STRUCTURAL EQUATION MODELING
2025, VOL. 32, NO. 2
ISSN: 1070-5511 | E-ISSN: 1532-8007

Evaluating Local Model Misspecification with Modification Indices in Bayesian Structural Equation Modeling

Mauricio Garnier-Villarreal[a] and Terrence D. Jorgensen[b]

[a]Vrije Universiteit Amsterdam; [b]University of Amsterdam

ABSTRACT

Model evaluation is a crucial step in SEM, consisting of two broad areas: global and local fit, where local fit indices are used to modify the original model. In the modification process, the modification index (MI) and the standardized expected parameter change (SEPC) are used to select the parameters that can be added to improve the fit. The purpose of this study is to extend the application of MI and SEPC to Bayesian SEM. We present how researchers can estimate posterior distributions of MI and SEPC using a posterior predictive model check (PPMC). We evaluated the effectiveness of these PPMCs with a simulation and found that MI can be used to detect the most relevant added parameters and that SEPC can be used as an effect size. Similar to maximum-likelihood estimation, the SEPC can overestimate the population value. Lastly, we present an example application of these indices.

KEYWORDS

Bayesian structural equation modelling; blavaan; model modification; modification index; standardized expected parameter change

In this paper, we propose and evaluate a method to detect local underspecification of a structural equation model (SEM) using Markov chain Monte Carlo (MCMC) estimation and Bayesian inference. SEM applications ubiquitously involve evaluating data–model correspondence (or "model fit") using global and local fit statistics. Global fit statistics have recently received much attention in Bayesian SEM methodology literature (e.g., Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018; Levy, 2011; B. O. Muthén & Asparouhov, 2012), but only limited suggestions have been provided on how to detect local misspecifications after global statistics indicate a problem with the model. A popular (though somewhat controversial) method in standard frequentist SEM (f-SEM) applications is to use score tests (or "modification indices") to detect specification errors (e.g., MacCallum, 1986; Saris et al., 1987; Sörbom, 1989; Whittaker, 2012). We show how this method can be incorporated into a Bayesian framework, which we evaluate using a Monte Carlo simulation study.

Although we presume some familiarity with SEM, we nonetheless provide a brief introduction to model-evaluation methods in a traditional frequentist framework as well as a Bayesian framework. We pay particular attention to the modification index (MI) and its corresponding expected parameter change (EPC). Indeed, these statistics are the focus of our proposal. We then describe our simulation study, present our results, and provide an example application to demonstrate how the Bayesian framework can add value to the use of MIs and EPCs.

1. Bayesian Methods

Bayesian methods have been gaining popularity in a variety of areas of research, leading to methodological advancements in areas such as Bayesian structural equation modeling (BSEM). The Bayesian framework can be useful in experimental psychology (Merkle & Wang, 2018), can improve estimation with small samples (Smid et al., 2020), can identify unknown change-points in genetic studies (Jeong & Kim, 2013), or can evaluate measurement invariance in a more flexible way (van de Schoot et al., 2013). User-friendly software, such as M*plus* (L. K. Muthén & Muthén, 1998–2012), Amos (Arbuckle, 2012), and the R (R Core Team, 2023) packages blavaan (Merkle et al., 2021), rstan (Stan Development Team, 2024), and brms (Bürkner, 2017) have improved accessibility to advanced Bayesian methods.

As BSEM methods increase in popularity and application (Depaoli, 2021; van de Schoot et al., 2017), it is necessary to improve our understanding of a variety of components related to model evaluation, given that model misspecification has an impact on model performance, accuracy, and inference (Kaplan, 1988). Previously, B. O. Muthén and Asparouhov (2012) proposed the use of small-variance priors for "nontarget" parameters to test whether they should be included in the model, which can provide evidence about whether a parameter should be estimated with a less restrictive prior (Depaoli, 2021; Garnier-Villarreal & Little, 2023; B. O. Muthén & Asparouhov, 2012). The posterior distribution of a nontarget parameter "can be used in line with modification

indices to free parameters for which the credibility interval does not cover zero" (B. O. Muthén & Asparouhov, 2012, pp. 316–317), despite being partially constrained by the prior information. However, further research has revealed that small-variance priors can limit the ability to detect misspecifications in Bayesian CFAs by artificially improving global fit indices. These small-variance priors add complications related to the propagation of bias (Jorgensen & Garnier-Villarreal, 2023; Jorgensen et al., 2019).

In applied research, the evaluation of model specification takes a main role. Decades of developing methods to evaluate model fit in f-SEM have yielded a range of methods to better detect misspecification in order to help researchers improve model accuracy. Recently, developments in BSEM to evaluate approximate global model fit (Garnier-Villarreal & Jorgensen, 2020) have complemented model testing with a posterior predictive model check (PPMC), but the tools and strategies to evaluate local model fit in the Bayesian framework are still lacking.

Given the limited utility of a Bayesian credible interval (BCI) to detect parameters estimated with small-variance priors as misspecified (i.e., with a prior too restrictive) and the lack of guidelines on BSEM in comparison with f-SEM, we sought to investigate the implementation of modification indices in BSEM and their ability to detect omitted parameters—providing a Bayesian analog to the methods commonly applied in f-SEM. This would allow researchers to apply the method with similar metrics.

2. Model-Evaluation Methods for SEM

The correspondence between a hypothesized structural equation model (SEM) and observed sample data can be evaluated globally using a χ^2 test statistic (calculated from various least-squares or likelihood-based discrepancy functions) and approximate fit indices (e.g., the comparative fit index: CFI; Bentler, 1990). If these global statistics or fit indices suggest inadequate data–model correspondence, the SEM could be respecified in light of theoretical considerations and evidence from the data, after which the modified SEM should be retested (MacCallum et al., 1992). Because SEMs are a priori models and based on theoretical relationships between observed and latent variables, poor fit of a model may be interpreted as evidence that the theoretical model is not plausible and should be rejected or that some particular aspect of the model is poorly specified. In the latter case, the model should be modified by further investigating the local sources of misspecification (e.g., correlation residuals or modification indices; Garnier-Villarreal & Jorgensen, 2020).

2.1. Global Indices

Global indices of model fit are commonly calculated based on the χ^2 statistic or the matrix of covariance (or correlation) residuals (Fan & Sivo, 2007; Pavlov et al., 2021; Ximénez et al., 2022). Here we expand on the methods related to the χ^2 test statistic, which is calculated from the discrepancy function used to obtain parameter estimates when fitting a hypothesized model to data. Because SEMs were traditionally developed as analyses of covariance structure, most discrepancy functions available in SEM software (Kline, 2011) are based on comparing the sample covariance matrix **S** (an unrestricted estimate of the population covariance matrix *Sigma*) to the model-implied covariance matrix $\Sigma(\hat{\theta})$ or simply $\hat{\Sigma}$.

The standard χ^2 statistic tests the null hypothesis (H_0): The hypothesized model perfectly reproduces the population covariance matrix. When using maximum likelihood estimation (MLE), this is a likelihood-ratio test (LRT) comparing the (null) hypothesized model to an alternative (H_A) model that freely estimates all (co)variances. Because theoretical models, by definition, merely approximate reality (MacCallum, 2003), the H_0 of exact fit is often considered a priori to be false. Because researchers cannot reasonably expect to retain H_0 in practice, the χ^2 test is often of limited general or practical interest (West et al., 2012).

Furthermore, the power of the χ^2 test to detect small (even negligible) inconsistencies with H_0 increases with N. To assess whether a model's misspecification is of any practical importance (i.e., whether predicted values are close enough to observed values to be useful in practice), several methodologists have proposed indices of approximate fit to complement the χ^2 significance test—functionally similar to providing measures of practical significance to complement significance tests in other contexts (e.g., Cohen's d to complement a t test). Most proposed fit indices make use of the χ^2 value by adjusting it or comparing it to another model's χ^2 to correct for model complexity, number of parameters, or overfitting, for example. So the χ^2 statistic has long remained the focus of overall model fit in SEM, even if indirectly.

In this paper, we focus on approximate indices that can be informally classified into two groups. Noncentrality-based fit indices include the root mean-squared error of approximation (RMSEA; MacCallum et al., 2006; Steiger & Lind, 1980) and "gamma-hat" ($\hat{\Gamma}$; Maiti & Mukherjee, 1990; Steiger, 1989). Incremental fit indices include the Tucker–Lewis index (TLI; Tucker & Lewis, 1973) and CFI (Bentler, 1990).

2.2. Posterior Predictive Model Checks in SEM

In Bayesian model evaluation, PPMC (Gelman et al., 1996) is a flexible method to test whether aspects of a model adequately capture features of the data. When MCMC estimation is used to estimate model parameters, a discrepancy function can be specified to capture the degree to which a meaningful feature of the observed data differs from its expected value, given the model parameters at iteration i of a Markov chain that has converged on the posterior distribution. If the model is not an adequate representation of the true data-generating process (or at least, cannot make sufficiently similar predictions about observed data), the realized value of the discrepancy function will be large for the observed data (D_i^{obs}). Although D is a function of parameters (and data) rather than an estimated parameter itself, D is evaluated using all samples from the posterior distribution of model parameters, resulting in "an empirical approximation to the posterior distribution of the discrepancy

measure, … [also] referred to as the *realized* values of the discrepancy measure" (Levy, 2011, pp. 672–673).

To quantify whether the discrepancy is larger than would be expected due to chance sampling fluctuations, a random sample of *replicated* data is drawn from the population implied by the model parameters (i.e., data that are predicted by the posterior distribution to occur if the model holds) at the same iteration i of the Markov chain. Because the replicated data are consistent with the model, the realized value of the discrepancy function for the replicated data (D_i^{rep}) only reflects sampling error. In contrast to the distribution of D^{obs}, D^{rep} empirically approximates the posterior *predictive* distribution of the discrepancy function—conditional on the data and estimated model parameters (i.e., the expected values of D if the H_0 of perfect data–model correspondence were true).

If the model is consistent with the population that generated the observed data, then $P(D^{\text{obs}} > D^{\text{rep}}) = 50\%$, but this probability will differ from 50% to the degree that the model over- or underpredicts the feature of the data specified by the chosen discrepancy function. The PPP estimates this probability as the proportion of $i = 1, 2, \ldots I$ samples from the posterior (i.e., postadaptation iterations of the Markov chain), for which $D_i^{\text{obs}} > D_i^{\text{rep}}$. Because the sampling distribution of PPP is not uniform in practice, application of traditional null-hypothesis testing criteria (α levels) yields conservative inferences (Levy, 2011); however, many researchers advocate its use as an informative diagnostic for identifying how a model fails (Gelman & Shalizi, 2013), rather than using traditional hypothesis testing.

BSEM software packages (Arbuckle, 2012; Merkle & Rosseel, 2018; B. O. Muthén & Asparouhov, 2012) uniformly report only PPP based on the $\chi 2$ statistic (PPP$\chi 2$), as demonstrated by Levy (2011), who also described other reasonable discrepancy functions to evaluate an SEM (e.g., the standardized root-mean-squared residual [SRMR] and model-implied factor correlations). The posterior predictive distribution of the χ^2 statistic has been used to obtain a distribution of realized values of global model-fit indices based on the χ^2, such as CFI, TLI, RMSEA, and $\hat{\Gamma}$ (Garnier-Villarreal & Jorgensen, 2020). These implementations have shown the power and flexibility of PPMC to evaluate global model fit.

2.3. Local Indices

Specification errors may be the inclusion of irrelevant relationships or the exclusion of relevant relationships (MacCallum, 1986). Most often, poor fit of a model constitutes the exclusion of relevant relationships and is considered by some to have more serious consequences than the inclusion of irrelevant relationships (Saris et al., 2009). This is a natural characteristic of research because no model or theory fully represents the complexity in the population. Applied researchers therefore often add parameters to an initial model to improve its fit. Although true data-generating processes are unknown in practice, and hypothesized models only represent approximations of the truth (Cudeck & Browne, 1983; MacCallum, 2003), theory-guided model modification can be conducted in order to find a model that adequately represents the relationships among observed and latent variables (Hayduk et al., 2005; Saris et al., 2009).

Regardless of the differing views on model modification, it is commonly agreed that modification of an SEM is no longer purely confirmatory or a priori in nature, but it is to some degree exploratory. Similar to subset selection in multiple regression, modifying and retesting an SEM solely on the basis of model fit indices capitalizes on the chance occurrences within the sampled data with which the models are tested. MacCallum (1986) demonstrated the failure of data-driven model modification to lead to a model that better represents the true data-generating process. Models should be modified on the basis of relevant theory. In addition, cross-validation is highly recommended to substantiate the predictive validity of modified SEMs (MacCallum et al., 1992).

In this paper, we take the position that *local model evaluation* is the process of investigating local sources of misspecification after a global test reveals evidence of nonignorable misfit in an attempt to modify and re-evaluate a hypothesized SEM. We see the process of model building as a dynamic conversation between theory and data, rather than a static one. If the H_0 model is rejected, a less restrictive alternative model might be found that is still more restrictive than the completely "saturated" H_A model used for the global χ^2 test of fit. The H_0 model tends to be a researcher's a priori hypothesis, so it follows that there are rarely other a priori hypothesized alternative models. This necessitates a process of "asking the data" how the H_0 model fails, then considering the theoretical plausibility of that new information.

2.4. Modification Index

During the dynamic process of modifying or reevaluating a model, various forms of information can be considered to help select parameters for inclusion in the model, aiming to enhance its goodness of fit. A useful piece of information commonly assessed is the modification index (MI)—a term used in SEM for Lagrange multipliers (or "score test" statistics), which are asymptotically equivalent to the LRT when the H_0 is true (Buse, 1982). The MI is a χ^2 test statistic that approximates how much the H_0 model's LRT statistic would decrease if a constrained parameter were instead freely estimated. Although asymptotically equivalent to the LRT, the Lagrange multiplier test only requires fitting the restricted model (M_0) instead of fitting two competing models (restricted model M_0 and unrestricted model M_1).

The MIs provided by most SEM software packages are $1-df$ Lagrange multipliers associated with each fixed parameter (or equality constraint). Score tests are calculated with a similar functional form as a Wald test statistic (W), which has already been incorporated into BSEM (Asparouhov & Muthén, 2021):

$$W = (\hat{\theta} - \theta_0)' V^{-1} (\hat{\theta} - \theta_0), \tag{1}$$

where $\hat{\theta}$ and θ_0 are the estimated and H_0 parameter values, respectively, and V is the estimated asymptotic covariance matrix of those parameters' sampling distributions. In the case of a single parameter (i.e., a χ^2 test with $df = 1$),

Equation (1) simplifies to the square of the familiar Wald z statistic provided per estimated parameter by any standard SEM software, assuming a $H_0 = 0$:

$$W = \left(\frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}\right)^2 = \left(\frac{\hat{\theta}}{SE(\hat{\theta})}\right)^2 = \frac{\hat{\theta}^2}{\mathrm{Var}(\hat{\theta})}. \quad (2)$$

The Wald test is also asymptotically equivalent to an LRT, comparing models M_0 and M_1, but it only requires estimating M_1.

Analogously, a score test statistic (S) for any fixed parameter can be calculated using information derived from model M_0. Although M_0 does not estimate $\hat{\theta}$, the first and second derivatives of the likelihood function can be calculated with respect to the constrained θ_0 (Chou & Bentler, 1990; Saris et al., 1987), which are typically parameters fixed to $\theta_0 = 0$. The first derivative is the gradient G, which is a vector with q elements (i.e., 1 per estimated parameter) minimized to 0 in order to obtain ML estimates. The matrix of second derivatives is the information matrix (I, the negative of the Hessian), which is used to verify that the algorithm has obtained a maximum (rather than minimum) likelihood solution. To obtain a score test for a parameter fixed to 0, the gradient is augmented by calculating the first derivative additionally for the fixed parameter (G_{q+1}), which was not minimized to 0 during estimation. The gradient is a local slope of the likelihood surface, "pointing" in a direction of the augmented parameter space that would further maximize the likelihood of the data.

If the H_0 is true, $G_{q+1} = 0$ is still possible due to sampling error, which can be estimated by similarly augmenting the information matrix. The more substantially the additional parameter would increase the likelihood, the greater the value r on the diagonal of an augmented information matrix:

$$I^{\mathrm{aug}} = \begin{bmatrix} P & Q \\ Q' & r \end{bmatrix}, \quad (3)$$

where P is the $q \times q$ information matrix of the estimated parameters in M_0, and r is the scalar value for the additional (fixed) parameter. The sampling variance of the augmented gradient G_{q+1} is calculated by partialing out its sampling covariance with the estimated model parameters, contained in the off-diagonal quadrant Q:

$$v = r - Q(P^{-1})Q'. \quad (4)$$

The score statistic is then calculated as:

$$S = \frac{G_{q+1}^2}{v}, \quad (5)$$

which resembles the Wald statistic in the last term of Equation (2).

Fixed parameters associated with a substantial MI value, for example, exceeding the $\chi^2(df = 1)$ critical value $= 3.84$ at $\alpha = .05$, are then examined to determine whether they are theoretically sound to include in the model by estimating freely. Typically, the process of using the MI in model respecification begins with assessing whether adding any fixed parameters significantly reduces the model's χ^2 test statistic. If so, researchers often explore the set of statistically significant potential modifications to identify which one results in the most significant reduction in the χ^2 statistic. If the identified modification is theoretically plausible, it may be incorporated to enhance the model fit; if not, then another substantial MI is considered. This process is repeated until adding fixed parameters no longer significantly reduces the χ^2 or when none of the statistically significant potential modifications align with theoretical feasibility (Bollen, 1989).

Research investigating the MI's accuracy in guiding researchers to the correct SEM has revealed limitations. For instance, MacCallum (1986) conducted a simulation study under various conditions, testing the MI's performance in identifying the correct SEM. The study included fitting different types of incorrectly specified SEMs to datasets generated based on the true SEM, with small ($N = 100$) and moderately small ($N = 300$) sample sizes. The MI's suggestions for adding parameters were recorded, and models were retested after adding the parameter recommended by the most significant MI (using $\alpha = .01$). Both restricted and unrestricted specification searches were conducted. The findings indicated that the MI-based specification searches were less successful when the initial model had more specification errors. Restricted searches yielded better outcomes than unrestricted ones, with no successful outcomes observed in the small sample size scenario.

Kaplan (1988) and Silvia and MacCallum (1988) conducted studies with similar results as MacCallum (1986), highlighting the limitations of the MI. In contrast, Chou and Bentler (1990) conducted a simulation study that yielded more promising results for the MI's performance. However, there were occasional Type I errors (i.e., a significant MI flagging an irrelevant fixed parameter to free), particularly with moderate to large sample sizes. The success of the MI in this study may have been influenced by the severity of model misspecification when assessing whether fixed parameters should be freely estimated to significantly improve the model's χ^2.

MacCallum et al. (1992) further demonstrated the instability of model modification through two sampling studies. These studies involved creating various sample size conditions (ranging from 100 to 1200) by randomly selecting cases from an existing dataset. The results indicated inconsistent model modifications across replication samples, especially with small sample sizes ($N = 100$).

Additionally, Hutchinson (1993) conducted a population study comparing the performance of univariate and multivariate MIs in arriving at the correct model. The results showed that both the univariate and multivariate MI tests correctly suggested freeing parameters in less misspecified models but became less reliable as model misspecification increased. Overall, the multivariate MI test did not outperform its univariate counterpart. It is important to note that incorrect parameters were intentionally included in the misspecified models in this study, which may have affected the results and their comparability to previous studies.

2.5. Expected Parameter Change

Hutchinson (1993) observed a potential limitation in her research regarding the omission of expected parameter

change (EPC) in conjunction with MI. This inclusion could have enhanced the effectiveness of specification searches. Consequently, she suggested the exploration of this combined approach in future studies. The concept of EPC was initially introduced by Saris et al. (1987) to aid in the selection of model parameters for potential inclusion, aiming to detect substantial model misspecifications. The EPC is a function of the score statistic and augmented gradient value:

$$\text{EPC} = \frac{S}{G_{q+1}} = \frac{G_{q+1}^2}{\nu} \times \frac{1}{G_{q+1}} = \frac{G_{q+1}}{\nu}, \tag{6}$$

which predicts the freely estimated value of the tested parameter. EPCs offer a direct estimation of the magnitude of misalignment for the restricted parameters. Model modification decisions using an MI and its EPC can incorporate both statistical and practical significance.

Saris et al. (1987) outlined four scenarios where the EPC could play a pivotal role in model respecification. First, when a statistically significant MI coincides with a substantially large EPC value for a parameter, it is reasonable to conclude that this parameter should be freely estimated. Second, if a statistically significant MI aligns with a small EPC value, it is not advisable to free estimate that parameter; this may be a negligible parameter that is nonetheless significant due to an adequate sample size to yield high power. Third, when a nonsignificant MI coincides with a substantial EPC value, the decision is unclear, and a power analysis is recommended. Fourth, if a nonsignificant MI aligns with a small EPC value for a parameter, it is not recommended to freely estimate that parameter because there is evidence of neither statistical nor practical significance. Notably, Saris et al. (1987) did not propose specific cutoff criteria for determining the magnitude of an EPC value that warrants the inclusion of a particular parameter.

Luijben and Boomsma (1988) conducted a simulation study comparing the performance of MIs and EPCs under various conditions, including different sample sizes ($N = 200$ and 400), factor loading magnitudes (.40, .60, and .80), factor loading structures (with observed variables loading on one or two factors), and factor intercorrelations (.30, .40, and .50). Their findings indicated that both MIs and EPCs improved as sample size, factor loading magnitude, and factor intercorrelations increased. Moreover, the EPC outperformed (higher power) MIs in suggesting the inclusion of factor intercorrelations in the model. Furthermore, Luijben and Boomsma (1988) demonstrated that the unstandardized EPC used in their study depended on the scaling of variables in the model. They conducted a smaller-scale simulation study, demonstrating a condition in which the unstandardized EPC no longer suggested estimating the missing factor intercorrelation due to its scale noninvariance.

Kaplan (1989) proposed a standardized version of the EPC, referred to as P-SEPC, and examined the combined use of MIs, EPCs, and P-SEPCs in specification searches for two models applied to existing datasets. The combination of substantial MIs and (P-S)EPCs for a fixed parameter resulted in significant improvements in model fit after these parameters were allowed to vary freely. Kaplan (1989) also concluded that MIs tended to suggest estimating improbable parameters, while (P-S)EPCs suggested estimating plausible parameters within a model. It is important to note that in this study, an EPC value greater than .10 was considered substantial.

Nonetheless, the partially standardized version of the EPC (P-SEPC) proposed by Kaplan (1989) still depended on the metric properties of variables in the model, as discussed by Chou and Bentler (1993) and Luijben and Boomsma (1988). Chou and Bentler (1993) recommended fully standardizing the EPC, now referred to as the SEPC, to ensure its invariance to the rescaling of observed and latent variables. The major difference between the two standardizations lies in the treatment of residual covariances between latent and observed variables, which are standardized using residual variances under SEPC standardization—as opposed to marginal variances under P-SEPC standardization. In essence, SEPC standardization scales everything to have unit variance.

2.6. Modification Indices in BSEM

In this paper, we assessed the use of MI and SEPC in BSEM. Following the same logic and process as the global approximate fit indices, we proposed the use of PPMC to estimate posterior distributions of MI and SEPC to evaluate the need (relevance) of adding an extra parameter to the model.

We expected that the MI posterior distribution should be useful to identify which parameter should be added into the model. Rather than relying on the theoretical χ^2 distribution, we used the PPMC to compare the MI at each iteration of a Markov chain to a corresponding M, calculated using data generated from the model parameters sampled at that iteration. This allowed us to calculate a posterior predictive p value (the PPP) for local fit assessment, similar to the PPP used to evaluate global fit. Similarly, a PPMC can utilize the SEPC as a discrepancy function, which can also have a PPP value.

In this project, we evaluated the performance of the PPMC method, using MI and SEPC as discrepancy measures to flag omitted parameters that could be considered for respecification, as well as estimate (with uncertainty) their expected size when estimated. We evaluated this proposal with a simulation study, presented next.

3. Simulation

The simulation was conducted in the `R` platform (R Core Team, 2023), using the following packages: `portableParallelSeeds` to control the seeds and random number generating process (Johnson, 2016), `simsem` for data simulation (Pornprasertmanit et al., 2021), `lavaan` for the estimation of the frequentist SEM with ML (Rosseel, 2012), `blavaan`[1] for the estimation of Bayesian SEM and calculation of PPMCs (Merkle et al., 2021), `semTools` (Jorgensen et al., 2023) for calculating additional ML fit indices not available from `lavaan`, and the `parallel` package (part of R's base

[1]There is a tutorial on how to estimate these modification indices with blavaan already on their website https://ecmerkle.github.io/blavaan/articles/mod_indices.html.

distribution) to run multiple analyses in parallel on a high performance computing cluster.[2] The goal of our study is to explore some frequency properties of our proposed PPMC for local-misfit detection, described in the Monte Carlo Outcomes section.

3.1. Design

We employed MIs across the full variety of SEMs, including path models (Saris et al., 2009), factor models (Whittaker, 2012), full SEMs (MacCallum, 1986; MacCallum et al., 1992), and growth curve models (Kwok et al., 2010). Because cross-loadings and residual correlations are so frequently used to modify measurement models (and have been a main argument in favor of flexible BSEMs; Asparouhov et al., 2015; B. O. Muthén & Asparouhov, 2012), we designed a study to investigate the degree to which posterior distributions of MIs and SEPCs could reliably flag such omitted parameters as potentially misspecified. In order to simulate data across a variety of realistic conditions, we specified population models with a cross-loading and residual correlation of randomly varying magnitudes, both of which would be constrained to zero in the fitted SEM. We did not employ a fixed factorial design because the goal is not to learn about any specific condition, as we might in a Monte Carlo power analysis. Rather, the goal is to demonstrate whether trends match expectations. That is, do larger samples and larger effect sizes lead to omitted parameters being more frequently flagged, and are SEPCs good estimates of the omitted parameters?

We simulated data from a three-factor CFA, with four primary indicators per factor. Standardized factor loadings were $\lambda = 0.7$, factor correlations were $\rho = 0.4$, and residual variances were $1 - 0.7^2$ (yielding observed variances = 1). In most conditions (described below), the population model also included one or two nonzero parameters that were fixed to 0 in the analysis model: a cross-loading and a residual correlation. The path diagram in Figure 1 provides population parameters. The analysis model is the same three-factor CFA, but without any cross-loadings or residual correlations (i.e., the grey paths in Figure 1 were fixed to zero). This allowed us to evaluate the local misspecification of our fitted model using a PPMC to detect the most common types of omitted parameters, using MI and SEPC for each fixed parameter as discrepancy functions.

The varying conditions in the simulation were sample size ($N = 100$, 200, 300, 400, or 500), magnitude of the standardized cross-loading (CL: $\lambda_{3,2} = 0.0$, 0.1, 0.2, 0.3, or 0.4), and magnitude of the residual correlation (RC: $\theta^*_{6,10} =$ 0.0, 0.1, 0.2, 0.3, or 0.4). The residual variance of y_3 varied with CL to maintain a marginal variance = 1 (i.e., standardized data). The RC of y_6 with y_{10} was scaled by 0.51 (i.e., square-root of the product of residual variances) to obtain residual covariances $\theta_{6,10} = 0$, 0.051, 0.102, 0.153, or 0.204 across RC conditions.

The magnitudes of misspecification (CL and RC) were chosen to represent a range of realistic situations, in which researchers attempt to respecify a model that has one or more specification errors of varying degrees. Correlations or standardized regression slopes (loadings) as small as 0.1 (Cohen, 1992) could be considered negligible enough to be ignored (B. O. Muthén & Asparouhov, 2012), whereas values in the medium range might be considered substantial enough to be included as unconstrained parameters. B. O. Muthén and Asparouhov (2012) simulated a similar range of CL and RC values, but only as high as 0.3.

To give a clearer picture about how to interpret the size of these misspecifications, Table 1 shows how the population SRMR was impacted by each combination of omitted nonzero CL and RC. An omitted CL presented about twice the effect on SRMR as the same value in the RC. Furthermore, Tables 2 and 3 present the power to detect each omitted parameter across all conditions of misspecification when $N = 150$ (arbitrarily chosen to illustrate the magnitude of misspecification conditions). We observed greater power to detect increasing levels of omitted CL compared to RC, likely because the actual parameter was not a correlation but a covariance, which was about half the size due to the residual variances being 0.51. Note that power ranged from about 5% (the nominal α level) when CL or RC = 0 to 100% when they equaled 0.4. Thus, our conditions represented the full range of power when $N = 150$.

We followed a random-sampling approach for the varying conditions by running the analysis for 30,000 replications, and at each replication we sampled a random N, CL, and RC for the population model. With a total of 125 combinations between these conditions, we expected random sampling to yield an average of $\frac{30,000}{125} = 240$ replications per condition, which we judged as sufficient to evaluate general patterns of interest in the results. We observed an average count of 240.1 replications per condition, with a range of 195–299. The first author conducted the simulations using the university's High Performance Computing cluster, which took 2.03 years of (distributed) computing time to complete.

3.2. Monte Carlo Outcomes

We evaluated our proposed PPMC to detect local misfit using the following criteria, calculated for each sample.

1. When rank ordering the posterior mean of MIs in each sample, did the largest MI correspond to one of the omitted parameters (i.e., $\lambda_{3,2}$ or $\theta_{6,10}$, when either parameter was = 0)? Because researchers typically explore numerous MIs for candidate parameters and because there were two valid candidates in most of our conditions, we also recorded whether $\lambda_{3,2}$ or $\theta_{6,10}$ was among the five largest (posterior-mean) MIs for a sample.
2. Similar to MIs, we recorded whether the largest (posterior-mean) SEPC corresponded to a valid candidate parameter and whether $\lambda_{3,2}$ or $\theta_{6,10}$ were among the top five largest SEPCs. We found that the highest SEPC did not perform as well as the largest MI, so we only focus on the MI criterion in our Results section.

[2]The code for the simulation can be found in the OSF site for this project .https://osf.io/kdq5y/.

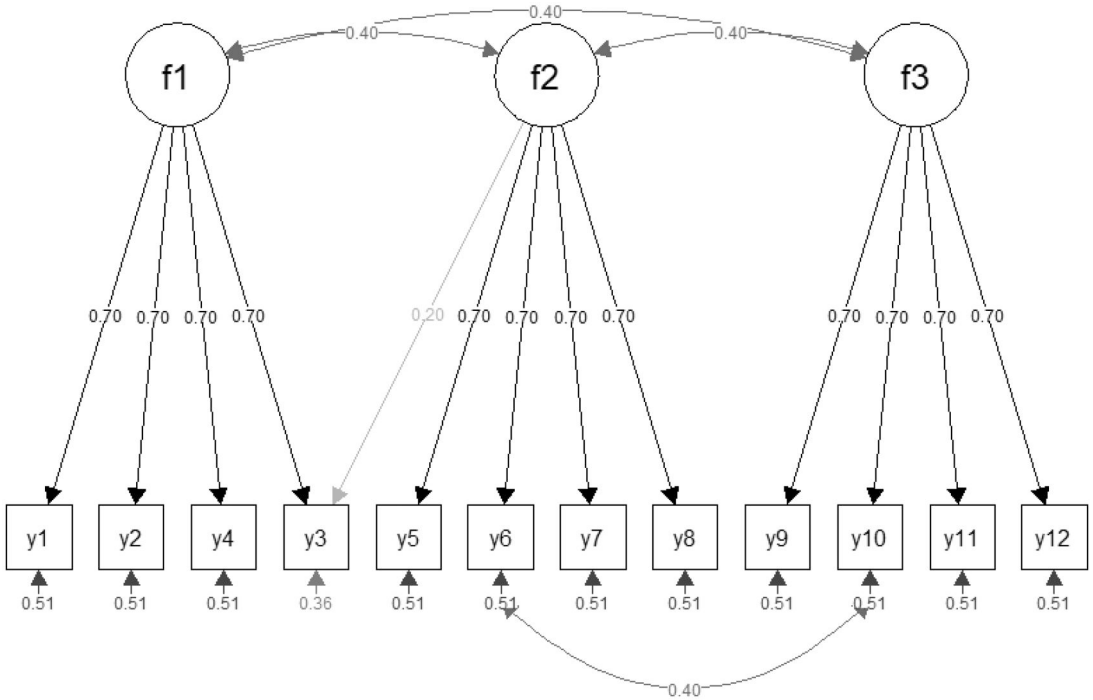

Figure 1. Path diagram including population parameters for data generation. The grey paths could vary across conditions but were always fixed to 0 in the analysis model. Factor variances were fixed to 1 in both the population and analysis models.

Table 1. SRMR when missing cross-loading and/or residual correlation.

	RC				
CL	0	0.1	0.2	0.3	0.4
0	0.000	0.005	0.011	0.016	0.022
0.1	0.011	0.012	0.015	0.020	0.024
0.2	0.021	0.021	0.024	0.027	0.030
0.3	0.031	0.031	0.033	0.035	0.038
0.4	0.044	0.044	0.045	0.047	0.049

Table 2. Power to detect the standardized cross-loading when $N = 150$.

	RC				
CL	0	0.1	0.2	0.3	0.4
0	0.056	0.056	0.066	0.062	0.084
0.1	0.244	0.244	0.266	0.216	0.226
0.2	0.710	0.726	0.768	0.712	0.748
0.3	0.982	0.978	0.978	0.976	0.978
0.4	1.000	1.000	1.000	1.000	1.000

Table 3. Power to detect the residual correlation when $N = 150$.

	RC				
CL	0	0.1	0.2	0.3	0.4
0	0.046	0.154	0.446	0.834	0.974
0.1	0.058	0.172	0.452	0.838	0.972
0.2	0.036	0.158	0.472	0.832	0.974
0.3	0.052	0.174	0.476	0.848	0.980
0.4	0.052	0.192	0.476	0.860	0.976

3. How biased were the posterior-mean SEPCs of $\lambda^*_{3,2}$ and $\theta^*_{6,10}$ relative to their population values? We calculated bias as the difference between the SEPC and the standardized population parameter.
4. The PPMC method also provides a PPP value, which could be used to conduct a test similar to a frequentist H_0 significance test. However, it has been observed that PPP does not behave quite like the frequentist p value (e.g., not uniformly distributed between 0–1 under H_0), so we did not calculate power or Type I error rates for it. Instead, we investigated the distribution of PPP values only among the fixed parameters flagged by MIs (as candidates to freely estimate) to ascertain whether they tended to be low enough to give researchers the impression that the H_0 ($\lambda_{3,2} = 0$ or $\theta_{6,10} = 0$) was false.

4. Results

4.1. Detecting an Omitted Parameter

We first looked at the MIs' ability to *find* the omitted parameters. Figure 2 shows the proportion of data sets where the omitted cross-loading or residual correlation was flagged as the first omitted parameter or in the top five parameters to be included (across population value and sample size). The first column presents the results when the cross-loading increased in effect size, holding the residual correlation to 0. Similarly, the top row presents the results when the residual correlation increased in effect size, holding the cross-loading to 0. The table shows the simple main effects of each type of parameter. The trends across other panels of the figure represent the interaction of nonzero effect size for both parameters.

When the population cross-loadings were 0, the omitted parameters were rarely found within the recommended added parameters, with a 5% chance of being in the top five of recommended parameters. When the population cross-loadings were higher than 0, we observed that as sample size increased, the proportion of data sets that found the omitted parameters increased as well. When the population value was 0.3, the MIs found the omitted parameters at least

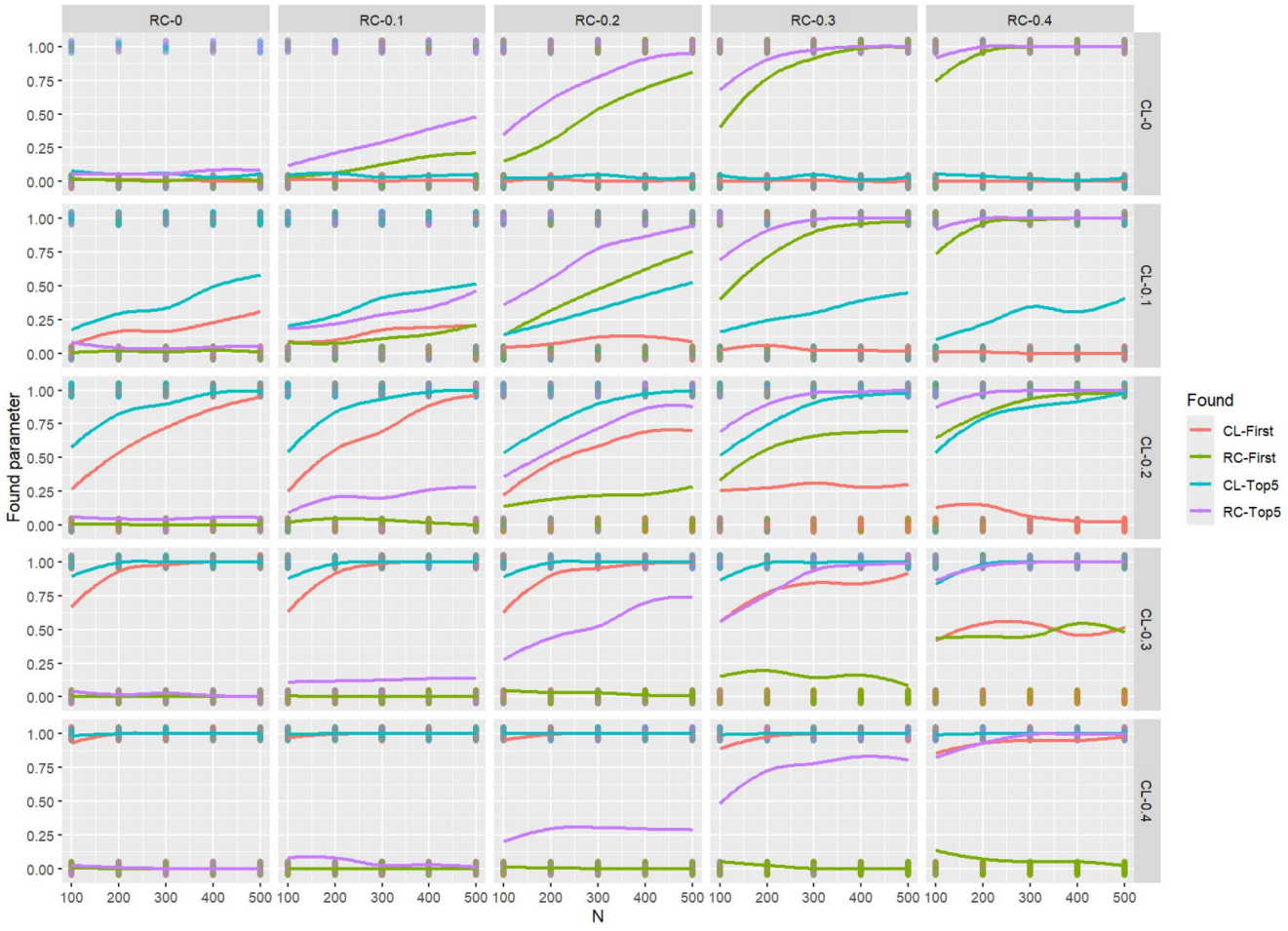

Figure 2. Detecting omitted parameters by MI.

87% of the time in the top five and had at least a 58% chance of finding the recommended parameter first, when $N = 100$. With higher sample sizes, we had at least an 80% chance of finding the omitted parameters as the first recommendation or 99% in the top five. When the population parameter was 0.4, we had an over 90% chance of finding the omitted parameters as the first recommendation. With higher sample sizes, we were able to find the omitted parameters as the first recommendation 99% of the time. In the top five, our chances increased to 100% of the time.

For the omitted residual correlations, where the omitted residual correlation was found, we observed that the proportion of the data sets was included as the first or in the top five parameters (across population value and sample size). When the population value was 0, we had low rates of finding the residual correlation, with a maximum chance of 5%. When the parameter was higher than 0, we observed that as sample size increased, the proportion of data sets in which we found the residual correlation increased as well. When the parameter was 0.3, percentages of found residual correlation ranged from 27% to 57%, as the first parameter, and from 62% to 96% in the top five. When the population parameter was 0.4, the percentages of found residual correlation ranged from 54% to 69%, as the first parameter, and from 88% to 99% in the top five.

Because our simulation design had conditions where both omitted parameters were not 0, it was necessary to look at the proportion of data sets in which both the cross-loading and residual correlation were flagged. When one of these was 0, the other parameter followed the same patterns as presented earlier. When both of these correlations were present, as their effect size increased, they had a greater chance of being detected in the top five. The cross-loading was found as the first parameter more often than the residual correlation when both were nonzero in the population—due to the latter having a smaller impact on model misfit (see Table 1).

4.2. SEPC Bias

When we looked at the bias of the posterior-mean SEPC (see Table 4), we noticed that in most cases as sample size increased the bias decreased. For the residual correlation, we saw that bias decreased as the population value increased, whereas for the cross-loading, we found the lowest bias when the population parameter was higher than 0 but lower than 0.4. When the omitted residual correlation was 0.4, the SEPC tended to overestimate it by about 0.3.

Figure 3 shows that as the population value increased for the cross-loading, the SEPC tended to overestimated it. The variability of the SEPC increased as well. But in the case of

Table 4. Bias of the SEPC when the parameter is found.

Population	N	CL	RC
0.00	100	−0.24	−0.25
0.00	200	−0.19	−0.19
0.00	300	−0.15	−0.15
0.00	400	−0.13	−0.14
0.00	500	−0.12	−0.12
0.10	100	−0.16	−0.17
0.10	200	−0.10	−0.11
0.10	300	−0.07	−0.08
0.10	400	−0.06	−0.06
0.10	500	−0.04	−0.05
0.20	100	−0.11	−0.11
0.20	200	−0.05	−0.04
0.20	300	−0.03	−0.03
0.20	400	−0.03	−0.02
0.20	500	−0.03	−0.02
0.30	100	−0.09	−0.05
0.30	200	−0.08	−0.01
0.30	300	−0.08	−0.01
0.30	400	−0.08	−0.00
0.30	500	−0.09	−0.00
0.40	100	−0.27	−0.01
0.40	200	−0.31	0.00
0.40	300	−0.34	−0.00
0.40	400	−0.35	−0.00
0.40	500	−0.35	−0.01

the residual correlations (see Figure 4), the bias was similar across different population levels, and the SEPC variability related to the sample size instead of the population value.

4.3. Distribution of PPP, Given Detection by MI

When we looked at the average PPP across omitted population parameters and sample sizes (see Table 5), we saw that as the population parameter increased the PPP decreased. Also, as sample size increased the PPP decreased. In comparison, the PPP was lower for SEPC than for MI.

When we looked at the distribution of the PPP, we saw that the MIs of the cross-loadings (see Figure 5) presented large variability unless the effect size and the sample size were large enough. We saw less variability (see Figure 6) for the residual correlations. As expected, based on these results, the MI's PPP values were almost always small ($< .05$) for the cross-loadings, given sufficient data (e.g., $N > 300$) and when the population value was least 0.2. Likewise, for the residual correlation MI and SEPC we found a PPP $< .05$ in almost all conditions. However, the PPP for the cross-loading SEPC was around or lower than 0.05 in most cases, with a few exceptions (e.g., when $N = 100$).

5. Example Application

For this example application we used the 25 personality items (BFI) from the International Personality Item Pool (Goldberg et al., 2006), which is available in the R package psychTools (Revelle, 2024). We used the R packages blavaan (Merkle et al., 2021) and loo (Vehtari et al., 2024) for the analyses.[3]

[3]The code for this example can be found in the OSF site for this project https://osf.io/kdq5y/.

We expected the BFI to follow a theoretical structure of the five personality factors: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. These would be estimated as correlated factors, with no cross-loadings or residual correlations, as presented in Figure 7. For this example, we randomly selected a subsample of $N = 500$. All models were estimated with weakly informative priors, of $\lambda \sim N(0.7, 2)$, $\nu \sim N(3, 2)$, $\theta_{sd} \sim \gamma(1, 1)$, $\rho \sim lkj(1)$.

We estimated the first model following the expected theoretical structure, and we observed that overall model fit was lower than desirable. Table 6 displays the posterior distributions of the three overall indices, which indicate that overall model fit can be improved. Among the different indices, it is evident that CFI presents the "worst" fit.

Next, we estimated the posterior distributions for the MI and SEPC. Table 7 shows the five parameters with the highest MI (based on the mean of the posterior). The posterior mean MI = 118.5 strongly suggests that the residual correlation between items N1 (Get angry easily) and N2 (Get irritated easily) is an omitted parameter that will improve the model fit the most, which also makes theoretical sense because these items focus on similar reactions.

We also recommend to evaluate the likely effect size by looking at the SEPC (Saris et al., 2009; Whittaker, 2012). Table 8 shows the five parameters with the highest SEPC (based on the mean of the posterior). The SEPC recommends the same residual correlation as with the average SEPC of 0.85 (PPP = 0). According to our data, the SEPC likely represents a high standardized value if the parameter is added.

Next, we added the recommended residual correlation ($N1 \sim\sim N2$) and evaluated the second model and the impact of the modification. The posterior distribution of the added parameter shows a mean of 0.79 ($SD = 0.09$, 95% CI = [0.61, 0.99], Std = 0.52). So, we see that the previous SEPC was close to the unstandardized posterior mean, but it overestimated the standardized mean ($Std = 0.52$).

In the approximate global fit indices, there is a small improvement when the original indices (see Table 6) are compared to those of the second model (see Table 9). Then we compare the overall model to a sample to determine predictive accuracy using the Leave-One-Out method (Vehtari et al., 2017). This comparison revealed a difference in expected log pointwise predictive density ($elpd$) of $\Delta elpd = -51.68$ ($SE = 13.87$). The ratio between the difference and its respective SE is 3.73, indicating that the model with the added residual correlation presents a meaningful improvement in the model's predictive accuracy. We would conclude to keep the added parameter.

As a next step, we can estimate at the posterior distributions of MI and SEPC from the second model. We can see the summary of the MI in Table 10: The highest improvements based on MI are between 30 and 46, and the SEPC values (see Table 11) are around 0.3.

For this example, we stop modifying the model further, but future researchers could continue modifying it if the model fit improves and if the changes are theoretically sound. With this example, we have shown the use and interpretability of modification indices in BSEM.

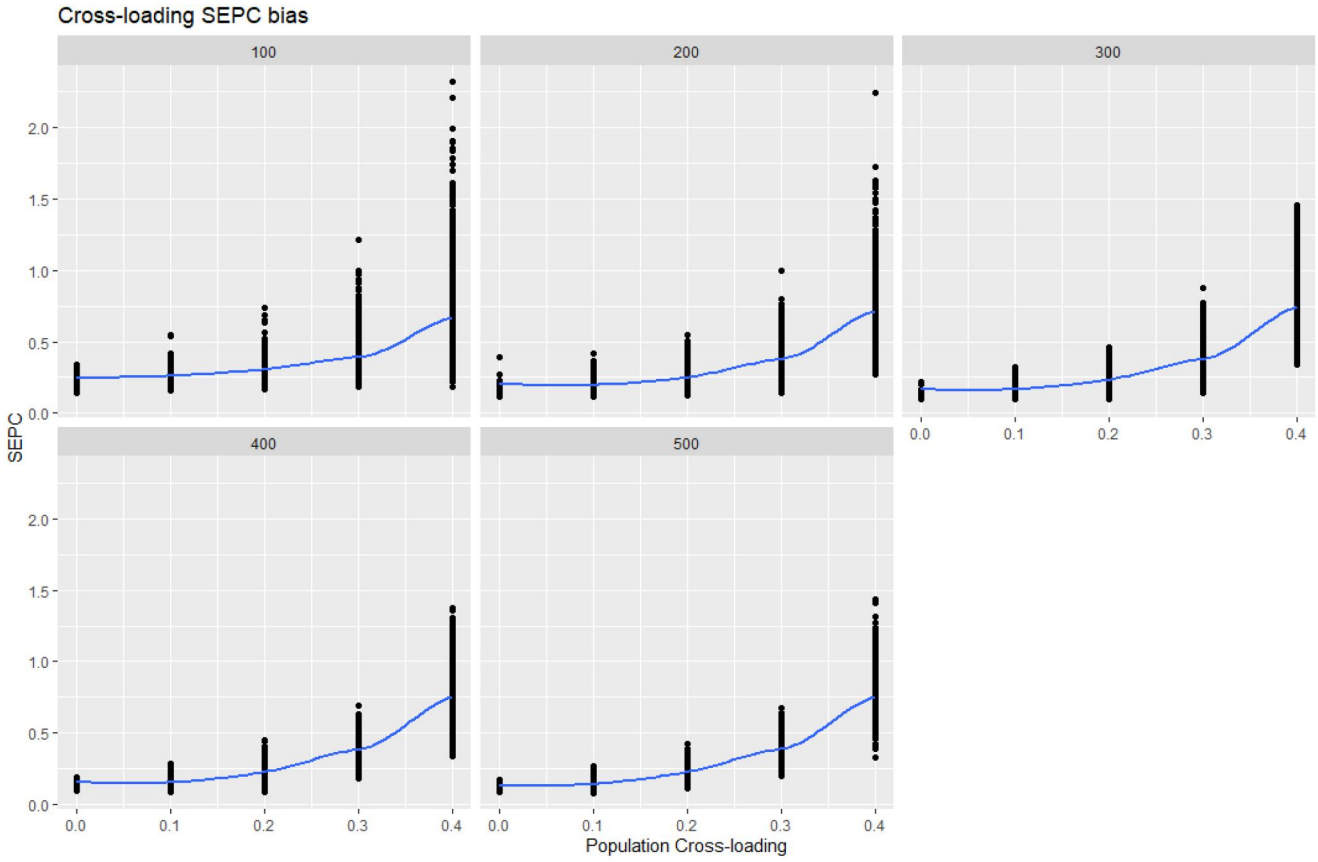

Figure 3. SEPC bias for cross-loadings.

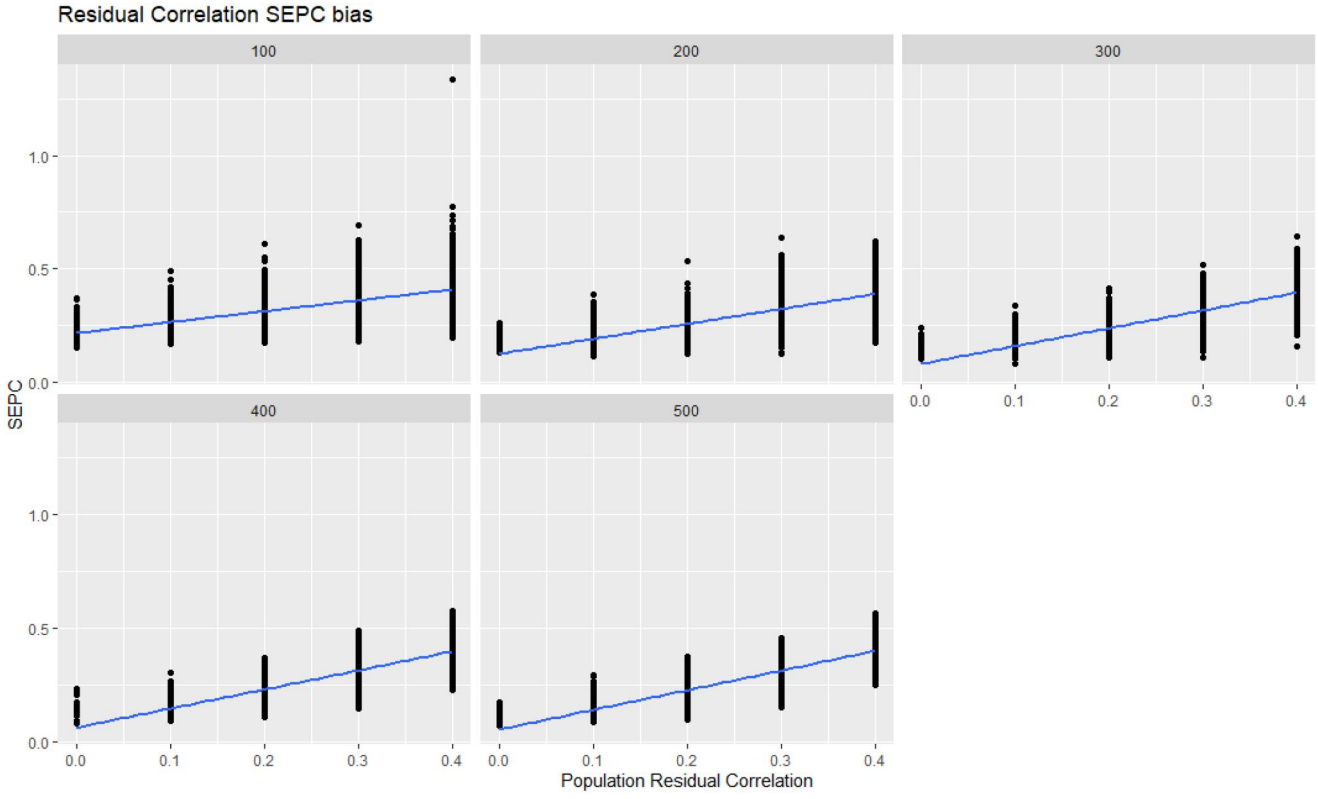

Figure 4. SEPC bias for residual correlations.

6. Discussion

In this project, we present the use of PPMC to estimate posterior distributions of modification indices and how these can be used to evaluate local model fit in BSEM. These modification indices were tested using a simulation study, across sample size, and a magnitude of two types of excluded parameters. In the simulation study, we found that the MIs in BSEM are good indicators to *find* parameters that should be included in the model, identifying the most likely parameters to improve the model fit based on posterior distribution of the Lagrange approximations.

In our Monte Carlo design, the MIs identified cross-loadings as more important than residual correlations due to their larger impact on the implied covariance matrix (see Table 1). A residual correlation will be found as the most relevant parameter if the corresponding residual covariance exceeds the omitted cross-loading.

We also tested the use of SEPCs to approximate the magnitude of the excluded parameters. We found that the SEPCs are adequate at indicating a possible range for the SEPC, but overestimation is more likely when the misspecification is greater (Mansolf et al., 2020) because the SEPC is calculated from the MI (which is greater when the omitted parameter is greater).

When looking at the PPP of MIs and SEPCs, we find that they are sensitive to sample size, effect size, and type of parameter. For this reason, we recommend that researchers be cautious in their interpretation and to use these tools as additional information, rather than the deciding statistic on whether to add a new parameter.

Table 5. PPP for MI and SEPC when the parameter is found.

		MI		SEPC	
Population	N	CL	RC	CL	RC
0.00	100	0.13	0.05	0.09	0.04
0.00	200	0.12	0.03	0.06	0.03
0.00	300	0.11	0.04	0.06	0.03
0.00	400	0.11	0.04	0.06	0.02
0.00	500	0.11	0.03	0.06	0.02
0.10	100	0.15	0.04	0.10	0.03
0.10	200	0.11	0.02	0.06	0.02
0.10	300	0.09	0.02	0.05	0.01
0.10	400	0.08	0.02	0.04	0.01
0.10	500	0.07	0.01	0.03	0.01
0.20	100	0.15	0.03	0.10	0.02
0.20	200	0.09	0.01	0.05	0.01
0.20	300	0.05	0.01	0.03	0.01
0.20	400	0.04	0.00	0.02	0.00
0.20	500	0.02	0.00	0.01	0.00
0.30	100	0.16	0.02	0.11	0.01
0.30	200	0.07	0.00	0.04	0.00
0.30	300	0.03	0.00	0.02	0.00
0.30	400	0.01	0.00	0.01	0.00
0.30	500	0.00	0.00	0.00	0.00
0.40	100	0.17	0.01	0.12	0.01
0.40	200	0.08	0.00	0.05	0.00
0.40	300	0.04	0.00	0.02	0.00
0.40	400	0.02	0.00	0.01	0.00
0.40	500	0.01	0.00	0.01	0.00

7. Recommendations

Based on these results, we can tentatively recommend that researchers utilize MIs and SEPCs in a PPMC framework, following similar guidelines as the ones suggested by Whittaker (2012). That is, the MI should be used to indicate

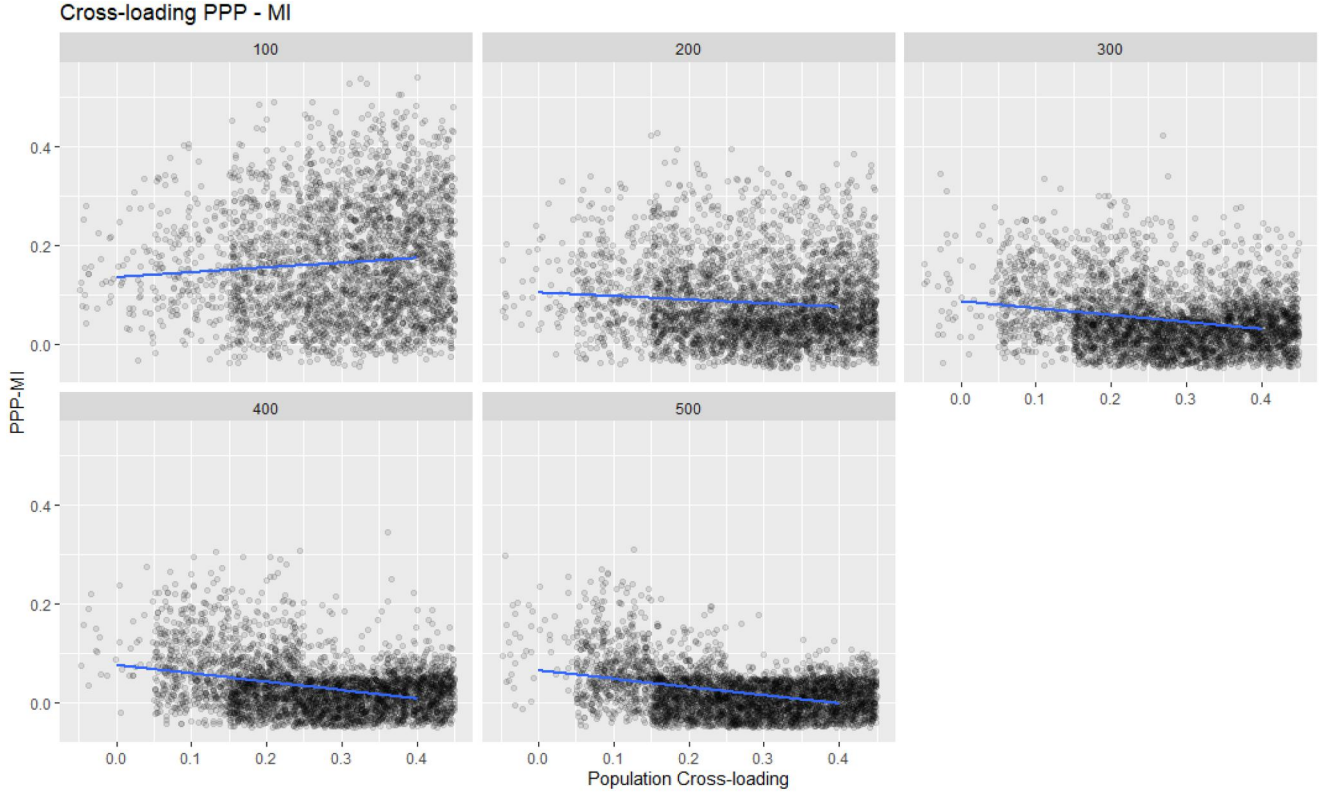

Figure 5. PPP of the cross-loadings MI.

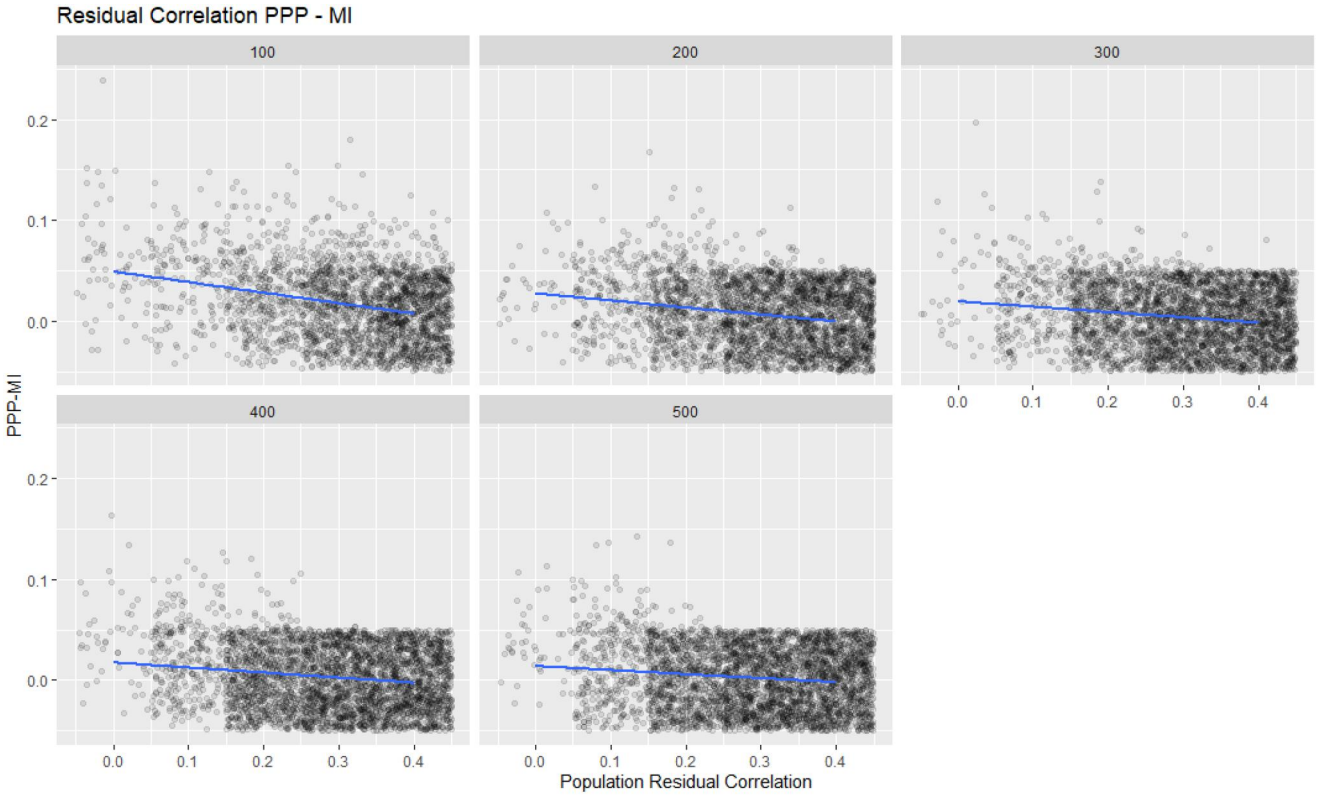

Figure 6. PPP of the residual correlation MI.

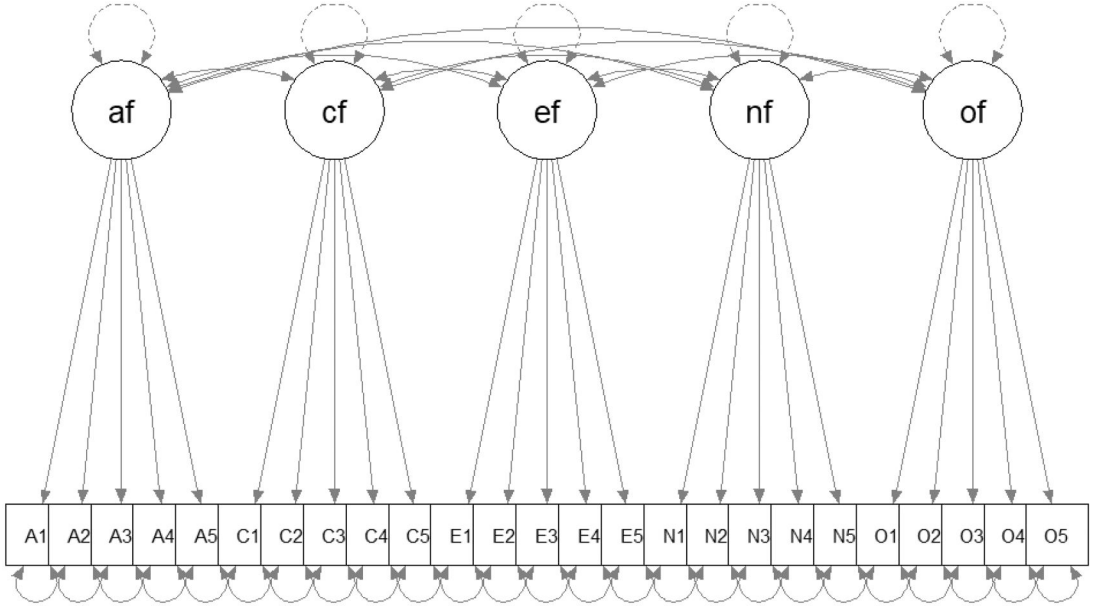

Figure 7. BFI theoretical structure.

Table 6. Approximate global fit indices for the original model.

	Mean	Median	SD	95% L	95% U
RMSEA	0.09	0.09	0.00	0.08	0.09
$\hat{\Gamma}$	0.87	0.87	0.00	0.87	0.88
CFI	0.73	0.73	0.00	0.73	0.74

Table 7. Posterior distribution of MI from the original model.

	Mean	Median	SD	95% L	95% U	PPP
N1∼∼ N2	118.47	119.85	10.62	102.94	135.09	0.00
A1∼∼ A2	45.85	45.79	4.49	38.11	52.03	0.00
nf =∼ C5	41.43	40.94	7.19	30.40	53.63	0.00
ef =∼ N4	41.03	41.04	7.16	28.43	52.55	0.00
of =∼ E3	36.73	36.41	8.02	22.35	48.70	0.00

whether adding a new parameter will substantially improve model fit, whereas the SEPC should be used to indicate whether the effect size of the freed parameter will be substantial. These metrics can be estimated using the `R` package `blavaan`. Researchers may watch a tutorial about this on the following website, https://ecmerkle.github.io/blavaan/articles/mod_indices.html.

Table 8. Posterior distribution of SEPC from the original model.

	Mean	Median	SD	95% L	95% U	PPP
N1∼∼ N2	0.85	0.85	0.06	0.73	0.97	0.00
A1∼∼ A2	0.37	0.37	0.02	0.33	0.41	0.00
ef =∼ A5	0.34	0.33	0.10	0.14	0.52	0.01
of =∼ E3	0.34	0.33	0.04	0.26	0.42	0.00
N3∼∼ N4	0.32	0.32	0.03	0.26	0.36	0.00

Table 9. Approximate global fit indices for the second model.

	Mean	Median	SD	lower	upper
BRMSEA	0.08	0.08	0.00	0.08	0.08
$\hat{\Gamma}$	0.89	0.89	0.00	0.88	0.89
BCFI	0.77	0.77	0.00	0.76	0.77

Table 10. Posterior distribution of MI from the second model.

	Mean	Median	SD	lower	upper	PPP
A1∼∼ A2	46.01	45.87	4.57	37.67	55.33	0.00
nf =∼ C5	43.42	43.60	7.77	28.86	58.87	0.00
of =∼ E3	37.74	36.84	8.28	22.03	52.91	0.00
E3∼∼ O3	31.59	31.49	3.58	24.19	38.16	0.00
ef =∼ N4	30.38	29.94	7.52	15.63	45.24	0.00

Table 11. Posterior distribution of SEPC from the second model.

	Mean	Median	SD	lower	upper	PPP
A1∼∼ A2	0.38	0.38	0.02	0.34	0.42	0.00
of =∼ E3	0.34	0.34	0.04	0.26	0.43	0.00
ef =∼ A5	0.33	0.33	0.09	0.19	0.52	0.02
nf =∼ C5	0.32	0.33	0.03	0.27	0.38	0.00
E3∼∼ O3	0.32	0.32	0.02	0.29	0.35	0.00

The PPMC framework has a potential added benefit of including measures of uncertainty about MIs and SEPCs. Researchers can calculate an uncertainty interval and posterior *SD* from the empirical posterior predictive distribution, allowing researchers to account for uncertainty in their interpretation. Future research is necessary to investigate how uncertainty measures can add value to the standard decision-making process (e.g., should an SEPC's entire interval include nonnegligible effect sizes?) and whether model-modification decisions can improve quality (e.g., fewer incorrectly freed parameters, more correctly flagged parameters to free). Due to the bias of the SEPC, because an omitted parameter increases in magnitude, we do not expect interval estimates to have nominal coverage rates for the true parameters, which would threaten their value for making model-modification decisions. However, the width of an uncertainty interval could still be informative for decision making.

References

Arbuckle, J. L. (2012). *IBM SPSS Amos 21 user's guide*. IBM.

Asparouhov, T., & Muthén, B. (2021). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 1–14. https://doi.org/10.1080/10705511.2020.1764360

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation Modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management*, *41*, 1561–1577. https://doi.org/10.1177/0149206315591075

Bentler, P. (1990). Fit indexes, Lagrange multipliers, constraint changes and incomplete data in structural models. *Multivariate Behavioral Research*, *25*, 163–172. https://doi.org/10.1207/s15327906mbr2502_3

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Inc.

urkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*, 153–157. https://doi.org/10.1080/00031305.1982.10482817

B..

Chou, C.-P., & Bentler, P. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research*, *25*, 115–136. https://doi.org/10.1207/s15327906mbr2501_13

Chou, C.-P., & Bentler, P. (1993). Invariant standardized estimated parameter change for model modification in covariance structure analysis. *Multivariate Behavioral Research*, *28*, 97–110. https://doi.org/10.1207/s15327906mbr2801_6

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*, 147–167. https://doi.org/10.1207/s15327906mbr1802_2

Depaoli, S. (2021). *Bayesian structural equation modeling*. Guilford Press.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*, 509–529. https://doi.org/10.1080/00273170701382864

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*, 46–70. https://doi.org/10.1037/met0000224

Garnier-Villarreal, M., & Little, T. D. (2023). Bayesian longitudinal structural equation modeling. *Longitudinal structural equation modeling (Second)*. Guilford Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–760. http://www.jstor.org/stable/24306036

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Hayduk, L. A., Pazderka-Robinson, H., Cummings, G. G., Levers, M.-J. D., & Beres, M. A. (2005). Structural equation model testing and the quality of natural killer cell activity measurements. *BMC Medical Research Methodology*, *5*, 1–9. https://doi.org/10.1186/1471-2288-5-1

Ximénez, C., Maydeu-Olivares, A., Shi, D., & Revuelta, J. (2022). Assessing cutoff values of SEM fit indices: Advantages of the unbiased SRMR index and its cutoff criterion based on communality. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 368–380. https://doi.org/10.1080/10705511.2021.1992596

Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement*, *78*, 537–568. https://doi.org/10.1177/0013164417709314

Hutchinson, S. R. (1993). Univariate and multivariate specification search indices in covariance structure modeling. *The Journal of Experimental Education*, *61*, 171–181. https://doi.org/10.1080/00220973.1993.9943859

Jeong, C., & Kim, J. (2013). Bayesian multiple structural change-points estimation in time series models with genetic algorithm. *Journal of the Korean Statistical Society*, *42*, 459–468. https://doi.org/10.1016/j.jkss.2013.02.001

Johnson, P. E. (2016). *Portableparallelseeds: Allow replication of simulations on parallel and serial computers* [R package version 0.97].

Jorgensen, T. D., & Garnier-Villarreal, M. (2023). Limited utility of small-variance priors to detect local misspecification in Bayesian structural equation models. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative Psychology. IMPS 2022*. Springer Proceedings in Mathematics & Statistics (pp. 85–95). Springer Nature. https://doi.org/10.1007/978-3-031-27781-8_8

Jorgensen, T. D., Garnier-Villarreal, M., Pornprasertmanit, S., & Lee, J. (2019). (2018). Small-variance priors can prevent detecting important misspecifications in Bayesian confirmatory factor analysis. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: The 83rd annual meeting of the psychometric society* (pp. 255–263). Springer. https://doi.org/10.1007/978-3-030-01310-3_23

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2023). *semTools: Useful tools for structural equation modeling* [R package version 0.5-6.928]. https://CRAN.R-project.org/package=semTools

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*, 69–86. https://doi.org/10.1207/s15327906mbr2301_4

Kaplan, D. (1989). Model Modification in Covariance Structure Analysis: Application of the Expected Parameter Change Statistic. *Multivariate Behavioral Research*, *24*, 285–305. https://doi.org/10.1207/s15327906mbr2403_2

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.

Kwok, O.-M., Luo, W., & West, S. G. (2010). Using modification indexes to detect turning points in longitudinal data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 216–240. https://doi.org/10.1080/10705511003659359

Levy, R. (2011). Bayesian data–model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 663–685. https://doi.org/10.1080/10705511.2011.607723

Luijben, T. C., & Boomsma, A. (1988). Statistical Guidance for Model Modification in Covariance Structure Analysis. In D. Edwards & N. E. Raun (Eds.), *Compstat* (pp. 335–340). Physica-Verlag HD.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107–120. https://doi.org/10.1037/0033-2909.100.1.107

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113–139. https://doi.org/10.1207/S15327906MBR3801_5

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Maiti, S. S., & Mukherjee, B. N. (1990). A note on distributional properties of the Jöreskog–Sörbom fit indices. *Psychometrika*, *55*, 721–726. https://doi.org/10.1007/BF02294619

Mansolf, M., Jorgensen, T. D., & Enders, C. K. (2020). A multiple imputation score test for model modification in structural equation models. *Psychological Methods*, *25*, 393–411. https://doi.org/10.1037/met0000243

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021). Efficient Bayesian structural equation modeling in Stan. *Journal of Statistical Software*, *100*, 1–22. https://doi.org/10.18637/jss.v100.i06

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*, 1–30. https://doi.org/10.18637/jss.v085.i04

Merkle, E. C., & Wang, T. (2018). Bayesian latent variable models for the analysis of experimental psychology data. *Psychonomic Bulletin & Review*, *25*, 256–270. https://doi.org/10.3758/s13423-016-1016-7

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. https://doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

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

Pornprasertmanit, S., Miller, P., Schoemann, A., & Jorgensen, T. D. (2021). *simsem: SIMulated structural equation modeling* [R package version 0.5-16]. https://CRAN.R-project.org/package=simsem

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Revelle, W. (2024). *Psychtools: Tools to accompany the 'psych' package for psychological research* [R package version 2.4.3]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psychTools

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, *17*, 105. https://doi.org/10.2307/271030

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

Silvia, E. S. M., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research*, *23*, 297–326. https://doi.org/10.1207/s15327906mbr2303_2

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 131–161. https://doi.org/10.1080/10705511.2019.1577140

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384. https://doi.org/10.1007/BF02294623

Stan Development Team. (2024). *RStan: The R interface to Stan* [R package version 2.32.5]. https://mc-stan.org/

Steiger, J. H. (1989).). *EzPATH: A supplementary module for SYSTAT and SYSGRAPH*.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors* [Paper presentation]. The Annual Meeting of the Psychometric Society (IMPS), Iowa City. IA.

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

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*, 770. https://doi.org/10.3389/fpsyg.2013.00770

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*, 217–239. https://doi.org/10.1037/met0000100

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2024). Loo: Efficient leave-one-out

cross-validation and Waic for Bayesian models [R package version 2.7.0]. https://mc-stan.org/loo/

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handobook of Structural Equation Modeling* (pp. 209–231). Guilford Press.

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, *80*, 26–44. https://doi.org/10.1080/00220973.2010.531299