40	
Mc	93	96	97	98	98	98	.90
Gamma	95	97	98	98	99	99	.95
Bollen □	84	90	94	96	97	98	
Bollen □	95	97	98	98	99	99	
SRMR	08	07	05	04	04	03	.08
RMSEA	15	12	10	08	07	06	.06

1 Higher cutoff values for True Models are more desirable for the fit indices (lower, for SRMR and RMSEA) because this implies 100% of the True models attained fit values at least as high as the tabled value. With higher cutoff values, the potential for rejecting false models while retaining True models is increased.

2 High values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 2

The Optimal [1,2] Index Value for Not Rejecting **Large True Models** Fitted to **CFA** Data (decimals omitted).

Large True CFA (15 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	88	93	96	98	99	99	
AGFI	83	91	95	97	98	99	
PGFI	70	74	77	78	79	79	
CFI	95	98	99	99	99	99	.95
NNFI/TLI	94	97	98	99	99	99	
NFI	88	94	97	98	99	99	
PNFI	70	75	77	78	79	79	
Mc	83	91	96	98	99	99	.90
Gamma	95	97	99	99	99	99	.95
Bollen □	91	92	96	98	99	99	
Bollen □	95	98	99	99	99	99	
SRMR	10	07	05	04	03	02	.08
RMSEA	07	05	03	03	02	02	.06

1 Higher cutoff values for True Models are more desirable for the fit indices (lower, for SRMR and RMSEA) because this implies 100% of the True models attained fit values at least as high as the tabled value. With higher cutoff values, the potential for rejecting false models while retaining True models is increased.

2 High values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 3

The Optimal [1,2] Index Value for Not Rejecting **Small True Models** Fitted to **LCARMA** Data (decimals omitted).

Small True LCARMA (6 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	94	96	98	99	99	99	
AGFI	85	92	96	98	98	99	
PGFI	49	50	51	51	52	52	
CFI	97	98	99	99	99	99	.95
NNFI/TLI	94	97	98	99	99	99	
NFI	95	97	98	99	99	99	
PNFI	49	51	51	52	52	52	
Mc	93	96	98	99	99	99	.90
Gamma	95	97	98	99	99	99	.95
Bollen □	91	95	97	98	99	99	
Bollen □	97	98	99	99	99	99	
SRMR	14	09	06	05	04	03	.08
RMSEA	12	09	06	04	03	02	.06

1 Higher cutoff values for True Models are more desirable for the fit indices (lower, for SRMR and RMSEA) because this implies 100% of the True models attained fit values at least as high as the tabled value. With higher cutoff values, the potential for rejecting false models while retaining True models is increased.

2 High values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Under the Large True LCARMA condition GFI, AGFI, NNFI(TLI), NFI, Bollen rho, Bollen Delta, CFI and Gamma values are about as acceptable as they are under all other model conditions (see Table 4). It can be said that, as far as Correctly Specified Models go (Small or Large, CFA or LCARMA), CFI and Gamma cutoffs set to retain 100% of the Correct Models are very close to Hu and Bentler recommended cutoffs. Mc cutoff values, on the other hand, for both CFA and LCARMA models at n=125 only meet Hu and Bentler's recommended values when Small True Models are examined. For the Large True CFA, sample size must reach 250 before meeting the .90 standard, and for the Large True LCARMA, a sample size of even 4000 was greeted by Mc cutoff values lower than .90 (at n=4000, the Mc cutoff is .87). Mc never reaches a high enough value for the Large True LCARMA model, despite the fact the Smaller nested version fitted to the same data (fewer variables of course) with a .90 at n=125.

A review of the parsimony indices (PGFI and PNFI) suggests that they disfavor greatly small true models over large true models. Moreover, the LCARMA models were more kindly treated by the CFA models simply because of the number of constraints imposed in LCARMA models. The reward for having imposed equality constraints lead the parsimony indices to suggest better fit results for true LCARMA models than true CFA models. This reward is unreasonably given that model types were true for the data at hand.

Question 2: What are the lowest possible fit index values (the highest for SRMR and RMSEA) resulting in the rejection of all misspecified (False) CFA and LCARMA models (i.e., 100% of the misspecified models are rejected) regardless of Model Size (Small or Large), and under various sample size and data distribution conditions?

Misspecified models (Small and Large, CFA and LCARMA) were set to have two levels of misspecification in terms of Saris-Satorra power calculations. In this way, severity of misspecification could be considered. Level 2 Misspecification was set to a power of .80; level 1 Misspecification was set to .56. Tables 5, 6, 7 and 8 present the fit index cutoff results for CFA and LCARMA conditions at both levels of misspecification. Each Table presents results for both the small and large models. Table 5 and 6 pertain to the Misspecified CFA models; Table 7 and 8, the Misspecified LCARMA models.

According to the results summarized in Tables 5, 6, 7 and 8, CFI cutoffs are never able to exclude 100% of all Misspecified models at the Hu and Bentler criterion of .95. That is, CFI cutoffs aiming to reject 100% of Misspecified models could not exclude models unless they were set to .96 or higher – indeed, often at .99 regardless of whether the model results were Small or Large, CFA or LCARMA or set to a Misspecified level of 1 or 2. Setting aside Hu and Bentler criterion, an independent evaluation of these cutoffs leads to the conclusion that an empirically derived CFI cutoff would have to be too high, approaching the limit of the scale for the CFI index. The same finding occurred for Gamma cutoff values when large or small Misspecified CFA models are fitted to data regardless of the degree of misspecification considered in this study. The situation for Gamma cutoffs was especially difficult for Small Misspecified CFA models according to Tables 5 and 6 because required cutoff values had to be set at .99 even at the highest sample size considered in this study (n=4000). Tables

Table 4

The Optimal [1,2] Index Value for Not Rejecting **Large True Models** Fitted to **LCARMA** Data (decimals omitted).

Large True LCARMA (15 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	85	91	93	95	96	96	
AGFI	82	88	92	94	95	96	
PGFI	77	82	85	86	87	87	
CFI	97	98	98	98	98	98	.95
NNFI/TLI	97	98	98	98	98	98	
NFI	91	95	97	98	98	98	
PNFI	82	86	88	89	89	89	
Mc	76	82	85	86	86	87	.90
Gamma	93	95	95	96	96	96	.95
Bollen □	90	94	97	98	98	98	
Bollen □	97	98	98	98	98	98	
SRMR	15	12	12	11	11	11	.08
RMSEA	08	06	06	06	06	05	.06

1 Higher cutoff values for True Models are more desirable for the fit indices (lower, for SRMR and RMSEA) because this implies 100% of the True models attained fit values at least as high as the tabled value. With higher cutoff values, the potential for rejecting false models while retaining True models is increased.

2 High values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 5

The Optimal [1,2] Index Value for Rejecting Level 2 Misspecified Models (**Power=.80**) Fitted to **Small CFA** Data respectively (decimals omitted).

Small CFA (6 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	99	99	99	99	99	99	
AGFI	96	96	96	96	96	96	
PGFI	59	59	59	59	59	59	
CFI	99	99	99	99	98	98	.95
NNFI/TLI	99	99	98	97	97	97	
NFI	98	98	98	98	98	98	
PNFI	59	59	59	59	59	59	
Mc	99	98	98	98	98	98	.90
Gamma	99	99	99	99	99	99	.95
Bollen □	97	97	97	97	97	97	
Bollen □	99	99	99	99	98	98	
SRMR	04	04	05	05	05	05	.08
RMSEA	01	05	06	07	08	08	.06

1 Lower cutoff values for False Models are more desirable for the fit indices (higher, for SRMR and RMSEA) because this implies 100% of the False models were rejected at the tabled value. With lower cutoff values, the potential for not rejecting True models is increased while rejecting False models.

2 Low values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 6

The Optimal [1,2] Index Value for Rejecting Level 2 Misspecified Models (**Power=.80**) Fitted to **Large CFA** Data respectively (decimals omitted).

Large CFA (15 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	91	93	95	95	96	96	
AGFI	87	91	93	94	94	94	
PGFI	75	77	79	79	79	79	
CFI	99	98	97	97	97	97	.95
NNFI/TLI	99	97	97	96	96	96	
NFI	92	94	95	96	96	97	
PNFI	76	78	79	80	80	80	
Mc	94	90	88	87	86	86	.90
Gamma	99	98	97	97	97	96	.95
Bollen □	90	93	94	95	95	96	
Bollen □	99	98	97	97	97	97	
SRMR	08	08	07	07	07	07	.08
RMSEA	03	04	05	05	05	06	.06

1 Lower cutoff values for False Models are more desirable for the fit indices (higher, for SRMR and RMSEA) because this implies 100% of the False models were rejected at the tabled value. With lower cutoff values, the potential for not rejecting True models is increased while rejecting False models.

2 Low values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 7

The Optimal [1,2] Index Value for Rejecting Level 1 Misspecified Models (**Power=.56**) Fitted to **Small CFA** Data respectively (decimals omitted).

Small CFA (6 variables)	**Sample Size**						
Fit Index	125	250	500	1000	2000	4000	*Hu and Bentler cutoffs*
GFI	99	99	99	99	99	99	
AGFI	97	97	97	96	96	96	
PGFI	59	59	59	59	59	59	
CFI	99	99	99	99	99	99	.95
NNFI/TLI	99	99	99	98	98	97	
NFI	99	99	99	99	99	99	
PNFI	59	59	59	59	59	59	
Mc	99	99	99	98	98	98	.90
Gamma	99	99	99	99	99	99	.95
Bollen □	98	98	98	97	97	97	
Bollen □	99	99	99	99	99	99	
SRMR	04	04	04	04	04	04	.08
RMSEA	01	03	05	07	07	07	.06

1 Lower cutoff values for False Models are more desirable for the fit indices (higher, for SRMR and RMSEA) because this implies 100% of the False models were rejected at the tabled value. With lower cutoff values, the potential for not rejecting True models is increased while rejecting False models.

2 Low values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 8

The Optimal [1,2] Index Value for Rejecting Level 1 Misspecified Models (**Power=.56**) Fitted to **Large CFA** Data respectively (decimals omitted).

Large CFA (15 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	92	94	95	95	96	96	
AGFI	88	92	93	94	94	94	
PGFI	77	79	80	80	80	80	
CFI	99	99	98	97	97	97	.95
NNFI/TLI	99	98	97	97	96	96	
NFI	93	95	96	96	96	96	
PNFI	78	80	80	81	81	81	
Mc	98	93	90	88	86	86	.90
Gamma	99	98	97	97	97	96	.95
Bollen □	91	94	95	95	96	96	
Bollen □	99	99	98	97	97	97	
SRMR	07	06	05	05	05	05	.08
RMSEA	02	04	05	05	05	06	.06

1 Lower cutoff values for False Models are more desirable for the fit indices (higher, for SRMR and RMSEA) because this implies 100% of the False models were rejected at the tabled value. With lower cutoff values, the potential for not rejecting True models is increased while rejecting False models.

2 Low values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

7 and 8, which summarized LCARMA results, indicate that Gamma cutoffs rejecting all misspecified models were possible regardless of which combination of LCARMA model size and misspecification level was considered. In fact, compared to Hu and Bentler's standard of .95, the required Gamma cutoffs are, in most cases, lower than the criterion suggesting that a lower cutoff is possible with the aim of rejecting 100% of all misspecified models. In fact, when sample size was set to 500 and Level 2 misspecified models were considered (power set to .80), Gamma cutoff could be set to .87 or .88 and still reject 100% of the misspecified LCARMA models. Independent of Hu and Bentler's standard of .95, optimal cutoff values of .87 or .88 suggest that the conventional .90 criterion of the past may be adequate under some conditions for ruling out misspecified models. Similar findings occurred for the GFI, AGFI, NNFI(TLI), NFI, Bollen rho, Bollen Delta indices as well.

Likewise, cutoff results for the SRMR are mixed for of Misspecified models. SRMR cutoffs calculated for Misspecified CFA models, compared to Hu and Bentler's standard are not satisfied when the Misspecified CFA Models were Small (across Misspecification severity) nor when Misspecification was set to level 1 (power=.56). In other words SRMR Hu and Bentler's cutoff of .08 only functioned well for the Large Misspecified CFA model condition at Misspecification level 2 (power=.80). Indeed, this result generalizes across sample sizes (see Table 3).

Incidentally, this is the condition that Hu and Bentler consulted when recommending a cutoff of .08 in the first place. These results affirm that Hu and Bentler's (1999) recommended cutoff for SRMR is idiosyncratic to the conditions of their study. Small CFA models require a tighter criterion. Hu and Bentler's (1999) cutoff does prove to be useful for the Small Misspecified LCARMA condition at both levels of Misspecification (Tables 7 and 8). In fact, when Large Misspecified LCARMA conditions are considered the empirically derived cutoff could be set a little higher.

Results for the RMSEA cutoff values were acceptable or better for rejecting any Misspecified LCARMA model under any condition using Hu and Bentler's recommended criterion of .06. The same (criterion=.06) was true for Misspecified CFA models when sample size was larger (N=500 for the Small CFA at power=.80; N=1000 for the Small CFA at power=.56; N=4000 for Large CFA models). Notably, RMSEA ranged from .03 to .05, by sample size for Large Misspecified CFA models up to N=2000. An empirically derived criterion of .05 for RMSEA permitted 100% of all Misspecified CFA models (and, of course, Misspecified LCARMA models) to be rejected across conditions when sample size was at least 500 (perhaps supporting the utility of the conventional standard; again, we are simply comparing our findings with extant standards, not recommending universal cutoffs). This result suggests that, regardless of the size of the model (CFA or LCARMA) and regardless of the level of misspecification, RMSEA cutoff values of .05 or .06 only failed to reject misspecified models when the sample size was set to 250 (our design condition). It may be observed that RMSEA is more "lenient" with larger models (15 variable rather than 6 variable models) but not to the point that the conventional criterion of .05 is compromised – at least for our model conditions. This finding is in stark contrast to the conclusion of Kenny and McCoach (2003) who report that RMSEA is upwardly biased for larger models, this increasing the chances of retaining misspecified models as p (number of variables) increases. Yet, we do not find, in consulting our tables that this is problematic for any model under any condition with respect to the conventional .05 criterion so long as sample size is set to .05 (see Tables 5, 6, 7 and 8).

Empirically derived cutoff values for Mc index permit 100% of Large Misspecified models to be rejected at .90 (Hu and Bentler's recommendation) so long as sample size is set to 500 (see Tables 5, 6, 8 and 9). In fact, Mc cutoff values never exceed .80 for Large Misspecified LCARMA models regardless of the degree of misspecification. Small Misspecified LCARMA models set to Misspecification level 2 (power=.80) were able to be rejected at .90 when sample size was 250, though at Misspecification level 1 (power=.56), this result was only possible at N=4000 (Degree of misspecification in this study has a strong effect on the Mc cutoff value for Small LCARMA models). Small Misspecified CFA models, conversely, require empirically derived cutoffs that are very restrictive (set at .98 or .99) in order to reject all such models.

Question 3: How do the empirically derived cutoffs for correctly specified and misspecified models compare? (Under what study conditions are 100% of the correct models retained and 100% of the misspecified models rejected?)

To make sense of what the conclusions in Questions 2 and 3 imply in terms of empirically derived cutoffs for the condition considered in this study, a joint consideration of related findings is here presented. Comparing Table 1 and 2 with Tables 5-8, and Tables 3 and 4 with Tables 9-12, we located the minimum cut-off value at which 100% of all Correctly Specified models were retained and 100% of the Misspecified Models were rejected and noted the corresponding sample size needed to accomplish this optimal outcome. Again, this is to understand how optimal fit index criteria vary by condition rather than to recommend specific cutoff values. We aim to identify the conditions under which empirically derived fit criteria function best to both retain Correctly Specified models and reject Misspecified models. Figures 5 through 8 present the condition distributions at the lowest and highest sample sizes (125 and 4000, respectively) for five of the fit indices (arguably the most frequently fit indices reported in the literature).

Table 13 summarizes the optimal points for five of the indices across the conditions of this study, but similar results can be constructed for any of the remaining 8 indices by comparing tables 1-12. The optimal cutoff value varies by Model Type, by Model Size, and by Sample Size regardless of the level of Misspecification. It can be seen that to use a Jointly optimal cutoff value for small CFAs, sample size needs to be rather large for all indices but the SRMR. An optimal cutoff for SRMR only required a sample size of 500 and a value of .04 or .05 in order to pronounce the status of a model, *assuming that we would only accept a perfectly correct model*.

Of course, this study is not recommending fit index cutoff values, but instead intends to study their behavior. It is reasonable to think, however, that an approximately correct Small CFA would require a larger sample size for fit indices to perform optimally unless one were to consult the SRMR. What the value would have to be for an SRMR under approximately true models is an open question which may, in the end, not be answerable. LCARMA models (especially with the number of constraints imposed in this study) do not require sample sizes higher than 500 for discrimination between Correctly Specified (perfectly) and Misspecified models. In fact, a Mc cutoff of .80 was capable of being an optimal choice. CFI cutoffs, although capable of distinguishing between Correct and Misspecified

Table 9

The Optimal [1,2] Index Value for Rejecting Level 2 Misspecified Models (**Power=.80**) Fitted to **Small LCARMA** Data respectively (decimals omitted).

Small LCARMA (6 variables)	**Sample Size**						
Fit Index[3]	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	92	92	92	92	92	92	
AGFI	88	88	88	88	88	88	
PGFI	79	79	79	79	79	79	
CFI	98	97	96	96	96	96	.95
NNFI/TLI	97	97	96	96	95	95	
NFI	96	96	96	96	96	96	
PNFI	82	82	82	82	82	82	
Mc	92	90	88	87	87	87	.90
Gamma	95	93	92	92	92	91	.95
Bollen □	95	95	95	95	95	95	
Bollen □	98	97	96	96	96	96	
SRMR	08	08	08	08	08	08	.08
RMSEA	09	11	12	12	12	12	.06

1 Lower cutoff values for False Models are more desirable for the fit indices (higher, for SRMR and RMSEA) because this implies 100% of the False models were rejected at the tabled value. With lower cutoff values, the potential for not rejecting True models is increased while rejecting False models.

2 Low values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 10

The Optimal [1,2] Index Value for Rejecting Level 2 Misspecified Models (**Power=.80**) Fitted to **Large LCARMA** Data respectively (decimals omitted).

Large LCARMA (15 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	81	84	85	86	86	86	
AGFI	79	82	84	85	85	85	
PGFI	84	87	89	90	90	90	
CFI	97	96	96	96	96	96	.95
NNFI/TLI	97	96	96	96	96	96	
NFI	93	94	95	95	95	95	
PNFI	97	98	99	99	99	99	
Mc	65	61	60	59	58	57	.90
Gamma	90	89	88	88	88	87	.95
Bollen □	93	94	95	95	96	96	
Bollen □	97	96	96	96	96	96	
SRMR	10	10	10	10	10	10	.08
RMSEA	08	08	09	09	09	09	.06

1 Lower cutoff values for False Models are more desirable for the fit indices (higher, for SRMR and RMSEA) because this implies 100% of the False models were rejected at the tabled value. With lower cutoff values, the potential for not rejecting True models is increased while rejecting False models.

2 Low values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 11

The Optimal [1,2] Index Value for Rejecting Level 1 Misspecified Models (**Power=.56**) Fitted to **Small LCARMA** Data respectively (decimals omitted).

Small LCARMA (6 variables)	**Sample Size**						
Fit Index	125	250	500	1000	2000	4000	*Hu and Bentler cutoffs*
GFI	94	94	94	94	94	94	
AGFI	91	91	91	91	91	91	
PGFI	81	81	81	81	81	81	
CFI	99	98	98	97	97	97	.95
NNFI/TLI	99	98	97	97	97	97	
NFI	97	97	97	97	97	97	
PNFI	83	83	83	83	83	83	
Mc	95	93	92	91	91	90	.90
Gamma	97	95	95	94	94	94	.95
Bollen □	96	96	96	96	96	96	
Bollen □	98	98	98	97	97	97	
SRMR	07	07	07	07	07	07	.08
RMSEA	07	09	09	10	10	11	.06

1 Lower cutoff values for False Models are more desirable for the fit indices (higher, for SRMR and RMSEA) because this implies 100% of the False models were rejected at the tabled value. With lower cutoff values, the potential for not rejecting True models is increased while rejecting False models.

2 Low values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 12

The Optimal [1,2] Index Value for Rejecting Level 1 Misspecified Models (**Power=.56**) Fitted to **Large LCARMA** Data respectively (decimals omitted).

Large LCARMA (15 variables)	**Sample Size**						
Fit Index	*125*	*250*	*500*	*1000*	*2000*	*4000*	*Hu and Bentler cutoffs*
GFI	83	87	89	89	90	90	
AGFI	82	86	88	89	89	89	
PGFI	87	90	92	93	93	93	
CFI	98	97	97	97	97	97	.95
NNFI/TLI	97	97	97	97	97	97	
NFI	94	95	96	96	97	97	
PNFI	98	99	99	99	99	99	
Mc	73	70	69	67	66	66	.90
Gamma	93	92	91	91	90	90	.95
Bollen □	94	95	96	97	97	97	
Bollen □	97	97	97	97	97	97	
SRMR	09	08	08	09	09	09	.08
RMSEA	07	07	07	08	08	08	.06

1 Lower cutoff values for False Models are more desirable for the fit indices (higher, for SRMR and RMSEA) because this implies 100% of the False models were rejected at the tabled value. With lower cutoff values, the potential for not rejecting True models is increased while rejecting False models.

2 Low values indicate better model fit for all except the 2 indices at the bottom (SRMR and RMSEA).

Table 13

The Joint Optimal Index Value for both Retaining 100% of the Correctly Specified Models and Rejecting 100% of the Misspecified Models (decimals omitted).

	CFA Model Joint Cutoff Results				LCARMA Model Joint Cutoff Results			
	Misspecification Level 2		Misspecification Level 1		Misspecification Level 2		Misspecification Level 1	
	Joint Optimal Cutoff	Required Sample Size	Joint Optimal Cutoff	Required Sample Size	Joint Optimal Cutoff	Required Sample Size	Joint Optimal Cutoff	Required Sample Size
Small Models								
CFI [1]	98	2000	99	2000	97	250	98	250
Mc	98	1000	98	1000	92	125	93	250
Gamma	99	2000	99	2000	95	125	95	250
SRMR	05	500	04	1000	08	500	07	500
RMSEA	08	2000	07	2000	11	250	09	250
Large Models								
CFI	98	250	98	500	97	125	98	125
Mc	90	250	90	500	80	250	80	250
Gamma	97	500	97	500	90	125	93	125
SRMR	08	250	05	500	N/A [2]	N/A	N/A	N/A
RMSEA	05	500	05	500	08	125	07	125

[1] CFI=Comparative Fit Index; Mc=McDonald's Measure of Centrality; Gamma=Steiger's Gamma Index; SRMR=Standardized RMR; RMSEA=Root Mean Squared Error Approximation. [2] N/A suggests that there was no optimal point for the SRMR values calculated for LCARMA Models

models, still required very high cutoffs for Small and Large LCARMA models (and the Large 3 factor, CFA model with crossloadings).

An RMSEA cutoff of .05 was capable of being optimal at n=500 for a Large 3 factor, CFA model with crossloadings and optimal at n=250 for Large or Small LCARMA models. RMSEA cutoffs could be less restrictive, at lower sample sizes, for LCARMA models (Small or Large).

Summary

The findings of this study provide a comprehensive view of how fit index behavior varies across model type, size, sample size, and severity of model misspecification. The results reinforce that no single set of cutoff values—such as Hu and Bentler's (1999) often-cited .95 for CFI and .08 for SRMR—can be universally applied across all structural equation modeling contexts.

One notable trend was that cutoffs traditionally used to assess model fit often failed to detect model misspecification at smaller sample sizes or in models with more subtle forms of misspecification. For example, the CFI consistently required values approaching .99 to fully distinguish correct from incorrect models in several small CFA and large CFA conditions, far exceeding the commonly applied .95 benchmark. Similarly, SRMR values below .08 were insufficient to detect misspecification in some models—particularly the small CFA—where a stricter cutoff of approximately .05 was necessary to reach complete discrimination.

The RMSEA demonstrated more consistent and robust performance across conditions, achieving accurate model rejection decisions with cutoffs near .05 in many contexts, including larger CFA and LCARMA models, even when sample size was as low as 250. This supports the continued utility of the .05 standard for RMSEA under a variety of modeling scenarios.

A key insight emerged concerning the parsimony-based indices PGFI and PNFI. These indices appeared to reward model complexity irrespective of true model status. Under correct model conditions, these indices consistently returned higher fit values for more complex models (such as LCARMA) than for simpler models, suggesting a bias in favor of models with more parameters. Researchers should be cautious in interpreting these indices, as they may encourage overfitting by favoring complex models without meaningful improvements in actual fit.

Another underappreciated factor influencing index performance was the severity of model misspecification. Models designed with lower power (e.g., .56) to detect misfit required substantially larger sample sizes or more extreme cutoff criteria to achieve satisfactory detection. For example, McDonald's index (Mc) could reject all misspecified small LCARMA models at N = 250 when power = .80, but required N > 4000 under less severe misspecification (power = .56). This finding illustrates that fit index thresholds should not be interpreted as fixed values but instead should reflect the expected size and impact of model misspecification.

Unexpectedly, RMSEA values were often lower for large CFA models than small CFA models under correct model conditions, contradicting the assumption that model complexity naturally increases misfit. SRMR similarly failed to meet conventional thresholds for the large true LCARMA model, highlighting how constrained, complex models can affect residual-based fit indices. These findings suggest that neither sample size nor model size alone reliably predict fit index behavior.

Finally, this study calls into question the practice of relying on static, conventional fit index cutoffs. The results show that acceptable thresholds for several commonly used indices—particularly CFI, Gamma, and Mc—varied widely depending on model structure, sample size, and the magnitude of misfit. While RMSEA appears relatively robust, even it can fail when model complexity or low misfit severity introduces challenges to detection. Overall, practitioners are advised to consider the context of their model—its size, constraints, and expected misspecification—when interpreting model fit, and to avoid rigid adherence to universal cutoff values.

References

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. (2002) *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute, Inc.

Fan, X., & Sivo, S. A. (2005). Sensitivity of Fit Indices to Misspecified Structural or Measurement Model Components: Rationale of Two-Index Strategy Revisited. *Structural Equation Modeling*, *12*, 343-367. https://doi.org/10.1207/s15328007sem1203_1

Fan, X., & Sivo, S. A. (2007). Sensitivity of Fit Indices to Model Misspecification and Model Types. *Multivariate Behavioral Research, 42*(3), 509–529. https://doi.org/10.1080/00273170701382864

Fan, X., Thompson, B, & Wang, L. (1999). The effects of sample size, estimation methods, and model specification on SEM fit indices. *Structural Equation Modeling, 6*, 56-83 . https://doi.org/10.1080/10705519909540119

Fan, X., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. *Educational and Psychological Measurement, 58*, 699-733. https://doi.org/10.1177/0013164498058005001

Gerbing, D. W. & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long, (Eds.), Testing structural equation models (pp.40-59). Newbury Park, CA: SAGE Publications, Inc.

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

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

Jöreskog, K., & Sörbom, D. (2001). *LISREL 8: User's reference guide*. Lincolnwood, IL: Scientific Software International, Inc.

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

Lykken, D. T. (1968). Statistical significance of psychological research. *Psychological Bulletin, 70*, 155-159. https://doi.org/10.1037/h0026141

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). En evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315-353). Erlbaum: Mahwah, NJ.

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. https://doi.org/10.1207/s15328007sem1103_2

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 181-204). Newbury Park, CA: SAGE Publications.

Satorra, A, & Saris, W. E. (1983). The accuracy of a procedure for calculating the power of the likelihood ration test as used within the LISREL framework. In C. O. Middendorp (Ed.), Sociometric research 1982 (pp. 129-190). Amsterdam: Sociometric Research Foundations.

Sivo, S.A., Fan, X., Witta, E.L., & Willse, J. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. Journal of Experimental Education, 74(3), 267-288. https://doi.org/10.3200/JEXE.74.3.267-288

Sivo, S.A., Saunders, C., Chang, Q., & Jiang, J.J. (2006). How low should you go? Low response rates and the validity of inferences in IS research using questionnaires. *Journal of the Association of Information Systems, 7*(6), 351-414. https://doi.org/10.17705/1jais.00093

Steiger, J.H., & Lind, J.C. (1980, May). Statistically based test for number of common factors. Paper presented at the Psychometric Society Meeting, Iowa City.

Wheaton, D.E., Muthén, B., Alwin, D.F., & Summers, G.F. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology 1977* (pp. 84-136). San Francisco: Jossy-Bass.

Yu, C. & Muthén, B. (2002, April). *Evaluation of the model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*,115–148. https://doi.org/10.1207/s15327906mbr4001_5