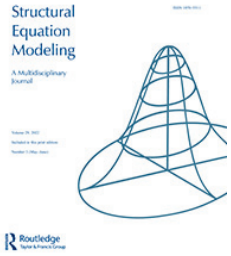
Routledge
Taylor & Francis Group



Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

A Model Comparison Approach to Posterior Predictive Model Checks in Bayesian Confirmatory Factor Analysis

Jihong Zhang, Jonathan Templin & Catherine E. Mintz

To cite this article: Jihong Zhang, Jonathan Templin & Catherine E. Mintz (2022) A Model Comparison Approach to Posterior Predictive Model Checks in Bayesian Confirmatory Factor Analysis, Structural Equation Modeling: A Multidisciplinary Journal, 29:3, 339-349, DOI: 10.1080/10705511.2021.2012682

To link to this article: https://doi.org/10.1080/10705511.2021.2012682

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2022, VOL. 29, NO. 3, 339–349
https://doi.org/10.1080/10705511.2021.2012682

Routledge
Taylor & Francis Group

A Model Comparison Approach to Posterior Predictive Model Checks in Bayesian Confirmatory Factor Analysis

Jihong Zhang, Jonathan Templin, and Catherine E. Mintz

University of Iowa, the American Board of Pediatrics

ABSTRACT

Posterior Predictive Model Checking (PPMC) is frequently used for model fit evaluation in Bayesian Confirmatory Factor Analysis (BCFA). In standard PPMC procedures, model misfit is quantified by comparing the location of an ML-based point estimate to the predictive distribution of a statistic. When the point estimate is far from the center posterior predictive distribution, model fit is poor. Not included in this approach, however, is the variability of the Maximum Likelihood (ML)-based point estimates. We propose a new method of PPMC based on comparing posterior predictive distributions of a hypothesized and saturated BCFA model. The method uses the predictive distribution of the saturated model as a reference and the Kolmogorov-Smirnov (KS) statistic to quantify the local misfit of hypothesized models. The results of the simulation study suggest that the saturated model PPMC approach was an accurate method of determining local model misfit and could be used for model comparison. A real data example is also provided in this study.

KEYWORDS

Bayesian analysis; posterior predictive modeling checking; posterior predictive *p* value; structural equation modeling

Bayesian estimation for structural equation modeling (SEM) is a viable alternative to frequentist SEM approaches (e.g., maximum likelihood), particularly for complex model specifications or for analyses with small sample sizes (e.g., Muthén & Asparouhov, 2012). The popularity of Bayesian approaches in SEM carries with it the need for Bayesian-based model fit investigations to examine global or local model misfits in Bayesian SEM analyses. Common types of SEM model misspecification include the omission of needed latent variables (e.g., Kaplan, 1988), the misspecification of which observed indicator variables measure which latent variables, or violations of latent variable distributional assumptions (e.g., Boomsma, 1987). Posterior predictive model checks (PPMCs) are tools used in Bayesian analyses to help detect model misspecifications by comparing statistics calculated from observed data with model-generated data based on the posterior estimates of parameters (e.g., Gelman et al., 1996; Levy, 2011; Levy et al., 2009). In this paper, we propose and investigate a novel method for investigating model fit in Bayesian CFA (and SEM) using PPMC with Kolmogorov-Smirnov statistic (KS-PPMC) for model comparison.

In PPMC, the value of an observed statistic is compared to the predictive distribution of the same statistic calculated from simulated data sets; these simulated data sets are generated by drawing from the posterior distribution of model parameters. To compare the observed and predictive statistics, the percentile rank of the observed statistic on the posterior predictive distribution, often called the posterior predictive *p* value (PPP value) is computed. The PPP value represents the location of an observed statistic relative to the posterior predicted distribution and is different from a maximum likelihood (ML)-based *p* value. The ML-based *p* value is a quantity derived from the likelihood function or the limiting distribution of model parameters. PPP values close to zero or one (i.e., at the tails of the posterior predictive distribution) typically indicate bad model-data fit. This concept was originated by Rubin and was later extended to include general discrepancies by Gelman et al. (1996).

One concern is that the PPP value may be heavily influenced by the reference point of the observed data, which is often the maximum likelihood estimate (MLE). For instance, when examining the local misfit between a pair of observed indicators in a Bayesian CFA model, the ML estimate of the Pearson correlation is commonly used as the observed statistic in a PPP analysis. In such analyses, the Pearson correlation between a pair of observed indicators is calculated using ML and then the percentile of that quantity is found using the predictive distribution of the correlation. Research on PPP values in latent variable modeling has yielded inconsistent results, in some cases finding Type I error rates to be less than nominal values and in other cases finding Type I error rates at or slightly below nominal values (Levy, 2011). The PPP value can also yield overly conservative results when the asymptotic mean of the test statistic T depends on parameter θ (Robins et al., 2000). Under small sample sizes, the empirical distribution of observed statistics used for calculating PPP values can be large, which may affect the accuracy of PPP value calculation. Additionally, as the observed test statistics are ML-based point estimates, their realized values may depend on asymptotic arguments to be consistent, while a posterior distribution (rather than a point estimate) does not have the same asymptotic dependency, at least in small samples (Levy, 2011).

CONTACT Jihong Zhang jihong-zhang@uiowa.edu Department of Psychological and Quantitative Foundations, University of Iowa, 224B1 Lindquist Center South, Iowa City, IA 52242.

Our proposed PPMC with the Kolmogorov-Smirnov statistic (KS-PPMC) replaces the ML-based point estimate from the observed data with the predictive distribution of a Bayesian-estimated saturated model, resulting in model fit being judged by the comparison between the two distributions. Prior research has investigated the use of the Kolmogorov-Smirnov (KS) statistic and other similar distance statistics for the testing model fit of structural equation models. For example, Wu et al. (2014) used Kullback–Leibler divergence to quantify the discrepancy between two realized posterior predictive distributions. Grønneberg and Foldnes (2019) employed the KS distance between the bootstrap distribution and the theoretical uniform distribution as the selection criterion of model fit indices. Marcoulides et al. (2020) made use of the Anderson-Darling (AD) metric for selecting the best test statistic. However, to the researchers' knowledge, this is the first time the KS-PPMC method has been used as a model comparison approach for Bayesian CFA. A detailed comparison among the performance of varied distance metrics is beyond the scope of the current study. Given the importance of such a comparison for future research, we will return to this issue later in this paper.

In our study, we chose the posterior predictive distribution of test statistics under the saturated model as the reference distribution of KS distance. In many ML-based test statistics, the saturated model is the basis for model fit comparisons globally (e.g., likelihood ratio tests and Root Mean Squared Error of Approximation values) and locally (i.e., standardized, normalized, and unstandardized residual covariances). In our proposed method, we estimate a saturated model with uninformative priors to use as a reference distribution. Instead of using PPP values formed by comparing point estimates of statistics to their respective posterior predictive distributions, our method seeks to quantify a measure of overlap between the posterior predictive distributions of the saturated and specified models. When these posterior predictive distributions have a high degree of overlap, the specified model can be considered to fit the data well. Alternatively, when these posterior predictive distributions show little overlap, model fit of the specified model can be considered poor.

The remainder of the paper defines KS-PPMC and examines its use in both simulation and empirical data analyses. First, we introduce various types of model fit indices implemented in either ML or Bayesian analyses. Next, we present KS-PPMC using the Kolmogorov–Smirnov (KS) statistic to quantify the degree of distributional overlap. To check the accuracy of the proposed measures, we describe the results of a simulation study, in which the performance of our proposed methods is compared with ML-constructed PPP values. We then apply our methods to an empirical example in order to show the performance of the proposed method in a real-world scenario. Finally, we discuss the advantages and the limitations of our new methods along with future extensions of these approaches.

Confirmatory factor models

Confirmatory factor analysis (CFA) models posit that a set of responses by a person p ($p = 1, \ldots, N$) to a set of observed indicator variables i ($i = 1, \ldots, I$), $\boldsymbol{Y}_p = \left[Y_{p1}, \ldots, Y_{pI}\right]^T$, is influenced by the value of a set of $k = 1, \ldots, K$ latent factors $\xi_{\boldsymbol{p}} = \left[\xi_{p1}, \ldots, \xi_{pK}\right]$ via a multivariate linear model:

$$\boldsymbol{Y}_p = \boldsymbol{\mu} + \boldsymbol{\Lambda}\xi_{\boldsymbol{p}} + \boldsymbol{\delta_p}, \tag{1}$$

where $\boldsymbol{\Lambda}$ is an $I \times K$ matrix of factor loadings, $\boldsymbol{\mu}$ is an $I \times 1$ vector of item intercepts, and $\boldsymbol{\delta}$ is an $I \times 1$ vector of item-specific residuals. For model identification, a set constraints are placed on the elements of $\boldsymbol{\Lambda}$ where $\lambda_{ik} = 0$ if item i does not measure factor k (e.g., McDonald, 1999). The residuals are assumed to follow a multivariate normal distribution with zero mean vector and error covariance matrix $\boldsymbol{\Psi}$.

Factors are assumed to follow a multivariate normal distribution, often with mean vector fixed to zero and factor covariance matrix $\boldsymbol{\Phi}$. Additional constraints may be placed on the item intercepts and factor loadings to estimate the factor means and variances, respectively. Coupling the assumed distribution of the factors (with zero mean vector) and item residuals with the linear model in Equation 1 results in the assumption that the data follow a multivariate normal distribution with mean vector equal to item intercepts $\boldsymbol{\mu}$ and covariance matrix given by:

$$\Sigma_0 = \boldsymbol{\Lambda\Phi\Lambda}^T + \boldsymbol{\Psi}. \tag{2}$$

The CFA model puts a specific hypothesized structure on the covariance matrix of the observed indicators. To test this hypothesized structure, the estimates of the hypothesized CFA model (which we label H_0) with covariance matrix Σ_0 are compared to a general, saturated model (which we label H_1) with no constraints on its covariance matrix Σ_1. As all possible CFA models are nested within the saturated model H_1, the comparison of these two models is conducted via mechanisms of nested model comparisons (such as likelihood ratio tests for ML-based analyses), where the CFA model with covariance matrix Σ_0 represents the null or hypothesized model (H_0) and the saturated model with Σ_1 represents the alternative model (H_1). In ML-based analyses, global fit statistics are derived from this comparison including the model Chi-Squared test and the root mean square error of approximation (RMSEA). Moreover, local model misfit is often conducted by inspection of the residuals (i.e., difference between Σ_0 and Σ_1) with raw, standardized, and normalized versions being used, the latter two involving estimates of the elements of unconstrained saturated model covariance matrix Σ_1.

Bayesian confirmatory factor models

In general, Bayesian estimation methods seek to find the posterior distribution of a set of parameters $\theta_{\boldsymbol{H}}$ for a hypothesized model H. This distribution is given by Bayes theorem in Equation 3:

$$\boldsymbol{p}(\theta_{\boldsymbol{H}}|\boldsymbol{Y}) \propto \boldsymbol{p}(\boldsymbol{Y}|\theta_{\boldsymbol{H}})\boldsymbol{p}(\theta_{\boldsymbol{H}}), \tag{3}$$

where $p(\theta_H)$ is the prior distribution of the parameters and $p(\boldsymbol{Y}|\theta_H)$ is the model likelihood function of the data given the parameters. In Bayesian CFA models, a conditional approach to estimation is frequently implemented where $p(\boldsymbol{Y}|\theta_H)$ is given by the multivariate normal density with mean vector equal to $\boldsymbol{\mu} + \boldsymbol{\Lambda\Xi}$ and covariance matrix $\boldsymbol{\Psi}$. In such conditional models, an additional step is needed to specify the likelihood of the unobserved factors $\boldsymbol{\Xi}$, which is specified as multivariate normal with zero mean vector and covariance matrix $\boldsymbol{\Phi}$.

To estimate the Bayesian CFA model, prior distributions are specified for each type of parameter, specifically, the item intercepts $\boldsymbol{\mu}$, the factor loadings $\boldsymbol{\Lambda}$, the unique variances $\boldsymbol{\Psi}$, and the factor variances and covariances in $\boldsymbol{\Phi}$. The latent variables Ξ are treated as parameters that have a prior distribution equal to their assumed factor distribution, which is multivariate normal with mean zero and covariance matrix $\boldsymbol{\Phi}$. Although prior distributions can vary for each type of parameter, often the item intercepts and factor loadings follow normal distributions (which are conjugate priors, enabling direct sampling from the posterior distribution) whereas the unique variances follow inverse gamma prior distributions (also conjugate priors). A conjugate prior for the factor covariance matrix is the inverse Wishart distribution.

Bayesian saturated models

In Bayesian statistics, the posterior distribution is obtained for all parameters. The Bayesian version of the saturated model (H_1) includes a set of prior distributions for the elements of the mean vector and elements of the covariance matrix Σ_1, which results in the construction of a posterior distribution for each unique parameter in the model. Herein lies a critical difference between ML-based and Bayesian methods of estimating the saturated model H_1: ML-based methods use the ML-based point estimate for all parameters of the saturated model H_1, whereas the Bayesian analog of the saturated model H_1 necessarily has a posterior distribution for all parameters. Although the posterior distribution will converge in distribution to the ML asymptotic distribution as the sample size goes to infinity, wide variability may exist for cases where sample sizes are small relative to the number of parameters.

The key feature of ML-based PPP values in this study seeks to investigate is how to incorporate the variability of the saturated model's posterior distributions into the model fit process. That is, when the saturated model H_1 (sample) means, variances, and covariances yield posterior distributions rather than point estimates, how do model fit indices change? Moreover, does variability in the posterior distribution need to be accounted for when evaluating the model fit of the specified model H_0?

Evaluation of Bayesian CFA model fit

To illustrate Bayesian Structural Equation Modeling (BSEM) and Bayesian CFA fit indices, Levy (2011) distinguished two separate types of Bayesian model fit evaluations in terms of their target measures: test statistics and discrepancy measures. The first approach focuses on the extent to which the model recovers or predicts features of the data, whereas the second aims to explicitly build in the comparison between the observed data and model-implied data characteristics. Our proposed KS-PPMC method employs the item-pair Pearson correlation as the foundation of a KS-based discrepancy measure that indicates item pairs with a local model misfit.

An advantage of Bayesian SEM fit indices over most of their frequentist counterparts is that the posterior distribution allows uncertainty to be quantified for any index. Despite different processes, Bayesian and ML-based model fit methods have shown to have similar rejection rates when sample sizes are large. For example, Garnier-Villarreal and Jorgensen (2019) compared the chi-square-based approximate fit indices that are commonly used in SEM to their Bayesian analogs through a simulation study and concluded that Markov Chain Monte Carlo (MCMC) with noninformative priors yields similar results to ML across varied levels of misspecification, sample sizes, and model types.

Model comparison posterior predictive model checking

Common Bayesian PPMC methods work to find the posterior predictive distribution conditional of the parameters of the specified model $p\left(Y_H^{rep}|\theta_H\right)$:

$$\begin{aligned} \boldsymbol{p}(Y^{\boldsymbol{rep}}|\theta_{\boldsymbol{H}}) &= \int_{\theta_H} \boldsymbol{p}(Y^{\boldsymbol{rep}}|\theta_{\boldsymbol{H}})\boldsymbol{p}(\theta_{\boldsymbol{H}}|Y^{\boldsymbol{obs}})\boldsymbol{d}\theta_{\boldsymbol{H}} \\ &\propto \int_{\theta_H} \boldsymbol{p}(Y^{\boldsymbol{rep}}|\theta_{\boldsymbol{H}})\boldsymbol{p}(\boldsymbol{Y}^{\boldsymbol{obs}}|\theta_{\boldsymbol{H}})\boldsymbol{p}(\theta_{\boldsymbol{H}})\boldsymbol{d}\theta_{\boldsymbol{H}}. \end{aligned} \tag{4}$$

Equation 4 shows the general form of the posterior predictive distribution, $p(\boldsymbol{Y}^{rep}|\theta_{\boldsymbol{H}})$, which is the integral of two components: the sampling distribution of the replicated data given the sampled values from the posterior distribution of parameters under model H, $p(\boldsymbol{Y}^{rep}|\theta_{\boldsymbol{H}})$, and the posterior distribution of parameter under model H, $p(\theta_{\mathrm{H}}|\boldsymbol{Y}^{\mathrm{obs}})$.

In practice, PPMC methods are implemented by generating predictive data based on the posterior distribution of estimates. To provide context, consider an example where a Bayesian CFA model of Equation 1 is estimated via MCMC. First, a standard MCMC estimation algorithm is run (with specifications of prior distributions, number of Markov chains, number of iterations, burn-in, etc.). Once the chains have been estimated and chain convergence is established, then a sample of parameters $(\theta_{\boldsymbol{H}})$ are drawn with replacement from the set of iterations of the Markov chains. For each sampled set of parameters, a set of data $(\boldsymbol{Y}^{rep})$ with sample size equal to that of the observed data are simulated by plugging the sampled parameters into the model H. Then, the test statistics $T(\boldsymbol{Y}^{rep})$ are calculated from the newly generated data. In our case, $T(\cdot)$ will be the Pearson correlation coefficients calculated for each pair of observed indicators. Across all replication samples of the posterior distribution of parameters, $T(\boldsymbol{Y}^{rep})$ is calculated, yielding, for each pair of observed indicators, the predictive distribution of the statistic. The principle of PPMC is to locate the position of the observed data statistic $\mathrm{T}(\boldsymbol{Y}^{\mathrm{obs}})$ in the posterior predictive distribution $p(\boldsymbol{Y}^{rep}|\theta_H)$. The key difference between KS-PPMC and PPMC with classical ML-based fit indices lies in the choice of the reference distribution (e.g., Lee et al., 2016). Under standard

PPMC methods, the reference distribution is the posterior predictive distribution, to which the often ML-based observed data statistic $T(\boldsymbol{Y}^{obs})$ is compared.

As standard PPMC methods typically use ML-based observed data statistics $T(\boldsymbol{Y}^{obs})$, they do not incorporate the uncertainty of the observed data statistics into the process. Such uncertainty of observed data statistics may come from sampling error which may especially be prevalent in situations where there are numerous cases of missing data or small sample sizes are present. Asparouhov and Muthén (2020) recently proposed a new approach by using the parameters of the saturated model (H_1) to generate the posterior predictive distribution, which could reduce the rejection error rate in such situations.

The principal motivation of this study is to replace the point estimate of realized data statistics, $T(Y^{obs})$, with the posterior predictive distribution of realized data statistics, $T(Y^{rep}|\theta_H)$. Then, we employ the KS statistic to quantify the distance between the posterior predictive distribution of realized data statistics and that of the saturated model. In our study, the KS-PPMC statistic for the cumulative distribution function under the saturated model is

$$KS_{PPMC} = \mathbf{sup}_T |F_n(\boldsymbol{Y}^{rep}|\theta_H) - F(\boldsymbol{Y}^{rep}|\theta_{H_1})| \quad (5)$$

where $\mathbf{sup}_T$ is the supremum of the set of distances. $F_n(\boldsymbol{Y}^{rep}|\theta_H)$ denotes the cumulative posterior predictive distribution of $T(Y^{rep}|\theta_H)$ under model H and $F_n(\boldsymbol{Y}^{rep}|\theta_{H_1})$ denotes the cumulative posterior predictive distribution of $T(\boldsymbol{Y}^{rep}|\theta_{H1})$ under the saturated model. The statistic computes the largest absolute difference between the two distribution functions across all realized data statistics. By the Glivenko–Cantelli theorem, if the sample comes from distribution under the saturated model, then KS_{PPMC} converges to zero almost surely in the limit when N goes to infinity. This approach not only depicts the discrepancy between the observed data with reference distribution under the saturated model H_1 but also the degree of uncertainty in the observed data statistics.

Proposed model comparison PPMC procedure

To get the posterior distribution of the covariance matrix $\boldsymbol{\Sigma}_1$ from the saturated model H_1, the first step is to use a Bayesian algorithm to estimate the saturated model using the observed data. Choices of prior distributions for the saturated model are critical as overly strict priors may result in saturated model posterior distributions far from what the data may suggest, which may cause bias in the model fit analysis. For our study, we model the observed data using a multivariate normal distribution, estimating each unique element of the mean vector and covariance matrix without constraints. For prior distributions, we specify a diffuse, uninformative prior of multivariate normal distribution for the mean vector with zero mean vector, zero-off diagonal elements of the prior covariance matrix and variances set to 100,000. For the saturated model covariance matrix $\boldsymbol{\Sigma}_1$, we also specify a diffuse, uninformative prior using an inverse Wishart prior distribution with parameters $\boldsymbol{\Psi}$ with zero-off diagonal elements of item variances and degree of freedom ν equals number of indicator variables.

Following estimation and successful convergence of the saturated model, the posterior predictive distributions of each of the means and covariances of the saturated model are formed using the typical PPMC process of sampling draws from the posterior distribution, using those parameters to generate simulated data Y^{rep}, and calculating the Pearson correlation to every item pair, forming $T(Y^{rep})$. We then quantify the distance between the alternative posterior predictive distribution with the reference posterior predictive distribution considered as a fully Bayesian analog of a traditional p value.

If the model is consistent with the population that generated the observed data, then the posterior predictive distributions should have considerable overlap. A nonparametric test of the equality of probability distributions, the Kolmogorov-Smirnov statistic (KS) is used to assess the distance between the current model with the saturated model. The PPMC with KS statistic is the maximum difference between the cumulative densities of the posterior predictive distribution of the specified model (H_0) and the saturated model H_1 across the space of the test statistic (Pearson correlation). A PPMC with KS value is obtained for each pair of observed indicator variables. Next, we test our new PPMC methods via a simulation study.

Monte Carlo simulation study

In this section, we report results from a simulation study designed to investigate the performance of KS-PPMC. Our study borrows simulation specifics from Hoofs et al. (2018). Data were generated using either one or two latent variables. For data generated with one latent variable, the correct model (the one-factor model) was then estimated and compared with an overspecified model (a two-factor model where equal numbers of items loaded onto both factors) as well as the saturated model using the KS-PPMC statistic. When data were generated based on a two-factor model, the correct model (a two-factor model) was then tested against one underspecified model (a one-factor model), two incorrectly specified models, and the saturated model (see Figure 1).

Data generation methods

The simulated data sets were generated based on three main experimental factors: (1) number of latent variables (i.e., one-factor structure—Model A0 and two-factor structure—Model B0; see Figure 1), (2) number of observed indicators (6 items or 12 items), and (3) sample size (25, 500, and 2,000) for a total of 12 conditions.

For the one-factor model, to mimic real data, the factor loadings were fixed to 0.4, 0.6, and 0.8. When 6-item tests were generated with a one-factor structure (Model A0), the factor loadings for all items were set as follows $\lambda_1 = \lambda_2 = 0.4$, $\lambda_3 = \lambda_4 = 0.6$, $\lambda_5 = \lambda_6 = 0.8$. The factor variance was set to 1. The residual variances of the indicators were set as follows $\psi_1 = \psi_2 = 0.84$, $\psi_3 = \psi_4 = 0.64$, $\psi_5 = \psi_6 = 0.36$. These values were picked to achieve observed indicator variables with varying levels of information about the latent trait. Similarly, when the population data matrix was generated based on 12 items and one factor (not shown in Figure 1),

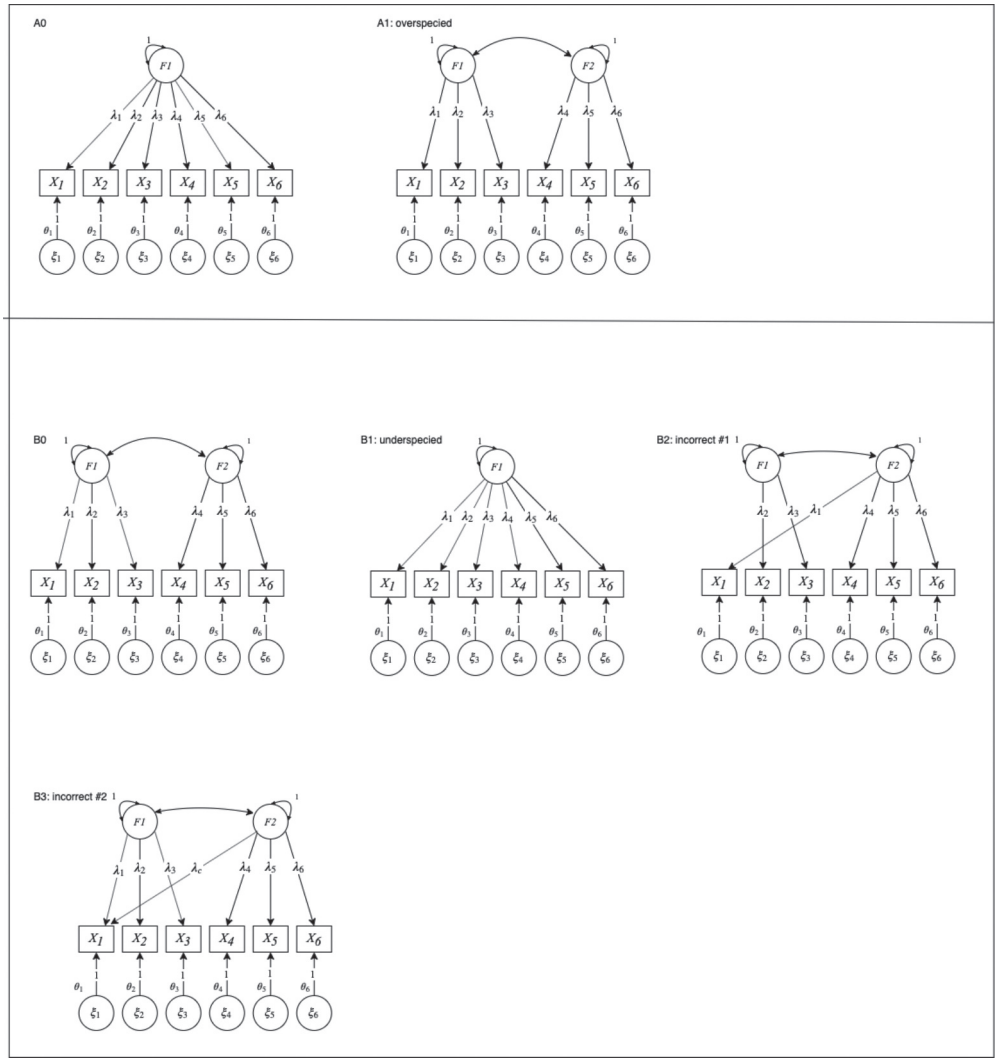


Figure 1. Simulation design: different models.

the factor loadings were set so that the first four items had a loading of 0.4, the second four items had a loading of 0.6, and the last four items had a loading of 0.8. Both factor variances were set to 1 while the factor covariance was set to $\phi_{12} = 0.4$. The residual variances of the indicators were set as $\psi = \{0.84, 0.64, 0.36\}$.

For the two-factor model, when data were generated with 6 items (see Figure 1 Model B0), the factor loadings were set as $\lambda_1 = \lambda_2 = \lambda_3 = 0.4, \lambda_4 = \lambda_5 = \lambda_6 = 0.8$. When data were generated with 12 items, the factor loadings were set as $\lambda_1 = \ldots = \lambda_3 = 0.4$, $\lambda_4 = \ldots = \lambda_6 = 0.8$, $\lambda_7 = \ldots = \lambda_9 = 0.4$, and $\lambda_{10} = \ldots = \lambda_{12} = 0.8$. There were no cross-loadings for any models used for simulating data. For all two-factor models, the factor covariance was set to $\phi_{12} = 0.4$. The residual variances of the indicators were set as $\psi_1 = \psi_2 = \psi_3 = 0.84$ for items 1 to 3, $\psi_4 = \psi_5 = \psi_6 = 0.64$ for items 4 to 6, and $\psi_7 = \psi_8 = \psi_9 = 0.35$ for items 7 to 9.

The item intercepts $\mu = (\mu_1, \ldots, \mu_i)^T$ for all generated data were fixed to zero. Item response data were generated using the CFA model given by Equation 1. The Bayesian estimation process used in each condition is explained in detail in the next section.

Simulation design

In this section, we show the factor structure of the misspecified model and the choices for prior distributions used in this study. In the first condition (data generated with a one-factor model), we analyzed the data with the saturated model and the misspecified two-factor model (Figure 1, Model B1: half of the items load onto the first factor and half of the items load onto the second factor). In the second condition, we estimated the model with three types of misspecification and the true model for comparison. As shown in Figure 1, the incorrect model B1 has one latent factor. The incorrect model B2 has one incorrect loading λ_1 for item 1. The incorrect model B3 has the correct number of dimensions but has one item with an additional, unnecessary factor loading, λ_c.

Each condition was replicated 100 times. All models were estimated using MCMC estimation via JAGS (Plummer, 2003) with uninformative priors. Specifically, we set the prior distribution of factor loadings using a normal distribution N($\mu = 0$, $\boldsymbol{\sigma} = 1$). The prior distribution of item means was set to be a normal distribution with mean zero and variance of 100,000; the unique variances were sampled from a gamma

prior distribution with alpha of .5 and beta of .059; the factor variance matrix in a three-factor model was sampled from the inverse Wishart distribution with Ψ as an identity matrix and three degrees of freedom. Each MCMC analysis had four estimation chains with 5,000 iterations of which 2,000 iterations were discarded as a burn-in phase. Following the analysis, 1,000 sets of parameters were then randomly drawn from the posterior distribution of MCMC estimates to generate posterior predictive data sets. To examine local misfit, for each pair of observed indicators, the KS-PPMC with KS statistic and traditional PPP values were calculated.

In order to compare with Bayesian estimation, we also fit the models using maximum likelihood estimation using the *Lavaan* package (Rosseel, 2012) in R version 4.0 (R Core Team, 2020).

Results

Global model fit

We first checked the global model fit to investigate the overall model-data misfit. Table 1 shows the average values and the rejection rates of four global model fit indices (SRMR, CFI, TLI, and RMSEA) across all 18 conditions (6 models by 3 sample sizes). The results suggested that when the true model had a one-factor structure, model A1 had lower SRMR/RMSEA, higher average CFI/TLI, and lower rejection rates than model A0, which means traditional global model fit indices were insensitive to over specification. Similar to that, when the true model had a two-factor structure, model B0 and model B3 had lower average SRMR/RMSEA and higher average CFI/TLI than models B1 and B2. It should be noted that model B3 had almost the same global model fit as Model B0, which is not surprising as model B3 had only one more cross-loading than model B0.

As for the influence of sample sizes, as sample size increases, all fit indices have more power to detect model misfit. Both models A0 and A1 had good model fit when sample sizes were larger than 25. Models B1 and B2 had poor model fit uniformly, even when the sample size was 2,000. Models B0 and B3 had acceptable model fit when the sample size was larger than 50.

KS-PPMC and PPP values

Since KS-PPMC statistics and PPP values have distinct criteria for misfit (KS-PPMC with KS measures near zero or PPP values near 0.5 indicate good fit), we transformed the PPP values to an absolute PPP (PPP*; Equation 6) so that lower absolute PPP values suggest better local fit.

$$\mathrm{PPP}^{*} = 2 * |PPP - 0.5| \tag{6}$$

When the KS-PPMC statistic of an item pair correlation is close to zero, the posterior predictive distribution of the discrepancy measure in the alternative model completely overlaps with that of the measure in the saturated model, meaning near perfect model-data fit. Similarly, if the KS-PPMC and PPP* values are close to one, local misfit is present. Figure 2 shows how KS-PPMC (red) and the PPP$*$ values (blue) perform differently for the same six-item test. The left-hand side panel shows the bar plot of PPP$*$ values and KS-PPMC statistics across all 100 replications for item pairs 1&2, 1&3, 1&5, and 4&5 (Model A0). The right-hand side panel of Figure 2 shows the results in the two-factor solution (Model A1). The results suggest that the PPP$*$ values and KS-PPMC have similar trends when the sample size increases but KS-PPMC values have lower variances than PPP$*$ values. Both indices also suggest different local model-data fit between two solutions. To be more specific, for item pair correlations between indicators both loading onto the same factor (i.e., indicators 1 and 2, indicators 1 and 3), the KS-PPMC and the PPP$*$ values are not affected by the small sample size in the one-factor solution (correct model, Model A0). In contrast, in the two-factor solution (Model A1), the KS-PPMC and the PPP$*$ values increase for some correlations (i.e., indicators 1 and 2, indicators 1 and 3) as the sample size gets larger. For instance, when sample size is 2,000, 6 out of 15 correlations in the overspecified model have higher PPP$*$ values than in the correct model; 12 out of 15

Table 1. Comparing global fit indices: average values and rejection rates for ML estimators.

Generated Model	Model	N	SRMR	CFI	TLI	RMSEA
One-factor structure	A0	50	.033/.060	.989/.020	.991/.055	.046/.410
		500	.010/.000	.999/.000	1.000/.000	.011/.000
		2000	.005/.000	1.000/.000	1.000/.000	.006/.000
	A1	50	.032/.045	.990/.020	.992/.060	.045/.390
		500	.010/.000	.999/.000	1.000/.000	.010/.000
		2000	.005/.000	1.000/.000	1.000/.000	.006/.000
Two-factor structure	B0	50	.065/.885	.965/.250	.962/.335	.063/.640
		500	.020/.000	.999/.000	1.000/.000	.007/.000
		2000	.010/.000	1.000/.000	1.000/.000	.003/.000
	B1	50	.197/1.000	.635/1.000	.578/1.000	.235/1.000
		500	.188/1.000	.664/1.000	.612/1.000	.221/1.000
		2000	.176/1.000	.675/1.000	.625/1.000	.215/1.000
	B2	50	.115/.995	.922/.760	.909/.805	.107/.950
		500	.095/.975	.955/.280	.946/.370	.080/.575
		2000	.091/.932	.956/.304	.947/.372	.078/.595
	B3	50	.063/.830	.965/.250	.962/.340	.063/.645
		500	.020/.000	.999/.000	1.000/.000	.007/.000
		2000	.010/.000	1.000/.000	1.000/.000	.003/.000

The values before the slash represent the average model fit; the values after the slash represent the proportion of models having unacceptable model fit among all repetitions.

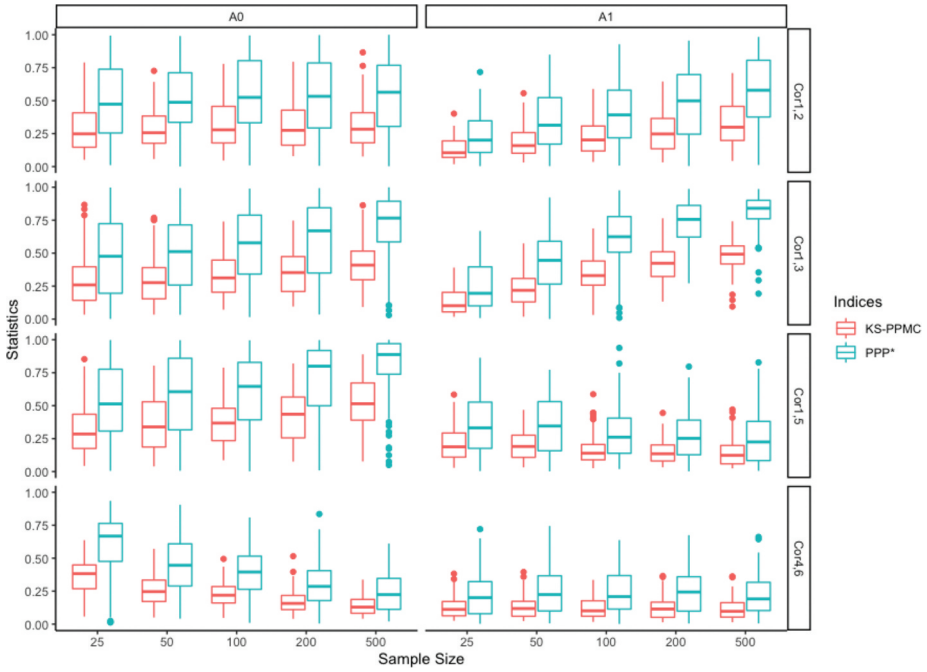


Figure 2. Transformed PPP values and KS-PPMC statistics in condition of 6-item with one-factor structure.

correlations in the overspecified model have higher KS-PPMC statistics than in the correct model. For $N = 25$, only two correlations are flagged as higher PPP∗ values in the overspecified model than in the correct model whereas five correlations have higher KS-PPMC statistics.

The results are consistent for the two-factor condition (Model B0) and other factor structures. Figure 3 shows the KS-PPMC and PPP∗ for six-item pair correlations (indicators 1 and 10, 1 and 4, 1 and 6, 10 and 11, 6 and 7, and 9 and 10) with Models B0, B1, B2, and B3 specifications. Here, PPP∗ values (blue) across all sample size conditions have similar patterns with the KS-PPMC statistics (red) but have higher variances than the other indices. Even for small sample sizes ($N = 25$), the PPP∗ values and the KS-PPMC statistics for Model B0 show good performance. For comparison, the underspecified model (Model B1) has relatively higher PPP∗ values and KS-PPMC statistics for item pairs with two observed indicators loading onto different factors. The incorrect model with one observed indicator loading onto the wrong factor (Model B2) also has large PPP∗ values and KS-PPMC statistics for all pairs that including observed indicator 1. Similarly, Model B3, the incorrect model with observed indicator 1 cross-loading on two factors, has relatively larger PPP∗ values and KS-PPMC statistics (see Rows 1 to 3, Column 4) than other models.

Even though the PPP∗ values and KS-PPMC statistics have similar patterns, there are some differences in their sensitivity. For example, in Model B2 (overspecified model), for $N = 2000$, the PPP∗ values of observed indicator pairs 1 with 10, 11, and 12 are close to their upper threshold which indicates overestimation when sample sizes are larger than 500 (see Rows 1 to 3, Column 3). However, KS-PPMC values for the same indicator pairs did not reach the upper threshold, which allows researchers to compare item pairs with worse model fit. KS-PPMC statistics also have relatively lower variances across all repetitions. Additionally, comparing PPP∗ values to KS-PPMC statistics reveal that standard PPP values larger than .95 or lower than .05 corresponds to KS-PPMC statistics larger than .5. Thus, .5 may be a fair cut score for the KS-PPMC approach.

Empirical data analysis

Method

In this section, we illustrate how the proposed PPMC approach could be used to obtain better model fit and select better model when using Bayesian confirmatory factor analysis. This section does not provide, however, a comprehensive overview of an actual Bayesian CFA. The goal of the empirical illustration was to demonstrate how the researchers could detect the local misfit of BCFA models or compare the models when multiple alternative models exist.

Data from Holzinger and Swineford (1939) were used. Test scores on 26 different measures were obtained from a total of 300 7th and 8th grade students in two schools. The Holzinger and Swineford (1939) data have been used as a model data set by many researchers. For example, Muthén and Asparouhov (2012) used the factor loading pattern of the four-factor model as shown in Table 2 . To be specific, 19 out of 26 items were intended to measure four correlated latent factors: (1) spatial (η_1) measured by visual perception, cubes, paper form board, and flags ($x_1 - x_4$), (2) verbal (η_2) measured by general information, paragraph comprehension, sentence completion, word classification, and word meaning ($x_5 - x_9$), (3) speed (η_3) measured by addition, code, counting groups of dots, and straight and curved capitals ($x_{10} - x_{13}$), and (4) memory (η_4) measured by word recognition, number recognition, figure word ($x_{14} - x_{19}$). To illustrate the performance of the proposed method, two models are estimated: a one-factor model and

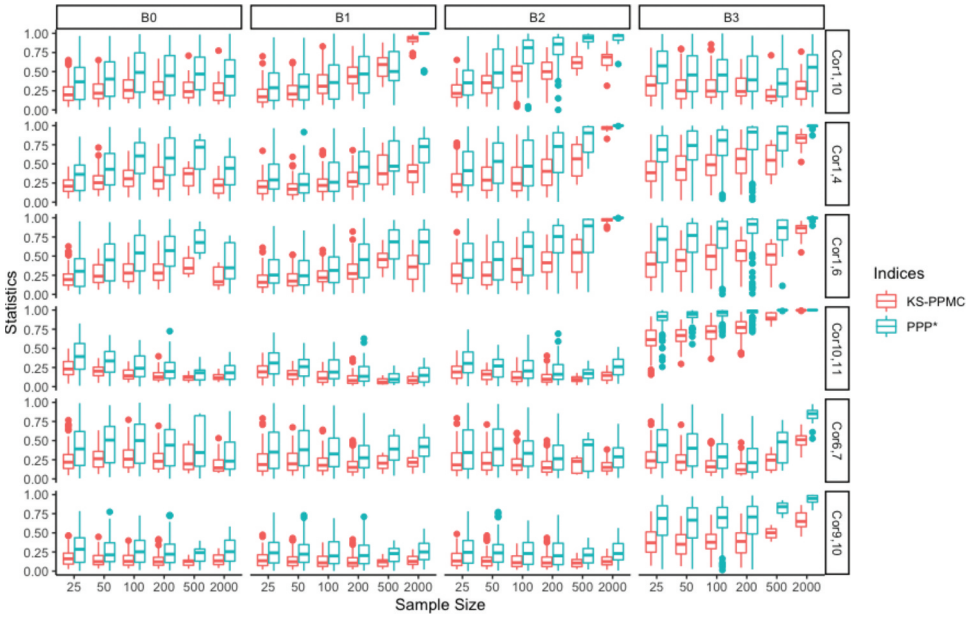


Figure 3. Transformed PPP values and KS-PPMC statistics: models B0, B1, B2, B3.

a four-factor model (Asparouhov & Muthén, 2020). The detailed specification for the four-factor solution is shown in Table 2. All test scores were standardized before the analysis.

The data were estimated with one-factor structure and four-factor structure using BCFA via MCMC estimation. The process used four chains with 10,000 iterations each, of which 2,000 were discarded as burn-in. The prior settings were as follows: item intercepts were normally distributed with mean 0 and variance 1; factor loadings were normally distributed with mean 0 and variance 1. For model identification, both the one-factor solution and four-factor solution employed the marker-item method, which means the factor loadings of the first item per factor were fixed to 1 while other factor loadings were freely estimated. After the posterior distributions for all parameters were estimated, 5,000 parameters from the posterior were randomly sampled and used to create the posterior predictive distribution for the 271 item-pair correlations.

Table 2. Factor structure of the Holzinger-Swineford example: four-factor solution.

	Spatial	Verbal	Speed	Memory
visual	X	0	0	0
cubes	X	0	0	0
paper	X	0	0	0
flags	X	0	0	0
general	0	X	0	0
paragrap	0	X	0	0
sentence	0	X	0	0
wordc	0	X	0	0
wordm	0	X	0	0
addition	0	0	X	0
code	0	0	X	0
counting	0	0	X	0
straight	0	0	X	0
wordr	0	0	0	X
numberr	0	0	0	X
figurer	0	0	0	X
object	0	0	0	X
numberf	0	0	0	X
figurew	0	0	0	X

Results

Using the Gelman-Rubin convergence diagnostic, both the one-factor solution ($R \leq 1.004$) and four-factor solution ($R \leq 1.033$) achieved convergence. Figure 4 presents the distribution of KS-PPMC statistics and PPP$*$ values across all item-pair correlations with the one-factor solution and the four-factor solution. As shown in the boxplot, the average KS-PPMC statistics were higher in the one-factor model ($\mu_{KS-PPMC} = .446$) than in the four-factor model ($\mu_{KS-PPMC} = .310$), which indicates worse local model fit in the one-factor solution than in the four-factor solution. In addition, the range and standard deviation of the KS-PPMC statistics in the one-factor model ($sd = .319$) were wider than in the four-factor model ($sd = .198$), which indicated the higher variation of KS-PPMC statistics in the one-factor model. From the one-factor model results, the highest KS-PPMC statistic was found in the correlation between indicator pair 10 and 12 (KS-PPMC = .999), while in the four-factor model, the highest KS statistics was found in the correlation between indicator pair 2 and 10 (KS-PPMC = .837). These results suggest that, according to the KS-PPMC results, the four-factor model fixed much of the local misfit in the one-factor model.

For comparison, Figure 5 shows the distribution of PPP values across the whole item-pair correlations with the one-factor and four-factor models. Here, similar to the KS-PPMC statistics, the average *PPP* value for the four-factor model was closer to .5 ($\bar{PPP} = .525$) than the one-factor model ($\bar{PPP} = .699$). However, the range and standard deviation for the one-factor ($sd = .297$) and four-factor models ($sd = .302$) were very similar. In addition, in the one-factor model, 25 item-pair correlations yielded *PPP* values equal to one. One of the problematic item-pair correlations included the correlation between items 10 and 12, which also had the highest KS-PPMC statistic. Indicator pair 2 and 10 had the lowest *PPP*

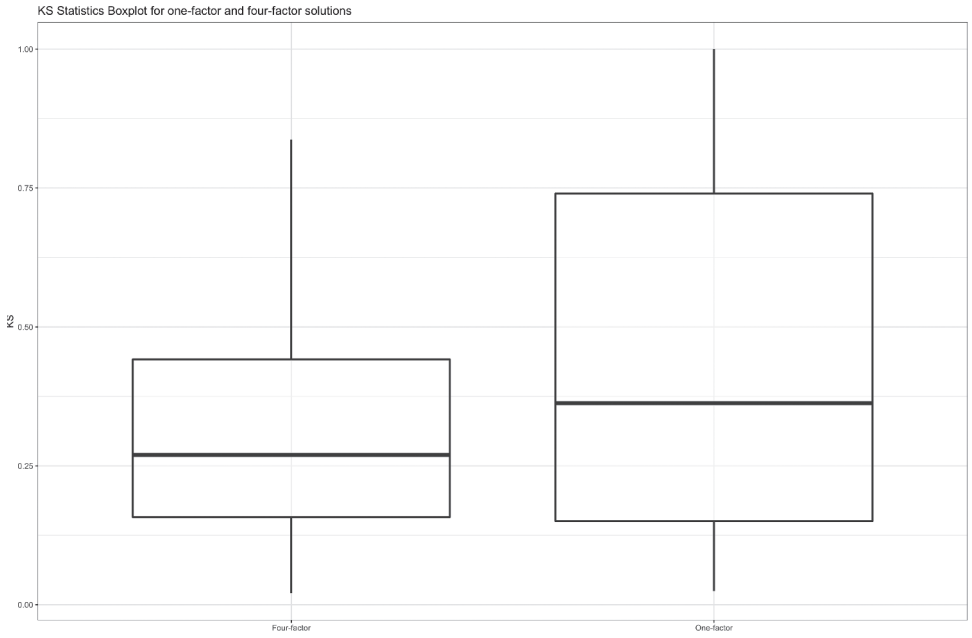


Figure 4. KS statistics boxplot for one-factor and four-factor solutions.

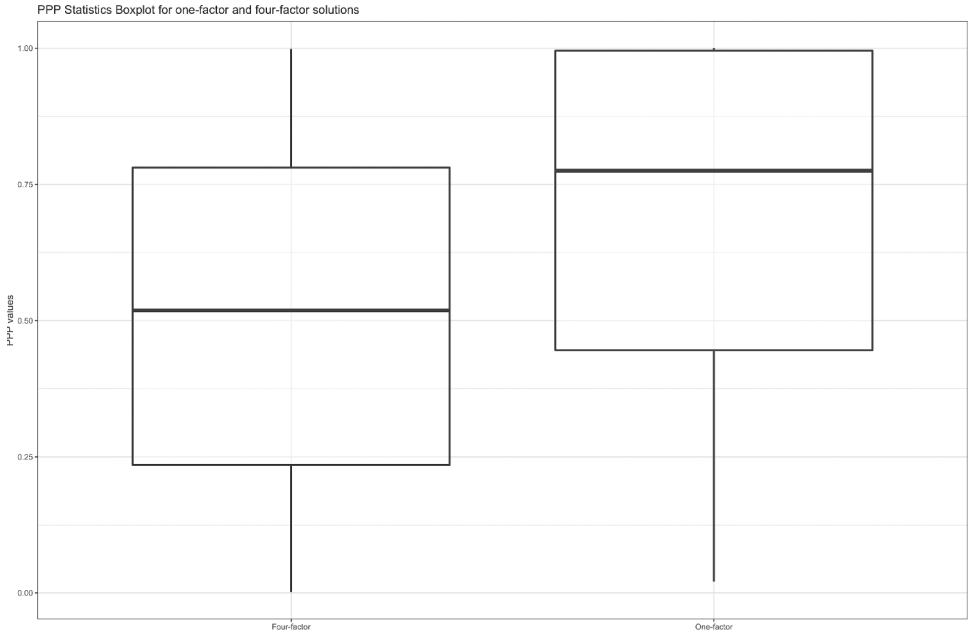


Figure 5. PPP statistics boxplot for one-factor and four-factor solutions.

value ($P\bar{P}P = .021$) in the one-factor model. For the four-factor model, indicator pair 2 and 10 had the lowest *PPP* value and indicator pair 8 and 10 had the highest *PPP* value.

Discussion

In this study, we proposed a model comparison approach to model checking in Bayesian CFA. In our investigation, we showed acceptable sensitivity of PPP values for three forms of misspecification (underspecified, overspecified, and wrongly specified) when the sample size was moderately large (i.e., $N = 500$), in accordance with previous studies (e.g., Hoofs et al., 2018). However, similar to the findings of previous research, PPP values were insensitive to small samples combined with an overspecified model. Our simulation study showed that the PPMC using KS statistics could be an alternative way of detecting local misfit in Bayesian CFA. For large sample sizes, KS-PPMC showed similar patterns with PPP$*$ values. When sample sizes are small, more indicator correlations showed higher KS-PPMC statistics in the overspecified model than those in the correct model; for comparison, less than half item correlations show higher PPP$*$ values.

In addition, we did not find that KS distance metrics are insensitive to the discrepancy between analyzed distributions, as shown in prior studies using re-sampling methods (e.g.,

Grønneberg & Foldnes, 2019; Marcoulides et al., 2020). Some methodological differences may explain the inconsistency. The posterior predictive distributions of the test statistic in our study are possibly more robust than the bootstrapping distribution or the empirical posterior predictive under saturated model are more sensitive to misfit compared to the theoretical uniform distribution used in the previous study. Further research is necessary to understand the interaction effects between KS statistic with varied model checking approaches (e.g., PPMC vs. bootstrap resampling).

PPMC using saturated model vs. PPP value

In summary, there are several similarities between KS-PPMC statistics and PPP values in this data analysis. First, both KS-PPMC statistics and PPP methods suggest that more local misfit exists in one-factor models than in four-factor models, which is consistent with previous research (Asparouhov & Muthén, 2020). Second, both KS-PPMC statistics and PPP values identified indicator pair 10 with 12 in the one-factor model and indicator pair 2 with 10 in the four-factor model as having the greatest amount of local misfit. However, there were also some differences between these two approaches. Using PPP values, some indicator pair correlations may reach the maximum of one (i.e., 25 item pair correlations have a PPP value 1 in the one-factor model). However, the KS-PPMC statistics were never as extreme, which makes model comparison possible.

The PPMC approach using a saturated model could be a very useful tool for detecting local misfit in a fully Bayesian framework. The underlying idea of comparing a saturated model to an alternative model is consistent with the ML-based model fit (i.e., RMSEA and SRMR). The only difference is that the model-data fit is represented by the overlap between the posterior predictive distribution of test statistics rather than a chi-square difference.

Both the PPP approach and KS-PPMC statistics can be good ways for checking local misfit in Bayesian CFA. When thefit is poor, KS-PPMC statistics may be more informative than PPP as KS-PPMC statistics never reach extreme values. In summary, KS-PPMC statistics could be a supplementary approach for PPP methods for checking local misfit in Bayesian CFA.

Limitations

This study has a few limitations which can be considered as future research directions. One limitation of this study is the missing cutoff scores for the KS-PPMC statistic. The universal cutoff scores for KS-PPMC statistics may not exist but false discovery rate—a method to control the error rate could be further investigated to find the criterion for KS-PPMC statistics. The second limitation is the computation time. For some CFA structures with large number of indicators, estimating posterior information of saturated model may be time-consuming, if not impossible. For this situation, the variational approximations method could be an alternative option (e.g., Dang & Maestrini, 2021). The last limitation is prior settings. Different choices of priors may affect the convergence of the model and even the detection error rates of PPMC. However, this problem has not been well researched. Future studies may focus on investigating the sensitivity of priors on the performance of PPMC methods.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Jihong Zhang http://orcid.org/0000-0003-2820-3734
Jonathan Templin http://orcid.org/0000-0001-7616-0973
Catherine E. Mintz http://orcid.org/0000-0002-8959-0013

References

Asparouhov, T., & Muthén, B. (2020). 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

Boomsma, A. (1987). The robustness of maximum likelihood estimation in structural equation models. In P. Cuttance and R. Ecob (Eds.), *Structural modeling by example: Applications in educational, sociological, and behavioral research* (pp. 160–188). Cambridge University Press.

Dang, K.-D., & Maestrini, L. (2021). *Fitting structural equation models via variational approximations. arXiv:2105.15036 [stat]*. Retrieved August 7, 2021, from http://arxiv.org/abs/2105.15036

Garnier-Villarreal, M., & Jorgensen, T. D. (2019). Adapting fit indices for bayesian structural equation modeling: comparison to maximum likelihood. *Psychological Methods*, *25*, 46–70. https://doi.org/10.1037/met0000224

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807 http://www.jstor.org/stable/24306036 .

Grønneberg, S., & Foldnes, N. (2019). Testing model fit by bootstrap selection. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 182–190. https://doi.org/10.1080/10705511.2018.1503543

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution*. Supplementary Educational Monographs, no. 48. Chicago: University of Chicago, Department of Education.

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

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

Lee, T., Cai, L., & Kuhfeld, M. (2016). A poor person's posterior predictive checking of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 206–220. https://doi.org/10.1080/10705511.2015.1014041

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*, 519–537. https://doi.org/10.1177/0146621608329504

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

Marcoulides, K. M., Foldnes, N., & Grønneberg, S. (2020). Assessing model fit in structural equation modeling using appropriate test statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 369–379. https://doi.org/10.1080/10705511.2019.1647785

McDonald, R. P. (1999). *Test theory: A unified approach*. Erlbaum.

Muthén, B., & 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

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. In Proceedings of the 3rd International workshop on distributed statistical computing.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association*, *95*, 1143–1156. https://doi.org/10.1080/01621459.2000.10474310

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Wu, H., Yuen, K.-V., & Leung, S.-O. (2014). A novel relative entropy–posterior predictive model checking approach with limited information statistics for latent trait models in sparse 2k contingency tables. *Computational Statistics & Data Analysis*, *79*, 261–276. https://doi.org/10.1016/j.csda.2014.06.004