

Adapting Fit Indices for Bayesian Structural Equation Modeling: Comparison to Maximum Likelihood

Mauricio Garnier-Villarreal
Marquette University

Terrence D. Jorgensen
University of Amsterdam



Abstract

In a frequentist framework, the exact fit of a structural equation model (SEM) is typically evaluated with the chi-square test and at least one index of approximate fit. Current Bayesian SEM (BSEM) software provides one measure of overall fit: the posterior predictive p value (PPP_{χ^2}). Because of the noted limitations of PPP_{χ^2} , common practice for evaluating Bayesian model fit instead focuses on model comparison, using information criteria or Bayes factors. Fit indices developed under maximum-likelihood estimation have not been incorporated into software for BSEM. We propose adapting 7 chi-square-based approximate fit indices for BSEM, using a Bayesian analog of the chi-square model-fit statistic. Simulation results show that the sampling distributions of the posterior means of these fit indices are similar to their frequentist counterparts across sample sizes, model types, and levels of misspecification when BSEMs are estimated with noninformative priors. The proposed fit indices therefore allow overall model-fit evaluation using familiar metrics of the original indices, with an accompanying interval to quantify their uncertainty. Illustrative examples with real data raise some important issues about the proposed fit indices' application to models specified with informative priors, when Bayesian and frequentist estimation methods might not yield similar results.



Translational Abstract

As Bayesian structural equation modeling (BSEM) has become more accessible with user-friendly software, there is still a need to research and present BSEM in a familiar setting for applied researchers. A basic step of SEM is the test of model fit, which in frequentist SEM is done with an exact fit chi-square test and at least one index of approximate fit. However, current BSEM model fit testing is limited. For this reason, the present study develops and tests 7 chi-square-based approximate fit indices (RMSEA, Gamma-Hat, adjusted-Gamma-Hat, Mc, CFI, TLI, NFI) for BSEM. These approximate fit indices are shown to be in the same metric as the frequentist counterparts, which allows one to have the same familiar interpretation of model-fit evaluation. The proposed BSEM approximate fit indices have an added advantage over their frequentist counterparts in that they allow quantifying uncertainty for each one with the posterior standard deviation and credible interval.

Keywords: Bayesian, fit indices, model fit, structural equation modeling, BSEM

Supplemental materials: <http://dx.doi.org/10.1037/met0000224.supp>


This article was published Online First June 10, 2019.

 Mauricio Garnier-Villarreal, College of Nursing, Marquette University;  Terrence D. Jorgensen, Department of Child Development and Education, University of Amsterdam.

We thank Marquette University for the use of the high-performance computing cluster, without which our simulation study would not have been possible, partly funded by National Science Foundation awards OCI-0923037 "MRI: Acquisition of a Parallel Computing Cluster and Storage for the Marquette University Grid (MUGrid)" and CBET-0521602 "Acquisition of a Linux Cluster to Support College-Wide Research & Teaching Activities." We also thank Ed Merkle for his assistance in updating his *blavaan* package to implement our pro-

posals, as well as Hao Wu for his feedback on earlier versions of this article. These results were presented as a paper in July 2018 at the 83rd annual International Meeting of the Psychometric Society in New York, New York. Data and syntax files associated with this project, as well as online supplementary figures, are available on the Open Science Framework (<https://osf.io/afkdw/>).

 The data are available at <https://osf.io/afkdw/>

 The experiment materials are available at <https://osf.io/afkdw/>

Correspondence concerning this article should be addressed to Mauricio Garnier-Villarreal, College of Nursing, Marquette University, 530 North 16th Street, Office 239, Milwaukee, WI 53233. E-mail: mauricio.garnier@marquette.edu

The availability of Markov chain Monte Carlo (MCMC) estimation in user-friendly structural equation modeling (SEM) programs such as *Mplus* (Muthén & Muthén, 1998–2017), Amos (Arbuckle, 2012), and the R (R Core Team, 2018) package *blavaan* (Merkle & Rosseel, 2018) has contributed to the increasing popularity of Bayesian SEM (BSEM; Muthén & Asparouhov, 2012; Song & Lee, 2012). Model-fit evaluation for BSEM has therefore become a topic of recent debate. In a frequentist framework, the exact fit of an SEM is tested with the chi-square statistic, typically complemented by reporting at least one index of approximate fit (Brown, 2006; Hu & Bentler, 1999; Kline, 2016). These familiar fit measures have not traditionally been provided as standard output in BSEM software. Building on the recent proposal of a Bayesian approximate fit measure (Hoofs, van de Schoot, Jansen, & Kant, 2018), we propose several approximate chi-square-based fit indices for BSEM that are calculated from a Bayesian analog of the chi-square statistic.

Hoofs et al. (2018) were motivated to define a fit index for BSEM for the same reason that motivated the development of numerous fit indices for SEM: to supplement a test of exact fit with a descriptive measure of approximate fit, which they noted is especially useful when large samples provide great power to detect trivial misspecifications. We were motivated to extend their ideas by defining fit measures that would behave consistently with the familiar fit measures in SEM, so we begin by considering the case of models with noninformative priors, in which case Bayesian and frequentist estimation routines provide asymptotically equivalent results.

Although it is also of interest to employ fit indices in the more general case, in which priors may be weakly or strongly informative, there are still advantages to fitting an SEM in a Bayesian framework that do not involve the incorporation of prior information. First, more complex (even intractable) models can be fit in a Bayesian framework that are not feasible with standard estimation routines based on covariance structure analysis, such as models that are nonlinear either in the latent variables¹ (e.g., including polynomial or interaction terms among latent variables) or in the parameters² or that account for complex dependencies among observations—examples of such BSEM applications can be found in Congdon (2009), van der Lans, van den Bergh, and Dieleman (2014), and Song and Lee (2006). The *blavaan* package, in particular, allows users to easily specify a basic SEM in *lavaan* (Rosseel, 2012) model syntax, then edit the automatically generated syntax from a general Bayesian modeling program to include features unavailable in standard SEM software (e.g., by specifying a beta regression for a proportion outcome). Second, models can be fit to smaller samples without violating an asymptotic assumption. Third, rather than relying on normal-theory confidence intervals (CIs) derived from the delta method for functions of parameters (e.g., indirect effects are products of direct effects), estimated posterior distributions can be used to calculate complex functions of parameters without assuming they must also have normal sampling distributions, yielding more robust credible intervals (the Bayesian analog of CIs; Y. Yuan & MacKinnon, 2009). Chi-square-based fit indices are very complex functions of model parameters, and most have unknown sampling distributions, so defining fit indices for BSEM allows access to intervals³ that can quantify uncertainty due to sampling error. Fourth, 95% credible intervals for model parameters have a more intuitive probabilistic

interpretation than a 95% CI (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). Fifth, Bayesian inference has multiple advantages over frequentist inference, even in simpler applications, such as the direct inference about the parameters of interest (θ) instead of inferring about a null hypothesis (related to the unintuitive interpretation of CIs); by conditioning on the data, the accuracy of a credible interval is not identified with the long-run behavior of the estimator (Kruschke, 2010); and frequentist inferential procedures tend to result in misinterpreting the p value and overestimating its information about the “significance” of a result (Matthews, 2001). Efron (1986), Matthews (2001), Wagenmakers, Lee, Lodewyckx, and Iverson (2008), and Kruschke (2010) provide details about the comparison of Bayesian and frequentist inference.

Thus, we consider it meaningful to define fit indices for BSEM at least in the limited case of noninformative priors (and not just in the case of very large samples; Hoofs et al., 2018), in which case decades of research on fit indices in SEM might also be relevant at least for this subset of BSEMs. As noted by van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli (2017) in their review of 25 years of Bayesian psychological science, only 12.6% of published articles reported the ability to specify priors as a motivating factor for using Bayesian methods. Because 31.1% of publications failed to report information about their priors at all (van de Schoot et al., 2017), it is difficult to discern how many BSEM publications have solely used noninformative priors since Scheines, Hoijtink, and Boomsma (1999) proposed their estimation, but 8.4% of publications using Bayesian methods seemed to imply that the authors relied on software packages’ default priors (typically noninformative; Arbuckle, 2012; Merkle & Rosseel, 2018; Muthén & Muthén, 1998–2017). So despite the popularity of small-variance priors proposed by Muthén and Asparouhov (2012) to account for trivial model misspecifications, we posit that it is not unreasonable to assume that BSEM is also commonly applied with noninformative priors.

We begin by reviewing some frequentist measures of model fit as well as fit measures currently provided by BSEM software. We then propose how to adapt familiar SEM fit measures for BSEM and compare our proposal with that of Hoofs et al. (2018). We present results from a simulation study designed to compare our proposed BSEM fit indices with their frequentist counterparts under maximum likelihood estimation (MLE) in various conditions, after which we enumerate important issues for further investigation—namely, the effects of missing data and informative priors. Using the well-known Holzinger and Swineford (1939) data set, we verify our expectations about these effects in the section Illustrative Examples, which demonstrate the importance of future

¹ Some restricted cases of nonlinearity in variables have been proposed, such as latent moderated structures for normally distributed latent variables (Klein & Moosbrugger, 2000), but are rarely implemented in SEM software packages (Muthén & Muthén, 1998–2017).

² Rare exceptions can be fitted with MLE by placing nonlinear constraints on certain estimated parameters. See Grimm and Ram (2009) for an example of nonlinear latent growth models.

³ Some resampling methods have been proposed for obtaining confidence intervals for SEM fit indices in a frequentist framework, such as bootstrapping (Cheng & Wu, 2017; Zhang & Savalei, 2016), permutation (Jorgensen, Kite, et al., 2018), and Monte Carlo simulation (Millsap, 2007; Pornprasertmanit, 2014; Pornprasertmanit, Wu, & Little, 2013).

Monte Carlo research into these issues. Our online OSF materials⁴ include R syntax to replicate these example analyses, which also show how to obtain the proposed fit indices with the R package *blavaan* (Merkle & Rosseel, 2018).

Frequency Fit Measures

The Chi-Square Statistic

The chi-square test statistic 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, 2016) are based on comparing the sample covariance matrix \mathbf{S} with the model-implied covariance matrix $\Sigma(\hat{\theta})$ (or simply $\hat{\Sigma}$). The most commonly implemented is the maximum likelihood discrepancy function:

$$F_{ML} = \log|\hat{\Sigma}| - \log|\mathbf{S}| + \text{trace}(\mathbf{S}\hat{\Sigma}^{-1}) - p + (\bar{\mathbf{x}} - \hat{\boldsymbol{\mu}})^T \hat{\Sigma}^{-1} (\bar{\mathbf{x}} - \hat{\boldsymbol{\mu}}), \quad (1)$$

where p is the number of variables in the model, $\bar{\mathbf{x}}$ is the vector of sample means, and $\hat{\boldsymbol{\mu}}$ is the vector of model-implied means. The corresponding statistic is calculated⁵ as $\chi^2_{ML} = N \times F_{ML}$.

More generally, χ^2_{ML} is a likelihood ratio test (LRT) statistic, calculated by plugging the n th observation's (perhaps incomplete) data vector \mathbf{y}_n into the multivariate normal log-likelihood (ℓ) function:

$$\ell_n = -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} (\mathbf{y}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{y}_n - \boldsymbol{\mu}), \quad (2)$$

using model-implied $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ derived from estimated model parameters in place of $\boldsymbol{\mu}$ and Σ above. Summing yields the log-likelihood of the hypothesized model (M_H),

$$\ell_H = \sum_{n=1}^N \ell_n. \quad (3)$$

Similarly, using the observed sample statistics $\bar{\mathbf{x}}$ and \mathbf{S} in place of $\boldsymbol{\mu}$ and Σ above yields the log-likelihood (ℓ_S) of the saturated model (M_S). The LRT statistic for M_H is also called its *deviance* from M_S :

$$\chi^2_{ML} = -2(\ell_H - \ell_S), \quad (4)$$

which is distributed as a chi-square random variable with $df = p^* - q$, where p^* is the number of nonredundant sample moments and q is the number of estimated parameters. In analyses of covariance structure only, $p^* = \frac{p(p+1)}{2}$, whereas $p^* = \frac{p(p+3)}{2}$ in mean and covariance structure (MACS) models.

MLE assumes the observed variables are multivariate normally distributed in order for the test statistic to be asymptotically distributed as a chi-square random variable. Other discrepancy functions can be applied and multiplied by N to calculate a model-fit statistic. Some discrepancy functions (e.g., unweighted least squares) allow the normality assumption to be relaxed (Browne, 1984). Robust corrections have also been developed to adjust the ML test statistic to be more approximately chi-square distributed when assumptions are violated (Savalei, 2014).

Regardless of the discrepancy function (or whether a robust correction was applied), the chi-square statistic tests the null hypothesis of exact model fit (H_0 : the hypothesized model perfectly represents the true data-generating process).

Because theoretical models are, by necessity, merely approximations of reality (MacCallum, 2003), the H_0 of exact fit is often considered a priori to be false. Because researchers typically cannot reasonably expect to retain H_0 in practice, the chi-square test is often of limited general or practical interest (West, Taylor, & Wu, 2012).

Furthermore, the power of the chi-square 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 chi-square 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 chi-square value by adjusting it or comparing it with another model's chi square, for example, to correct for model complexity, number of parameters, or overfitting. So the chi-square statistic has long remained the focus of overall model fit in SEM, even if indirectly.

Non-Centrality-Based Fit Indices

When H_0 is false, the model-fit test statistic follows a noncentral chi-square distribution, with noncentrality parameter λ . When H_0 is true, $\lambda = 0$ and the expected value of chi square is its degrees of freedom (df), whereas the expected value of a noncentral chi square is the sum of its df and λ . Thus, the difference between a model's chi-square statistic and df is a sample estimate of the model's noncentrality parameter: $\hat{\lambda} = \chi^2 - df$. Several fit indices are based on a rescaling of $\hat{\lambda}$; we present the most popular ones below.

Each df represents a restriction on how estimated model parameters can reproduce the observed sample moments $\bar{\mathbf{x}}$ and \mathbf{S} . Thus, greater df implies a more parsimonious model, which might therefore deviate even more from the true data-generating process—referred to as *approximation discrepancy* (MacCallum, 2003). Steiger and Lind (1980) proposed expressing model misfit as an average across the number of restrictions the model made. To express this average misfit per df in the metric of the discrepancy function (F_{ML}), $\hat{\lambda}$ is divided by N as well as df :

$$\begin{aligned} \text{RMSEA} &= \varepsilon \\ &= \sqrt{\max\left[0, \frac{\hat{\lambda}}{df \times N}\right]} \\ &= \sqrt{\max\left[0, \frac{\chi^2_{ML} - df}{df \times N}\right]} \\ &= \sqrt{\max\left[0, \frac{F_{ML}}{df} - \frac{1}{N}\right]}. \end{aligned} \quad (5)$$

⁴ Online materials can be found at <https://osf.io/afkew/>.

⁵ Early software such as LISREL (Jöreskog & Sörbom, 2006) and EQS (Bentler, 2006) applied Wishart-theory likelihood to analyses of covariance structure only, so they used $N - 1$ instead of N . More recently developed software such as *Mplus* (Muthén & Muthén, 1998–2017) and *lavaan* (Rosseel, 2012) include a mean structure by default, so they apply normal-theory likelihood, which requires N as the multiplier (Widaman & Thompson, 2003).

Note that the population root mean square error of approximation (RMSEA; ϵ) is independent of sample size:

$$\text{RMSEA}_{\text{pop}} = \epsilon = \sqrt{\frac{F_{\text{ML}}}{df}}. \quad (6)$$

Unlike most fit indices, the RMSEA has a known sampling distribution (Browne & Cudeck, 1992), so a CI can be provided to test null hypotheses about specific population values of RMSEA (MacCallum, Browne, & Cai, 2006). Higher values of RMSEA correspond to worse fit. Common interpretations are that $\text{RMSEA} < .05$ indicates close fit and $\text{RMSEA} > .10$ indicates poor fit, with varying intermediate values proposed to indicate acceptable levels of approximate fit (Browne & Cudeck, 1992; Hu & Bentler, 1999; MacCallum, Browne, & Sugawara, 1996).

McDonald (1989) also proposed dividing $\hat{\lambda}$ by N but also exponentiated to express misfit in terms of likelihood rather than log-likelihood. Taking the reciprocal (i.e., a negative exponent) results in an index that measures goodness (rather than badness) of fit, with a theoretical upper bound of 1 indicating excellent model fit,

$$\text{Mc} = e^{-\frac{1}{2}(\frac{\hat{\lambda}}{N})}. \quad (7)$$

Steiger (1989) proposed another goodness-of-fit index as a function of $\hat{\lambda}$ (again, divided by N to express misfit in the metric of the discrepancy function) and the number of variables p ,

$$\hat{\Gamma} = \frac{p}{p + 2\frac{\hat{\lambda}}{N}}. \quad (8)$$

Like McDonald's (1989) centrality index (Mc), values of "gamma-hat" ($\hat{\Gamma}$) approaching 1 indicate better fit. Maiti and Mukherjee (1990) described this as an unbiased estimator of the population goodness-of-fit index (GFI; Jöreskog & Sörbom, 2006). An adjusted version ($\hat{\Gamma}_{\text{adj}}$) is less likely to be positively biased in small samples, although Hu and Bentler (1998) and Fan and Sivo (2007) showed that $\hat{\Gamma}$ is already very independent of sample size,

$$\hat{\Gamma}_{\text{adj}} = 1 - \frac{p}{df}(1 - \hat{\Gamma}). \quad (9)$$

Incremental Fit Indices

Incremental fit indices are based on the idea that there is a continuum between a worst-fitting (but still theoretically defensible) baseline model (M_0 ; typically an independence model, in which all correlations are fixed to zero) and the best-fitting model (M_S ; represented by the saturated model, in which all observed covariances are freely estimated). A hypothesized model (M_H) should lie somewhere between these two extremes, and incremental fit indices indicate where along the continuum M_H is located. Values closer to zero indicate M_H is closer to the poor-fitting M_0 , and values closer to 1 indicate M_H is closer to the perfect-fitting M_S . This allows nonnested hypothesized models to be compared, so long as they are both nested within the same M_S (true by definition) and the same specified M_0 is nested within all competing models (Bentler & Bonett, 1980; Widaman & Thompson, 2003).

The earliest incremental fit index was proposed by Tucker and Lewis (1973) as a reliability index to assist selecting the number of factors in exploratory factor analysis. Bentler and Bonett (1980) later referred to the Tucker-Lewis index (TLI) as the nonnormed fit index (NNFI) because its values can fall outside the 0 to 1 range (e.g., $\text{TLI} > 1$ when $\chi_H^2 < df_H$),

$$\text{TLI} = \text{NNFI} = \frac{\frac{\chi_0^2}{df_0} - \frac{\chi_H^2}{df_H}}{\frac{\chi_0^2}{df_0} - 1}, \quad (10)$$

where the subscripts "H" and "0" indicate to which model the chi-square and df belong. Bentler and Bonett also proposed a normed fit index (NFI) that is bound to a 0 to 1 scale,

$$\text{NFI} = \frac{\chi_0^2 - \chi_H^2}{\chi_0^2}. \quad (11)$$

Although NFI has the advantage over TLI of being restricted to a 0 to 1 scale, NFI is unfortunately heavily influenced by sample size, whereas TLI is relatively independent of sample size (Fan & Sivo, 2007; Hu & Bentler, 1998).

Bentler (1990) later proposed a similar index based on noncentrality—the comparative fit index (CFI)—which is also normed to fall between 0 and 1,

$$\text{CFI} = \frac{\max(0, \hat{\lambda}_0) - \max(0, \hat{\lambda}_H)}{\max(0, \hat{\lambda}_0)} = 1 - \frac{\max(0, \hat{\lambda}_H)}{\max(0, \hat{\lambda}_0)}. \quad (12)$$

CFI has the advantage over both TLI and NFI by being both normed and relatively independent of sample size (Fan & Sivo, 2007; Hu & Bentler, 1998).

On the Interpretation and Application of SEM Fit Indices

We find it reasonable to interpret fit indices as merely descriptive measures of approximate fit, and to interpret questionable values as indicating not that the model should be rejected outright but that further investigation of local sources of misspecification (e.g., correlation residuals or modification indices) would be warranted. Fit indices were not proposed to function as test statistics but rather to "provide important adjunct information in evaluating models" (Bentler & Bonett, 1980, p. 604). Thus, similar to measures of effect size in other contexts (e.g., Cohen, 1992, provided guidelines to interpret correlations, standardized mean differences, and proportions of variance explained as small, medium, or large), guidelines were sought for interpreting the magnitude of fit indices. Because the H_0 tested by χ_{ML}^2 is that there is no discrepancy between the model-implied (mean and) covariance structure and the true population structure, a large "effect size" would indicate that a significant χ_{ML}^2 test has identified a large discrepancy (i.e., "badness of fit") between a hypothesized model and the true data-generating process. Initially proposed guidelines were largely heuristic, based on experience (see Bentler & Bonett, 1980, p. 600; Browne & Cudeck, 1992, p. 239; MacCallum et al., 1996, p. 134). However, Hu and Bentler (1998, 1999) used Monte Carlo simulation results to develop a more objective set of criteria for interpreting fit indices, based on a hypothesis-testing rationale (i.e.,

minimizing Type I and II errors), which has been met with much criticism (Beauducel & Wittmann, 2005; Fan & Sivo, 2005; Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011; Heene, Hilbert, Freudenthaler, & Bühner, 2012; Marsh, Hau, & Wen, 2004).

In our view, there are at least three major limitations preventing any of the previously proposed cutoffs from being generally useful to “test hypotheses” about approximate fit. First, they do not account for sampling variability. When claiming fit indices are relatively unaffected by sample size (e.g., Hu & Bentler, 1998), that only refers to the mean of their sampling distribution; as with any statistic, their sampling variance shrinks with larger N (Marsh et al., 2004). Second, sampling distributions of fit indices (including their means) vary across characteristics of the data (e.g., missing data: Davey, 2005; categorical data: Sass, Schmitt, & Marsh, 2014) and the model (e.g., number of factors and indicators; Jorgensen, Kite, Chen, & Short, 2018). Third, it is rarely clear what the numerical value of an index means in any absolute sense, so even when using a fit index (typically RMSEA’s 90% CI) to actually test a less restrictive H_0 (MacCallum et al., 2006; K.-H. Yuan, Chan, Marcoulides, & Bentler, 2016), there is little justification for preferring a particular value for the H_0 .

By approximating a sampling distribution consistent with M_H , “Bayesian variant[s]” (Hoofs et al., 2018, p. 543) of SEM fit indices would provide a viable solution to the first two problems (as resampling methods have; Cheng & Wu, 2017; Jorgensen, Kite, et al., 2018; Zhang & Savalei, 2016). Although we do not set out to resolve the third issue in this article, we later note how Bayesian fit indices could be used to avoid the third major limitation by removing the need for a cutoff value at all. However, these three issues are contingent on applying cutoff guidelines as critical values to test a H_0 of approximate fit. Again, we do not explicitly endorse this interpretation, instead encouraging researchers to consider the fit indices descriptive of model fit. No index will be fully descriptive, so any questionably poor values (including within the 90% CI of RMSEA) should be considered by researchers as an invitation to explore how their model fails using tools for detecting local misspecification.

Bayesian Model-Fit Assessment

Levy (2011) concluded in his review of BSEM model evaluation approaches that techniques for assessing model fit remain underdeveloped. Thus, although the increased availability of BSEM software for applied research holds great promise, particularly for complex modeling situations, the full scope of BSEM’s practical application remains limited due to the corresponding lack of general guidelines and best practices for model evaluation for applied users.

Although Hoofs et al. (2018, p. 543) recently proposed a “Bayesian variant of the RMSEA” (BRMSEA), most BSEM software currently provides only one measure of overall fit: the posterior predictive p value (PPP; Gelman, Meng, & Stern, 1996) based on the familiar chi-square model-fit statistic. Most BSEM software also provides only two indices for model comparison—the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) and Schwarz’s (1978) so-called Bayesian information criterion (BIC)—although *blavaan* also provides additional, more recently developed information criteria (Vehtari, Gelman, & Gabry, 2017; Watanabe, 2010). Our proposed

Bayesian fit indices are based on the same quantities utilized by Hoofs et al. to calculate BRMSEA, which include aspects of posterior predictive model checking (PPMC) as well as the effective number of parameters (pD) used to calculate information criteria such as DIC. We review these measures below, before discussing Hoofs et al.’s proposed BRMSEA and introducing our proposed Bayesian indices of overall model fit.

Posterior Predictive Checks

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 under-predicts the feature of the data specified by the chosen discrepancy function. The PPP estimates this probability as the proportion of $i = 1, 2, \dots, I$ 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 advocate its use as an informative diagnostic for identifying how a model fails (Gelman & Shalizi, 2013) rather than for traditional hypothesis testing.

PPMC is flexible regarding the discrepancy function, which can evaluate any feature of the model from which predictions can be derived. For example, in general linear models, the discrepancy function can be summary measures such as the mean, standard deviation, minimum, or maximum values of the model’s outcome variable (Gelman et al., 2014). Levy (2011) described some reasonable discrepancy functions for evaluating an SEM, such as the

SRMR, the model-implied factor correlations, and the familiar chi-square model-fit statistic (i.e., the likelihood ratio comparing the hypothesized model with a saturated model). Because BSEM software packages (Arbuckle, 2012; Merkle & Rosseel, 2018; Muthén & Asparouhov, 2012) uniformly report only PPP based on the chi-square statistic—and because this paper focuses only on chi-square-based fit indices—we confine our discussion of PPP to this special case, denoting it as PPP_{χ^2} .

Information Criteria Based on the Posterior Distribution

Spiegelhalter et al. (2002) proposed DIC as a Bayesian generalization of Akaike's information criterion (AIC), which adjusts a likelihood-based measure of model fit (e.g., the log-likelihood of the model) by taking the model's complexity into account. Complexity can be quantified by the number of estimated parameters (p), although that oversimplifies how models can be parsimonious (Preacher, 2006). In the SEM framework, AIC can be calculated⁶ as $\chi^2 + 2p$, but in BSEM, the number of estimated parameters cannot simply be represented as an integer. For example, estimating three parameters with restrictively informative priors effectively limits the parameter space much more than estimating the same three parameters using uninformative priors, so the number of estimated parameters should differ in these situations.

The effective number of parameters pD (or \hat{p} ; Vehtari et al., 2017) can be estimated from the posterior distribution by calculating the deviance (i.e., χ^2) using each vector of parameters ($\tilde{\theta}_i$) sampled from the posterior distribution during MCMC estimation, the same way one would calculate chi square using ML estimates of the model parameters. But sampling $\tilde{\theta}_i$ repeatedly from the joint posterior yields a posterior distribution of the deviance, with posterior-mean deviance

$$\bar{D} = \frac{1}{I} \sum_{i=1}^I D(\tilde{\theta}_i). \quad (13)$$

A different point-estimate for the model's deviance, $D(\bar{\theta})$, can be calculated using $\hat{\mu}$ and $\hat{\Sigma}$ implied by the posterior means of the parameters $\bar{\theta}$, which is analogous to the χ^2_{ML} calculated using the ML point estimates $\hat{\theta}$ (which are the mode rather than the mean of the likelihood function). The sampled parameters $\tilde{\theta}_i$ at a particular iteration i in a Markov chain will rarely (if ever) be identical to the posterior mean $\bar{\theta}$, so the discrepancy function D_i calculated at iteration i is expected to exceed $D(\bar{\theta})$ because values of $\tilde{\theta}_i$ further from $\bar{\theta}$ yield smaller (log-)likelihoods of the data. Spiegelhalter et al. (2002) called the degree to which \bar{D} exceeds $D(\bar{\theta})$ the *effective number of parameters*,

$$pD_{DIC} = \bar{D} - D(\bar{\theta}). \quad (14)$$

Conceptually, pD quantifies the degree to which the fit of the model improves due to allowing unknown values to vary across iterations of the Markov chain rather than fixing them to hypothesized values. Thus, it is a crude measure of model complexity, but it is not a count of the number of parameters (as in the frequentist framework). Another definition for pD is as a function of the variance of D_i across iterations of the Markov chain (Vehtari et al., 2017), which perhaps more clearly illustrates that pD represents the uncertainty about how well a hypothesized model actually fits

the data, and that uncertainty increases as we place fewer restrictions (or less restrictive priors) on our fitted model.

DIC itself can be calculated analogously to AIC, with the deviance evaluated at the posterior mean $D(\bar{\theta})$ substituted for the analogous χ^2_{ML} , and the effective number of parameters pD substituted for the number of parameters in MLE:

$$DIC = D(\bar{\theta}) + 2pD = \bar{D} + pD. \quad (15)$$

Alternatively, DIC can be calculated⁸ as a function of the posterior-mean deviance \bar{D} , as shown on the far right-hand side of Equation 15.

Hoofs et al.'s (2018) BRMSEA

Although there are not established fit indices in BSEM that are equivalent to their frequentist counterparts (which might not even be possible, given the lack of df or an integer number of estimated parameters in a Bayesian context), a recently proposed Bayesian analog of RMSEA (Hoofs et al., 2018) showed promise for evaluating approximate fit in large samples ($N > 1,000$). Hoofs et al. (2018) presented an intriguing approach to BRMSEA that attempts to estimate the same parameter that RMSEA estimates in a frequentist context. They replaced the quantities χ^2_{ML} and df in Equation 5 with analogous quantities representing a model's misfit and complexity, as conceptualized in terms of posterior predictive model checking (PPP_{χ^2}) and the effective number of parameters (pD) in DIC. Specifically, rather than using the discrepancy function evaluated at the posterior mean— $D(\bar{\theta})$, which is analogous to the chi square when priors are noninformative—model misfit is represented by the difference in discrepancies of observed and replicated data ($D_i^{obs} - D_i^{rep}$) at each iteration in the Markov chain, as in posterior predictive model checks (hence, our “PPMC” superscript in Equation 16 below). Model complexity (as represented by $df = p^* - q$ in Equation 5) is calculated as $p^* - pD$,

$$BRMSEA_i^{PPMC} = \sqrt{\max\left[0, \frac{(D_i^{obs} - D_i^{rep}) - (p^* - pD)}{(p^* - pD) \times N}\right]}. \quad (16)$$

Hoofs et al. (2018) proposed their $BRMSEA^{PPMC}$ to complement the PPP_{χ^2} , because in large samples, PPP_{χ^2} rejects all models with even minor misspecification (similar to χ^2 in SEM). The results from their simulation show that under those conditions (large N , minor misfit, “significantly” small PPP_{χ^2}), their $BRMSEA^{PPMC}$ would indicate approximately well-fitting models are acceptable, according to commonly used cutoffs (i.e.,

⁶ Some SEM software, such as *Mplus* and *lavaan*, calculate AIC as $-2 \times \log(\text{likelihood}) + 2p$, which yields equivalent rankings of competing models because each model's chi square is calculated relative to the same saturated model.

⁷ Or more generally, further from the posterior mode of θ , which is the same as the posterior mean when the posterior distribution is symmetric and unimodal (e.g., normal).

⁸ Vehtari et al. (2017) compared pD as developed for DIC with an expression developed for Watanabe's (2010) widely applicable information criterion (WAIC)—which Vehtari et al. called more “fully Bayesian in that it uses the entire posterior distribution” (p. 1414)—as well as a Pareto-smoothed importance sampling (PSIS) approximation of leave-one-out (LOO) cross-validation. We do not discuss the more complex calculations of pD presented by Vehtari et al., but we compare them in our simulation study.

RMSEA < .08 is acceptable, RMSEA < .05 indicates close fit; Browne & Cudeck, 1992).

To operate as a measure of effect size to accompany the statistical significance test offered by chi square, a fit index should be sensitive only to misspecification (Fan & Sivo, 2007). Although BRMSEA^{PPMC} has been the first attempt to translate a frequentist fit index to BSEM, it does appear sensitive to sample size as well as model size. For example, when Hoofs et al. (2018) simulated data from a six-indicator confirmatory factor analysis (CFA), the BRMSEA 90% credible interval (with noninformative priors) implied a similar amount of sampling variability as implied by the RMSEA 90% CI under MLE. However, the 90% credible intervals for BRMSEA were much narrower than the 90% CIs for RMSEA when fitting larger, 12-indicator CFA models to smaller samples, leading to much lower power of BRMSEA than RMSEA to detect even severe misspecifications. Thus, when fitting models that are large relative to a smaller sample size, neither the BRMSEA nor PPP χ^2 would have power to detect important problems with the model.

This has important implications for multidimensional assessment of (approximate) model fit in BSEM, especially in the common case when samples are substantially less than 1,000. When samples are particularly small, the χ^2_{ML} might have very little power, yet the 90% CI of RMSEA will be wide enough to prevent rejecting the hypothesis of inadequate or poor fit (RMSEA > 0.08 or 0.10). An example of this can be found in the SEM textbook by (Kline, 2016, p. 257, Table 9.9), in which a multigroup CFA with full factorial invariance was not rejected by the χ^2_{ML} test ($p = .229$). However, the upper confidence limit of RMSEA was 0.103, prompting Kline (2016) to inspect local sources of misfit, where he found 16 correlation residuals that exceeded .10 (recall this practice is consistent with our recommendation to consider any indication of questionable overall fit as worthy of further attention, not as a reason to outright reject a model).

We therefore argue that there are common situations (e.g., small to moderate sample size) when researchers would want a Bayesian analog of RMSEA to behave as RMSEA would under MLE, as opposed to indicating very certainly (i.e., very narrow credible intervals⁹) that a model with important misspecifications fits well. In the following section, we explore further methods to translate commonly used fit indices for use in BSEM, which can be expected to behave similarly to their MLE counterparts when noninformative priors are used for BSEM parameters (in which case, the posterior distribution is proportionally equivalent to the likelihood function). We consider a few cases of informative priors in the Illustrative Examples section, on the basis of which we offer important considerations for future research.

Proposed Adaptation of Fit Indices for BSEM

There is no ideal measure of fit in SEM. Rather, different indices evaluate different dimensions of model fit, leading many experts to propose supplementing the χ^2_{ML} test statistic with at least two additional fit indices (Brown, 2006; Hu & Bentler, 1998; Kline, 2016). The addition of a BRMSEA index allows Bayesian models to be evaluated relative to their complexity, which complements simply using PPP χ^2 to evaluate whether the observed data are consistent with the model. To allow for models to be evaluated across additional dimensions of (approximate or relative) fit, we

propose how additional non-centrality-based and incremental fit indices from SEM can be incorporated into BSEM. As previous research has shown (see Fan & Sivo, 2007, for a review of some issues), all fit indices have limitations and (dis)advantages in different situations, so an array of indices would allow for more nuanced evaluation of how a model fails than the BRMSEA alone.

We propose BSEM fit indices developed in a fashion similar to Hoofs et al. (2018), by using $p^* - pD$ in the role of df to represent model complexity (or rather, model parsimony). However, our proposals differ from Hoofs et al. in how we represent model misfit. A Bayesian analog of RMSEA that can be expected to behave like its frequentist counterpart (at least when using noninformative priors) might differ from Equation 16 by using the Bayesian analog of chi square: the deviance evaluated at the posterior mean (hence, the superscript “DevM”),

$$\text{BRMSEA}^{\text{DevM}} = \sqrt{\max\left[0, \frac{D(\bar{\theta}) - (p^* - pD)}{(p^* - pD) \times N}\right]}. \quad (17)$$

We posit that compared with BRMSEA^{PPMC}, BRMSEA^{DevM} would more closely estimate the same quantity that RMSEA estimates (i.e., ϵ ; Browne & Cudeck, 1992) in a frequentist framework, using an analogous (though not equivalent) definition. However, using $D(\bar{\theta})$ does not allow for the advantage of obtaining a credible interval for BRMSEA because Equation 17 only contains summaries of the data (N and p^*) and the posterior ($D(\bar{\theta})$ and pD).

A similar quantity can be obtained that does vary across iterations, by using $D(\bar{\theta}_i) = D_i^{\text{obs}}$ instead of $D(\bar{\theta})$. By substituting the right-hand side of Equation 13 for \bar{D} in Equation 14, rearranging terms provides a distribution centered at $D(\bar{\theta})$:

$$D(\bar{\theta}) = \frac{1}{I} \sum_{i=1}^I D_i^{\text{obs}} - pD. \quad (18)$$

Thus, substituting $(D_i^{\text{obs}} - pD)$ for $D(\bar{\theta})$ in Equation 17 yields a distribution of BRMSEA:

$$\begin{aligned} \text{BRMSEA}_i^{\text{DevM}} &= \sqrt{\max\left[0, \frac{(D_i^{\text{obs}} - pD) - (p^* - pD)}{(p^* - pD) \times N}\right]} \\ &= \sqrt{\max\left[0, \frac{D_i^{\text{obs}} - p^*}{(p^* - pD) \times N}\right]}. \end{aligned} \quad (19)$$

The subscript i indicates that BRMSEA^{DevM} in Equation 19 varies across samples of parameters drawn from their posterior distribution. BRMSEA^{DevM} resembles Hoofs et al.’s (2018) BRMSEA^{PPMC} but replaces D_i^{ep} by pD . Interestingly, the two occurrences of pD in the numerator of Equation 19 cancel out, resulting in a numerator that expresses misfit simply as the discrepancy at iteration i rescaled by the number of observed sample moments.

Two important observations are worth mentioning here. First, the proposed approach in Equation 19 yields neither a posterior distribution (because it is a function not only of the estimated parameters but also of the observed data) nor a posterior predictive distribution (because it is a function only of observed data, not data simulated from model parameters); rather, it yields a distribution of realized values (Levy, 2011) of a chi-square-based discrepancy

⁹ See Results for the 12-indicator Models C–E in Figure 2, Model C in Figure 3, and Model F2 in Figure 4 of Hoofs et al. (2018).

measure, which could therefore be used in a posterior predictive model check, unlike $\text{BRMSEA}^{\text{PPMC}}$. We provide further details in a later section. Second, it follows from Jensen's inequality¹⁰ that the mean of \sqrt{X} generally underestimates the square root of \bar{X} . So although the equivalence in Equation 18 holds, the mean of the distribution in Equation 19 will generally underestimate $\text{BRMSEA}^{\text{DevM}}$ in Equation 17. However, if the difference turns out to be negligible in practice, we could still expect $\text{BRMSEA}^{\text{DevM}}$ to behave approximately like the frequentist RMSEA (e.g., under MLE), which our simulation study is designed to investigate.

We similarly propose Bayesian versions of other approximate fit indices by replacing chi square and df with $(D_i^{\text{obs}} - pD)$ and $(p^* - pD)$, respectively in Equations 5 to 12; furthermore, the noncentrality parameter $\hat{\lambda}$ is replaced by $(D_i^{\text{obs}} - pD) - (p^* - pD) = D_i^{\text{obs}} - p^*$. With these substitutions, we are able to derive distributions of realized values not only for RMSEA but also for Mc , $\hat{\Gamma}$, $\hat{\Gamma}_{\text{adj}}$, TLI, NFI, and CFI (see formulas in the Appendix). As in Hoofs et al. (2018), the distribution of $\text{BRMSEA}_i^{\text{DevM}}$ (and other indices in the Appendix) can be summarized using a measure of central tendency, such as the mean (expected a posteriori; EAP), mode (modal a posteriori; MAP), or median of their distribution, which should typically be quite close to the value obtained using $D(\hat{\theta})$ as in Equation 17. Uncertainty about EAP, MAP, or median estimates can be represented with their respective standard deviations and with 5th and 95th percentile from their empirical distributions, similar to defining a 90% credible interval for a model parameter.

Incremental fit indices also require fitting a null model (M_0) to the data, which is traditionally an *independence model* in which means and variances are freely estimated but covariances are constrained to zero. Although Hoofs et al. (2018, p. 26) brought up the concern that an independence model would be hard to justify (even contradictory) when incorporating prior information, we currently only consider the case of noninformative priors. But we recommend defining a meaningful null model appropriate for answering a specific research question, taking into account the functional form of the hypothesized model and characteristics of the data (e.g., Widaman & Thompson, 2003, discussed alternative null models for multigroup and longitudinal data; Lai & Yoon, 2015, proposed a modified version of Rigdon's, 1998, null model specifically for evaluating measurement invariance). Similar to differences between discrepancies in posterior predictive checks, the differences (or ratios) between discrepancies of the hypothesized and null models in Equations 10 to 12 are calculated at each iteration of each model's Markov chain(s). Thus, for computational purposes, the same number of samples should be drawn from the posterior distribution of both models fit to the same data.

Distributions of the Proposed Fit Indices

Because D^{obs} is evaluated across the estimated posterior distribution of model parameters, uncertainty about fit indices is "borrowed" from uncertainty about the model parameters. Following from Levy (2011), the realized values D_i^{obs} approximate the posterior distribution of the discrepancy function, so the posterior distributions of derived quantities like $\text{BRMSEA}_i^{\text{DevM}}$ can be seen as realized values of fit indices defined analogously in Bayesian and frequentist frameworks. Empirical distributions of fit indices help (but may not fully) resolve the first major limitation of interpreting the magnitude of fit indices

relative to rule-of-thumb guidelines that we identified previously. And because the distribution of a fit index is derived from the estimated posterior distribution of the model parameters, features of the model are already taken into account, partially resolving the second major limitation (assumptions must still be made about the data distribution).

There typically exists some actual discrepancy between the hypothesized model and the true data-generating process, even if the model parameters were to be estimated using the population covariance matrix as input data (ruling out discrepancy due to sampling error; Cudeck & Browne, 1992; MacCallum, 2003). Although, in that sense, there is a population parameter ϵ (Equation 6) that the sample formula (Equation 5) estimates (likewise for other fit indices), we do not assert that $\text{BRMSEA}^{\text{DevM}}$ would consistently estimate ϵ . We only argue that in the special case of noninformative priors being used to estimate the same model parameters with MCMC as with MLE, the Bayesian analog of RMSEA should provide a reasonable approximation of ϵ by virtue of MCMC and MLE converging on equivalent parameter estimates in this case.

It was brought to our attention during the review process that referring to a "posterior distribution of a fit index" could be construed as misleading given that the discrepancy function D is a function of not only estimated parameters but also the data on which those parameters were conditioned. Although it would be possible to estimate data-model discrepancy as a parameter in a BSEM framework (e.g., as *adventitious error*; Wu & Browne, 2015), we do not develop such an approach here. Instead, we show how the familiar measures of approximate data-model correspondence in SEM (i.e., fit indices derived from the χ^2_{ML} statistic) can be conceptualized in a Bayesian framework by plugging analogous quantities into the same formulas derived in a frequentist framework.

Thus, we tend to refer to our proposed fit indices as having "distributions" with "intervals" around their means. For brevity in some instances (e.g., plots and tables), we sometimes refer to a "posterior distribution" of BRMSEA (or its 5th and 95th percentiles as constituting a 90% "credible interval"), following the existing convention of referring to a "posterior distribution for a discrepancy measure" (Levy, 2011, p. 672). But readers should recall that $\text{BRMSEA}^{\text{DevM}}$ estimates a quantity only analogous (not equivalent) to the frequentist RMSEA (likewise for other fit indices), and we use the terms "posterior distribution" and "credible interval" only in the sense that BRMSEA is a function of the data and model parameters that is evaluated across the estimated posterior distribution of model parameters.

The analogous definitions in Equations 5 and 19 yield similar interpretations for $\text{BRMSEA}^{\text{DevM}}$ and RMSEA, so it is tempting to apply the same rules of thumb for interpreting their magnitudes. However, one should not expect any proposed guidelines for interpreting the magnitude of RMSEA (or of any other fit index) to generalize¹¹ to $\text{BRMSEA}^{\text{DevM}}$. Likewise, although a 90% interval estimate could be used to test a hypothesis about a fit index (e.g., similar to MacCallum et al., 2006), the hypothesized value should not be derived from MLE-based guidelines when informative

¹⁰ An example proof can be found here: <https://math.stackexchange.com/questions/1204484/average-of-square-rootss-sum-vs-square-root-of-average>

¹¹ Guidelines proposed for fit indices under MLE have also been shown not to generalize to other frequentist estimators, such as DWLS with a mean- and variance-adjusted chi square for ordinal data (Sass et al., 2014).

priors are used. Furthermore, researchers should refrain from applying the standard interpretation of a 90% credible interval (i.e., conditional on the data, there is a 90% probability the parameter lies between these limits) until more is understood about the quantities that are consistently estimated by the proposed indices.

This brings us back to the third major limitation of proposed cutoff values for fit indices: It is not immediately apparent how to derive a cutoff value that can be meaningfully interpreted as indicating maximally ignorable or minimally important data-model misfit. Rather than relying on fixed cutoffs proposed under MLE, the magnitude of the realized values of fit indices could instead be compared with values consistent with the model simulated using PPMC, avoiding (though not resolving) the third major limitation of interpreting fit indices relative to cutoff values. As this issue falls beyond the scope of our current study, we return to this point in the Discussion to provide details as a direction for future research.

Monte Carlo Simulation Study

We borrowed design factors from Fan and Sivo (2007) to compare the sampling behaviors of fit indices under MLE and MCMC estimation across a range of sample sizes, model types (and sizes), and levels of model misfit. We chose their design elements because they synchronized levels of misspecification across model types and sizes such that the power to detect moderate and large amounts of misfit was always 51% and 88%, respectively, when $N = 100$ (see Table 1). Similar to Hoofs et al. (2018), we simulated data from CFA models with cross-loadings (CFA-A) and simple structure (CFA-B), but we also simulated data from large (12 indicators, four factors) and small (six indicators, two factors) structural regression models (SEM-A and SEM-B, respectively). Thus, we worked with models that had similar numbers of factors and indicators as Hoofs et al. worked with, but with the advantage of holding the practical impact of misfit constant across types of model and misspecification.

Figures 1, 2, 3, and 4 present path diagrams of the four population models, the population values used to generate multivariate-normal data, and which parameters were fixed to zero in analysis models that were moderately or severely misspecified. Because

Hoofs et al. (2018) found their BRMSEA behaved similar to RMSEA when $N \geq 1,000$, we did not generate samples larger than 1,000. We included five sample-size conditions: $N = 75, 100, 250, 500$, or 1,000. Our full-factorial $5 (N) \times 4 (\text{model types}) \times 3 (\text{levels of misfit: none, moderate, and severe})$ design therefore included 60 conditions, and we generated 1,000 replications in each condition. Data were generated using the `simulateData()` function in the R (R Core Team, 2018) package `lavaan` (Rosseel, 2012). Models were fit to data using MLE in `lavaan` and using MCMC estimation (specifically, Gibbs sampling) available in the R package `blavaan` (Merkle & Rosseel, 2018), which optionally uses JAGS (Plummer, 2017) or Stan (Carpenter et al., 2017) as the general Bayesian estimation program in the back end.

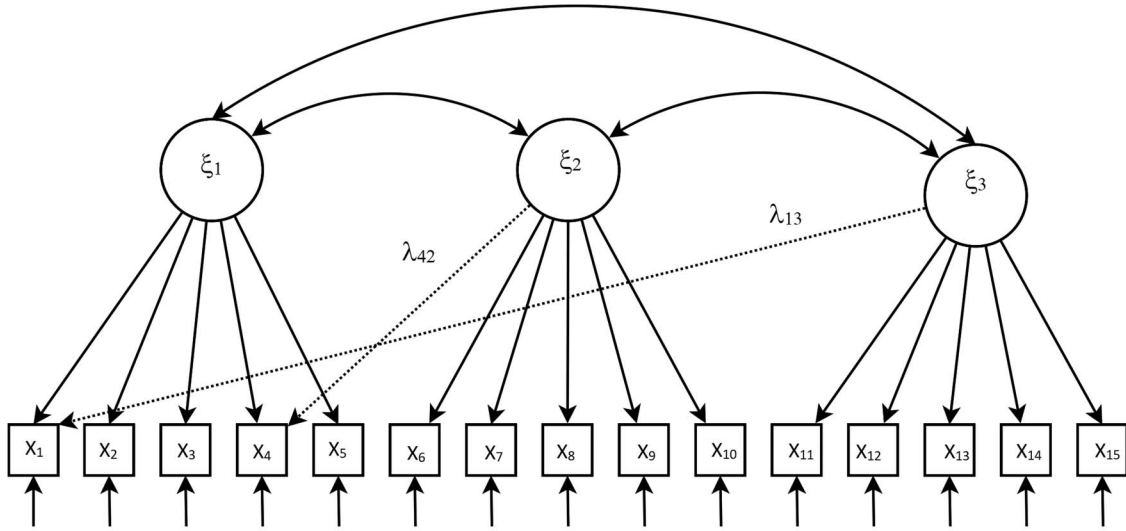
We used noninformative priors to analyze the simulated data. “Noninformative” indicates the prior distributions for the parameters provided little to no information above and beyond the information provided by the data, so that the posterior distributions are estimated around the information from the data instead of prior knowledge (Gelman, Simpson, & Betancourt, 2017). Priors for factor loadings, indicator intercepts, and regressions were $\sim \mathcal{N}(\mu = 0, \sigma^2 = 100)$ and priors for indicator residual standard deviations were $\sim \text{half-Cauchy}(\mu = 0, \sigma^2 = 2.5)$. For the CFA models, factor variances and covariances were $\sim \text{Wishart}^{-1}(\Psi = \mathbf{I}, \nu = nf + 1)$, where \mathbf{I} represents an identity matrix of dimension equal to the number of factors (nf), and degrees of freedom (ν) equal to number of factors plus 1. For the SEM models, factor covariances followed the same inverse-Wishart distribution as the CFA models, and the factor residual standard deviations followed the same half-Cauchy priors as the indicator residual standard deviations. It is pertinent to mention that these priors are noninformative with respect to the model parameters and the scale of the data in this simulation, but these priors could be considered informative in different conditions (e.g., data with larger scales).

Each model started with 30,000 burn-in iterations; if the model did not converge, this was iteratively increased by 5,000 until the model converged. Convergence of each Markov chain to the same posterior distribution was evaluated using the potential scale reduction factor (PSRF), also known as “univariate R-hat” (Gelman & Rubin, 1992). It was determined that the

Table 1
Comparison of Estimated to Population RMSEA

Misfit (power)	Model type	$F_{ML}(df)$	ϵ	RMSEA	BRMSEA	ΔML	$\Delta MCMC$
Level 0 ($\alpha = 5\%$)	CFA-A	0 (84)	0	.0172	.0202		
	CFA-B	0 (87)	0	.0169	.0194		
	SEM-A	0 (46)	0	.0166	.0251		
	SEM-B	0 (4)	0	.0225	.0453		
Level 1 (51%)	CFA-A	.2419 (85)	.0533	.0537	.0547	-.0004	-.0013
	CFA-B	.2416 (89)	.0525	.0548	.0562	-.0022	-.0036
	SEM-A	.1868 (48)	.0623	.0630	.0665	-.0006	-.0041
	SEM-B	.0726 (5)	.1204	.1129	.1140	.0075	.0064
Level 2 (88%)	CFA-A	.4521 (86)	.0725	.0746	.0754	-.0021	-.0029
	CFA-B	.4627 (90)	.0717	.0736	.0803	-.0019	-.0086
	SEM-A	.3557 (49)	.0852	.0861	.0890	-.0009	-.0038
	SEM-B	.1831 (8)	.1512	.1362	.1504	.0150	.0008

Note. Power = power of χ^2_{ML} test when $N = 100$; F_{ML} = maximum likelihood fit function; df = degrees of freedom; ϵ = population RMSEA; RMSEA = estimated RMSEA under MLE; BRMSEA = Bayesian variant of RMSEA using “DevM” formulation; ΔML = difference between ϵ and RMSEA (omitted when $\epsilon = 0$); $\Delta MCMC$ = difference between ϵ and BRMSEA (omitted when $\epsilon = 0$); Misfit = level of misspecification.



Level 1 Misspecified Model $\lambda_{42} = 0$

Level 2 Misspecified Model $\lambda_{42} = 0, \lambda_{13} = 0$

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

$$\Lambda_X(\text{transposed}) = \begin{vmatrix} .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .50 & .00 & .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 \\ .50 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .70 & .00 & .70 & .75 & .80 & .80 & .00 \end{vmatrix}$$

$$\Theta_\delta(\text{diagonal}) = \begin{vmatrix} .51 & .51 & .4375 & .36 & .36 & .51 & .51 & .4375 & .36 & .36 & .51 & .51 & .4375 & .36 & .36 \end{vmatrix}$$

Figure 1. Path diagram and population parameters for Confirmatory Factor Model A (CFA-A). Dotted paths represent nonzero population values that were fixed to zero in the analysis model. λ = factor loading matrix; Φ = factor covariance matrix; θ = residual covariance matrix.

model converged when $R\text{-hat} < 1.10$ for each parameter (Brooks & Gelman, 1998). Although the total (including burn-in) number of iterations could differ across replications, we always saved the same number of post-burn-in iterations (5,000 from each chain) to estimate the posterior distributions. For every analysis, we used three chains, yielding 15,000 iterations for drawing inferences. We did not “thin” the chains by discarding samples because although thinning decreases the autocorrelation between iterations (thus decreasing the Monte Carlo error of the posterior distribution’s sample statistics), it does not affect the posterior distribution itself nor inferences drawn from it (Gelman et al., 2014).

Because Hoofs et al. (2018) already showed via Monte Carlo simulations that their $\text{BRMSEA}^{\text{PPMC}}$ and the RMSEA under MLE do not converge until sample size is quite large ($N \geq 1,000$), the main goal of our simulation study was to evaluate whether our proposed $\text{BRMSEA}^{\text{DevM}}$ behaves as expected when priors are uninformative (i.e., similar to the RMSEA under MLE, which

converges on ϵ at much smaller N ; Curran, Bollen, Chen, Paxton, & Kirby, 2003). Thus, when comparing SEM fit indices under MLE with our proposed BSEM fit indices in the Results section, we omit the “DevM” label for brevity, referring only to BRMSEA, BCFI, and so forth, except in the case of explicitly comparing the “DevM” and “PPMC” formulations.

Monte Carlo Results

For each of the 60 conditions, we had 1,000 replications that converged for both the MLE and Bayesian models ($\text{PSRF} < 1.1$). The total computation time was 3,175.33 days (8.69 years), distributed across several parallel computers. To calculate BSEM fit indices with $p^* - pD$ in place of df , we chose pD_{LOO} rather than pD_{WAIC} or pD_{DIC} because it is preferred by Vehtari et al. (2017). But because we estimated models with noninformative priors, any pD was expected to be very close to the number of parameters under MLE. Table 2 shows that pD_{LOO} and pD_{WAIC} were both

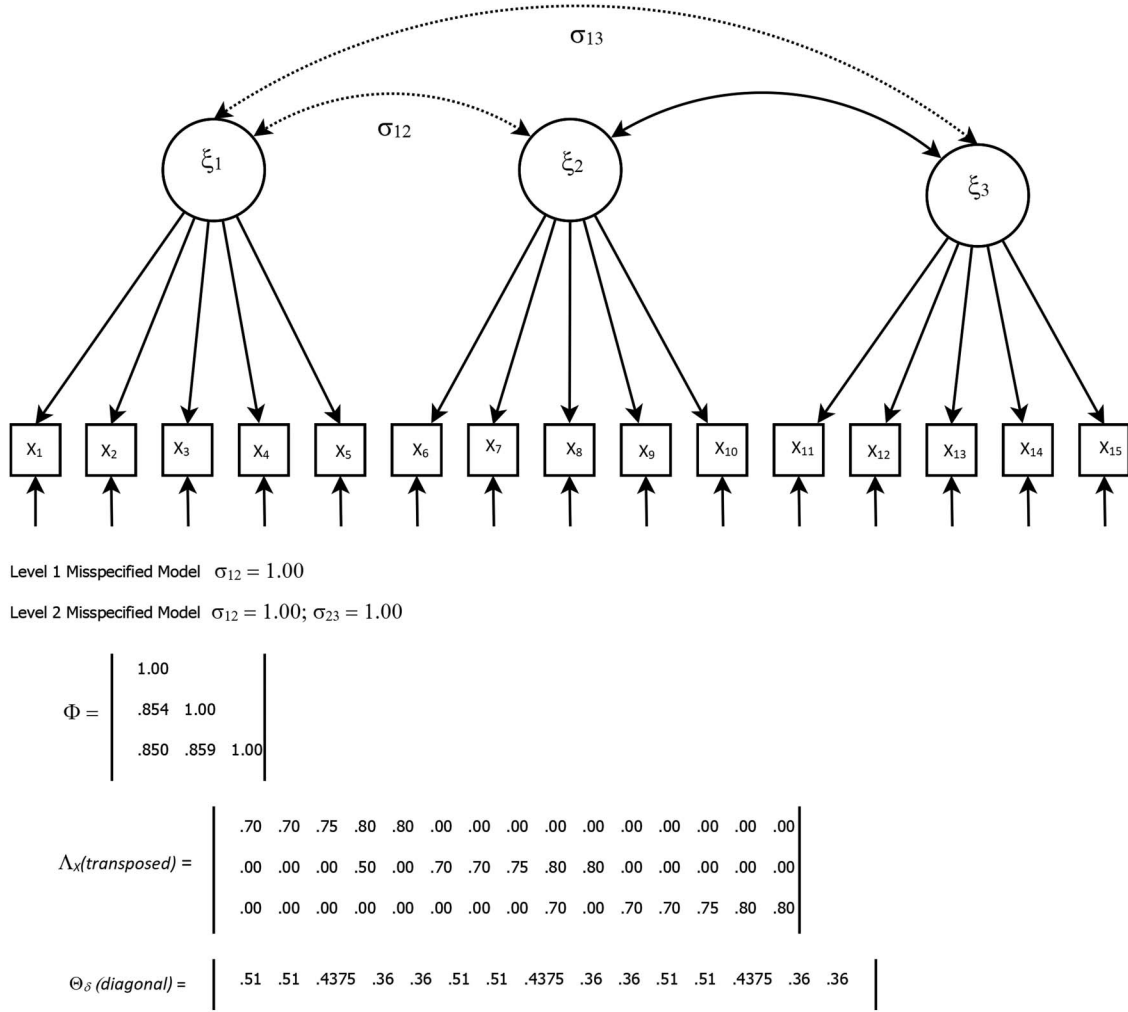


Figure 2. Path diagram and population parameters for Confirmatory Factor Model B (CFA-B). Dotted paths represent nonzero population values that were fixed to zero in the analysis model. λ = factor loading matrix; Φ = factor covariance matrix; θ = residual covariance matrix.

very similar to the number of parameters under MLE across model types and misspecification levels, whereas pD_{DIC} substantially underestimated the number of parameters for model SEM-A. Given this result, we recommend using either pD_{LOO} or pD_{WAIC} in practice.

Population RMSEA for each analysis model (i.e., ϵ with varying across levels of misspecification) was calculated by fitting the model to the population covariance matrix implied by its associated population-model parameters. F_{ML} (which is independent of sample size) in each condition was plugged into Equation 6, reported in Table 1 along with the average RMSEA and BRMSEA (averaged across sample size conditions). Compared with (Hoofs et al., 2018, see their Figures 2–4), Table 1 shows that across population models and levels of misspecification, both RMSEA and BRMSEA closely approximate ϵ . The largest deviations ($\hat{\epsilon} - \epsilon$) were when there was no misspecification ($\epsilon = 0$), which was due to a floor effect because RMSEA is bound below at zero. When this floor effect was not present, the absolute value of $\hat{\epsilon} - \epsilon$ ranged from 0.0008 to 0.0086, showing that, on average, BRMSEA

(like RMSEA under MLE) reproduced ϵ within two to three decimal places. The solid lines in Figure 5 further show that across sample sizes, RMSEA and BRMSEA tended to converge quickly on the same value. The larger deviations between RMSEA and BRMSEA in the small SEM model (SEM-B in Figure 4) may be related to the more erratic sampling variability of RMSEA when either N or df (but especially when both) are small (Kenny, Kaniskan, & McCoach, 2015).

Table 3 compares each BSEM fit index with its ML counterpart. Recall that $D(\hat{\theta})$ is the Bayesian analog of chi square. The BSEM fit indices consistently indicate slightly worse fit, at least in part because $D(\hat{\theta}) > \chi^2$ in 93.3% of replications; however, in most cases, this difference is minimal because the values are equal to the second decimal place. Looking at these differences as paired comparisons, we can estimate the Cohen's d as the standardized mean difference, representing the mean difference in units of SD . Following standard guidelines (Cohen, 1992), these are small to medium effect sizes. Table 3 also shows that fit indices have nearly equivalent standard deviations, and Figure 5 shows that for

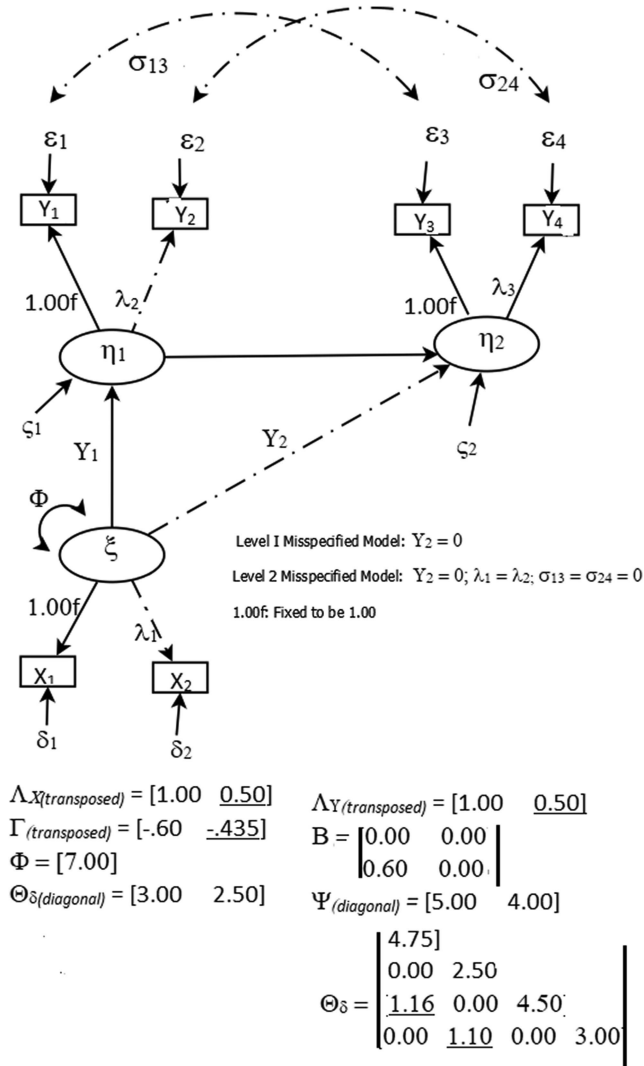


Figure 4. Path diagram and population parameters for Structural Equation Model B (SEM-B). Dotted paths represent nonzero population values that were fixed to zero in the analysis model. λ_x = factor loading matrix for exogenous factors; λ_y = factor loading matrix for endogenous factors; Φ = factor covariance matrix for endogenous factors; Ψ = factor covariance matrix for exogenous factors; θ = residual covariance matrix; Γ = regressions from exogenous factors; B = regressions from endogenous factors.

research. BSEM fit indices generally tended to be more sensitive to N and model type, and less sensitive to misspecification, than their ML counterparts. The differences tended to be small, but two substantial differences can be seen: BRMSEA seems much more affected by model type than RMSEA, and $\hat{\Gamma}_{adj}$ appears less sensitive to misspecification under MCMC than MLE. Consistent with past research (Fan & Sivo, 2007), (B)CFI and (B) $\hat{\Gamma}$ were substantially affected ($\eta^2 > 10\%$) only by model misspecification, which is ideal behavior for a fit index (Hu & Bentler, 1998). Finally, PPP_{χ^2} was affected by the interaction between misspecification level and sample size. Table 5 presents the Bayesian fit indices across misspecification levels, as misspecification increases the fit indices present

worse model fit. Table 6 shows that for models with minor misspecification, PPP_{χ^2} tended to reveal model misspecification as sample size increased. With minor misspecification at $N = 75$, $PPP_{\chi^2} < .01$ for 4.5% of replications compared with 95.3% of replications with $N = 500$. In the case of severe misspecification, 28.6% of replications had $PPP_{\chi^2} < .01$ at $N = 75$, compared with 100% of replications when $N = 500$. This is analogous to increased power of χ^2_{ML} when sample size increases, rejecting models even with trivial levels of misspecification, whereas small sample sizes are unable to detect severe levels of misspecification consistently.

To understand the nature of the differences between estimators, Figure 5 shows the average (B)RMSEA across conditions. To evaluate the estimated sampling variability in BRMSEA implied by its posterior distribution, the average of its 5th and of its 95th posterior percentiles (corresponding to 90% credible-interval limits) were plotted along with the average 90% confidence limits of RMSEA. BRMSEA appears generally less variable (more precisely estimated) than RMSEA, except when SEM-B was perfectly specified. However, recall from Table 3 that BRMSEA is just as variable as RMSEA, so the estimated distribution of BRMSEA actually underestimates its true sampling variability. The means, however, were similar between estimators, showing greater variability at smaller sample sizes, especially for model SEM-B. Compared with $BRMSEA^{PPMC}$ (Hoofs et al., 2018), we see that $BRMSEA^{DevM}$ is less sensitive to sample size and provides similar information about model fit as its MLE counterpart. This contrasts with Hoofs et al. (2018), who showed that their $BRMSEA^{PPMC}$ under noninformative priors provided much lower values than RMSEA under MLE, failing to detect misspecification in smaller samples. In Table 7 we see the linear relation between the PPP_{χ^2} and the approximate Bayesian fit indices, it presents the stronger relation with $\hat{\Gamma}$, and the lowest relation with NFI.

We found the same patterns of results for other fit indices, which can be seen in plots available on the Open Science Framework.¹² Because only RMSEA has analytically derived confidence limits under MLE, we represented sampling variability of the other MLE fit indices by the 5th and 95th percentiles from their Monte Carlo distributions in each condition.

Limitations of the Monte Carlo Design

We can conclude from our Monte Carlo study that the BSEM fit indices proposed here can be expected to provide similar information about data-model fit as their frequentist counterparts do, at least in the special case of effectively equivalent models under MLE and MCMC with noninformative priors. But there are some important limitations worth noting.

First, our Monte Carlo study utilized only D_i^{obs} and $D(\hat{\theta})$ as defined by the marginal likelihood of the data, not the conditional likelihood, which is consistent with the chi square (likelihood ratio) test statistic in SEM. Bayesian latent variable models can directly sample latent variables through data augmentation (Merkle, Furr, & Rabe-Hesketh, n.d.; Merkle & Rosseel, 2018; Song & Lee, 2012), and the conditional likelihood treats the latent variable scores as parameters (called *person parameters* in the item-response theory framework). In contrast, the latent variable scores can be integrated out to yield the marginal likelihood.

¹² Online files associated with this project are available at <https://osf.io/afkew/>

Table 2
Comparison of Alternative Estimators of the Effective Number of Parameters (pD)

Misfit	Model types	q	pD_{LOO}	pD_{WAIC}	pD_{DIC}
Level 0 (none)	CFA-A	51	50.01	49.90	49.42
	CFA-B	48	46.05	45.96	45.31
	SEM-A	44	43.17	43.06	26.49
	SEM-B	23	22.62	22.57	21.96
Level 1	CFA-A	50	49.21	49.10	48.23
	CFA-B	46	44.75	44.67	44.05
	SEM-A	42	41.10	41.00	19.89
	SEM-B	22	21.61	21.57	20.96
Level 2	CFA-A	49	48.48	48.38	47.59
	CFA-B	45	44.55	44.48	38.15
	SEM-A	41	40.53	40.43	11.24
	SEM-B	19	19.28	19.24	18.78

Note. q = number of parameters estimated in the analogous model fitted with maximum likelihood estimation; LOO = leave-one-out information criterion; WAIC = widely applicable information criterion (or Watanabe's AIC); DIC = deviance information criterion; pD = effective number of parameters; Misfit = level of misspecification.

Information criteria based on marginal likelihoods have been shown to behave more consistently with expectations across conditions than those based on conditional likelihoods (Merkle et al., n.d.), and the interpretation applies more generally to any new data point rather than only to new observed data from cases with the same factor scores. We therefore considered marginal likelihood to be preferable (a view shared by developers of BSEM software such as *Mplus* and *blavaan*), but future researchers might find it reasonable to investigate the sampling properties of BSEM fit indices calculated under conditional likelihoods.

Additionally, we simulated only complete data. Missing data are quite common in applied research (Enders, 2010; Little, 2013), so more research is needed about how the proposed fit indices behave with incomplete data. Because BSEM software uses data augmentation to deal with missing data (Merkle & Rosseel, 2018; Muthén & Muthén, 1998–2017), the proposed fit indices can still be estimated with incomplete data. Missing values are treated as unknown parameters to be estimated at each iteration of the Markov chain, effectively imputing the data before calculating the likelihood (Merkle, 2011). This would make the effective number of parameters pD larger, but *blavaan* marginalizes the missing data by integrating it out, just as it does with latent variable scores, so the pD it reports should be relatively unaffected by missing data. The χ^2_{ML} statistic and fit indices based on it have been shown to indicate better fit when incomplete data are analyzed using full-information maximum likelihood (FIML; Davey, 2005), but it remains to be seen whether multiple imputation in SEM or data augmentation in BSEM affect fit measures the same way. The numerous other factors that could impact BSEM fit indices (e.g., proportion missing data, fraction of missing information, mechanisms of missingness) warrant deeper attention than would fit within the scope of the current article, but two of our illustrative examples explore some effects of missing data on fit measures.

Finally, our Monte Carlo study investigated only noninformative priors because the variety of ways that informative priors could effect BSEM fit indices would have required an unwieldy number of design factors (thus, these issues warrant their own investigation). Given how commonly informative priors are applied (van de Schoot et al., 2017), it is important to understand how

the proposed indices can be expected to behave. In general, we expect priors would only have a noticeable effect on the posterior distribution in small to moderate samples, because in larger samples, the likelihood overwhelms the prior. This is consistent with results reported by Hoofs et al. (2018). But this also depends on how informative the priors are (Gelman et al., 2017). Greater precision will shrink pD , which can affect fit to different degrees depending on how closely the prior matches the actual parameter (Muthén & Asparouhov, 2012). Equation 19 shows that pD cancels out in our “DevM” formulation of BRMSEA, whereas in the “PPMC” formulation (Equation 16) D_i^{obs} is rescaled by D_i^{rep} . Thus, the shrinking of pD by increasing the precision of prior distributions should have differential effects on the different proposals for BRMSEA (and other χ^2 -based fit indices). We explore some assumptions using illustrative examples, and we encourage future researchers to take these issues into account when designing Monte Carlo studies to investigate the effect of informative priors on BSEM fit indices, as well as comparing the “PPMC” with “DevM” formulations.

Illustrative Examples

We encourage future Monte Carlo research into the issues involving informative priors and missing data discussed above, and our illustrative examples serve as preliminary investigations into the details we expect to warrant immediate attention. We utilize the popular Holzinger and Swineford (1939) data set, available¹³ in the R packages *lavaan* (Rosseel, 2012) and *blavaan* (Merkle & Rosseel, 2018). The data set consists of mental ability test scores of $N = 301$ seventh- and eighth-grade children from two different schools. The CFA consisted of three latent cognitive-ability constructs (visual, textual, and speed), each of which was defined by three indicators.

The model was estimated¹⁴ with MLE in *lavaan* (Rosseel, 2012), and several Bayesian models were estimated with the No-U-Turn Sampler (NUTS)—an extension to Hamiltonian Monte

¹³ A description and path diagram of the model can also be found on the *lavaan* tutorial: <http://lavaan.ugent.be/tutorial/cfa.html>

¹⁴ The R scripts to replicate this analysis are available on <https://osf.io/afkew/>

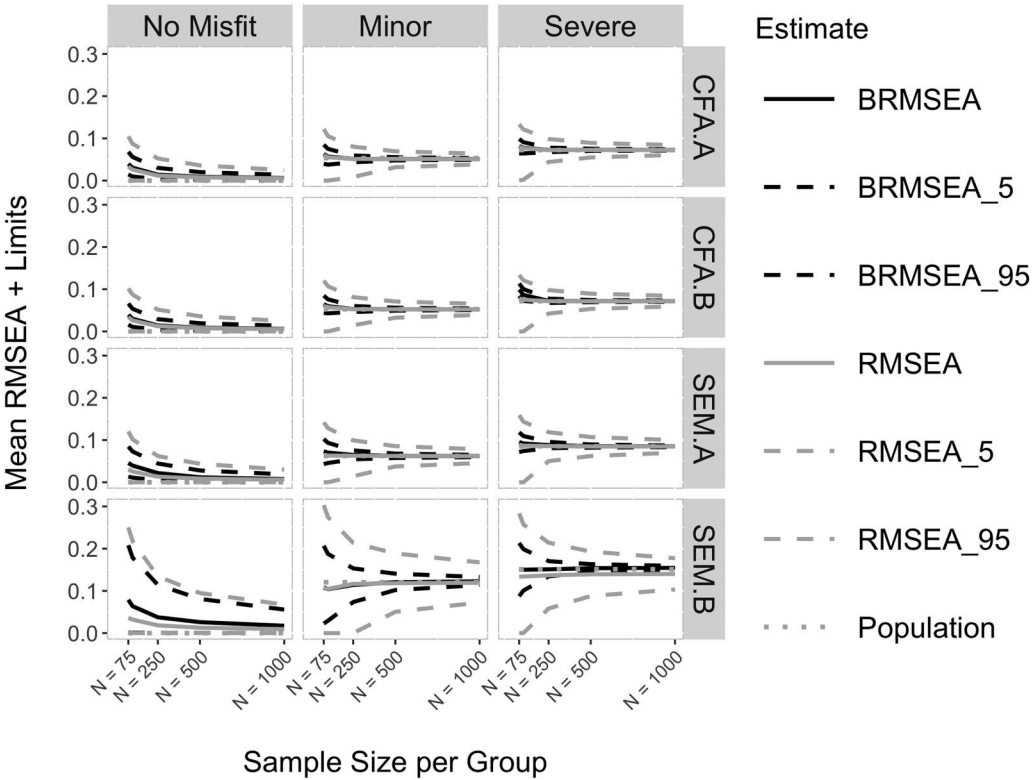


Figure 5. (B)RMSEA mean (solid lines) and average 90% confidence (or credible) bounds (dashed lines) across simulation conditions. The dotted gray line represents the population RMSEA (ϵ) in that condition.

Carlo (Gelman et al., 2014; Hoffman & Gelman, 2014)—with Stan (Carpenter et al., 2017; Stan Development Team, 2018) as the general Bayesian software employed by *blavaan* (Merkle & Rosseel, 2018) in the back end. For both MLE and MCMC (NUTS), we identified the latent scales and locations by fixing factor variances to 1 and factor means to 0. For MCMC estimation, model parameters were sampled using three chains, discarding 10,000 burn-in iterations from each, and retaining 10,000 post-burn-in iterations from each chain. Model convergence was as-

sessed by inspecting traceplots for adequate mixing chains and verifying $R\text{-hat} < 1.10$ for every parameter (Brooks & Gelman, 1998). For all models, we calculated BSEM fit indices using pD_{LOO} as the measure of effective number of parameters. The number of parameters estimated with MLE was always 30 ($df = 24$), and different Bayesian models yielded different estimates of the effective number of parameters.

Results using *blavaan*'s default priors (except for residual standard deviations, where the default prior is $\sim \Gamma(1, .5)$ for the

Table 3
Comparison of Mean (SD) Fit Indices Using Maximum Likelihood and MCMC Estimation

Index	MLE (SD)	MCMC (SD)	ΔMean (SD)	Cohen's d	% Overlap	r
χ^2	128.3104 (119.784)	132.3457 (119.277)	-4.0352 (5.874)	.6869	98.21	.9988
PPP χ^2	.167 (.289)	.198 (.248)	-.030 (.095)	.322	81.89	.936
RMSEA	.0607 (.044)	.0664 (.044)	-.0056 (.012)	.4465	93.44	.9585
$\hat{\Gamma}$.9688 (.028)	.9635 (.029)	.0052 (.007)	.6598	99.73	.9627
$\hat{\Gamma}_{adj}$.9445 (.054)	.9334 (.058)	.0111 (.026)	.4221	99.26	.8924
Mc	.9072 (.091)	.8928 (.094)	.0144 (.023)	.6237	99.19	.9692
CFI	.9713 (.027)	.9670 (.030)	.0043 (.007)	.5476	99.76	.9663
TLI	.9580 (.046)	.9512 (.053)	.0068 (.017)	.3997	99.51	.9490
NFI	.9367 (.050)	.9323 (.053)	.0043 (.007)	.5888	99.74	.9921

Note. ΔMean = MLE - MCMC; % Overlap = percentage of the MLE sampling distribution that overlaps with the MCMC sampling distribution; d = standardized ΔMean (Cohen's d); r = the Pearson correlation between indices under MLE and MCMC; MLE = maximum likelihood estimation; SD = Standard deviation; MCMC = Markov Chain Monte Carlo estimation; χ^2 = chi-square; PPP χ^2 = chi-square based posterior predictive p -value; RMSEA = root mean square error of approximation; $\hat{\Gamma}$ = gamma-hat; $\hat{\Gamma}_{adj}$ = adjusted gamma-hat; Mc = McDonald's centrality index; CFI = comparative fit index; TLI = Tucker-Lewis index; NFI = normed fit index.

Table 4

Percentage of Variance (η^2) in Fit Indices Accounted for by Simulation Conditions

Condition	χ^2	RMSEA	$\hat{\Gamma}$	$\hat{\Gamma}_{adj}$	Mc	CFI	TLI	NFI
MLE								
(MI)	22.7	49.5	61.5	49.7	49.8	53.0	39.6	17.3
(N)	24.8	.9	1.9	1.2	1.8	2.1	.9	36.1
(MT)	24.6	15.5	2.1	5.2	14.3	7.8	13.2	23.9
MI \times N	17.3	1.2	.0	.0	.0	.1	.0	.0
MI \times MT	2.5	5.8	.3	3.8	5.2	3.9	7.5	1.5
N \times MT	2.9	.2	.7	.3	.8	.7	.3	11.0
MI \times N \times MT	2.1	.1	.0	.0	.0	.0	.0	.1
MCMC								
(MI)	23.9	44.7	58.6	39.6	46.5	49.5	35.4	16.6
(N)	23.9	2.9	8.6	5.9	7.2	7.0	3.3	40.6
(MT)	24.8	24.0	1.7	12.3	15.7	8.6	17.0	20.8
MI \times N	17.8	2.1	.1	.4	.1	.1	.2	.0
MI \times MT	2.4	3.5	.8	1.7	5.4	4.6	6.9	1.7
N \times MT	2.4	.1	1.2	.2	2.0	1.4	.5	10.5
MI \times N \times MT	1.7	.9	.9	1.2	.6	1.1	.7	.2
								PPP χ^2
(MI)								66.6
(N)								8.4
(MT)								.9
MI \times N								11.5
MI \times MT								.8
N \times MT								.8
MI \times N \times MT								.6

Note. For rows under MCMC, the column labels should be understood as representing their Bayesian analogs (e.g., $D(\bar{\theta})$ in place of χ^2 , BRMSEA in place of RMSEA, and so on). MLE = Maximum Likelihood Estimation; MCMC = Markov Chain Monte Carlo estimation; χ^2 = chi-square; PPP χ^2 = chi-square based posterior predictive p -value; RMSEA = root mean square error of approximation; $\hat{\Gamma}$ = gamma-hat; $\hat{\Gamma}_{adj}$ = adjusted gamma-hat; Mc = McDonald's centrality index; CFI = comparative fit index; TLI = Tucker-Lewis index; NFI = normed fit index; MI = misspecification; MT = model type; N = sample size.

residual variance) are presented in Table 8, which compares our proposed fit indices with those using the PPMC method proposed by Hoofs et al. (2018) as well as the ML fit indices. These results provide no new information beyond the conclusions of the Monte Carlo study, but they provide users with a real-data example with accompanying syntax in our online OSF materials. They also serve as a baseline to highlight how informative priors and missing data can affect the resulting posterior distributions of the proposed BSEM fit indices in subsequent illustrative examples. We summarize our expectations of these effects below:

- We expect results not to differ between alternative noninformative prior specifications.
- We expect sufficiently informative priors (i.e., informative enough not to be overwhelmed by the likelihood) that are consistent with the data to yield narrower intervals than

noninformative priors. When the posterior distributions of estimated parameters vary less, so will D^{obs} . However, when the posterior is influenced by the prior, pD will be lower, which could affect the expected values of some fit indices (see Equations 19 and 20–25). Holding other quantities constant, smaller pD should yield smaller BRMSEA^{DevM} and BNFI^{DevM} (indicating better and worse fit, respectively) but larger B- $\hat{\Gamma}_{adj}^{DevM}$ (indicating better fit), whereas the posterior means of B- $\hat{\Gamma}^{DevM}$ and BCFI^{DevM} should be unaffected. Given where pD_H appears twice in Equation 23, how BTLI^{DevM} might be affected would probably depend on the average magnitude of D_H^{obs} relative to p^* .

- We expect that if an informative prior is not consistent with the data, the BSEM fit indices will indicate poorer

Table 5

PPP χ^2 and Bayesian Approximate Fit Indices Across Misspecification Conditions

Index	Level 0		Level 1		Level 2	
	Mean	90% CI	Mean	90% CI	Mean	90% CI
PPP χ^2	.469	[.122, .807]	.095	[.000, .463]	.028	[.000, .181]
RMSEA	.028	[.001, .078]	.072	[.037, .141]	.098	[.062, .171]
$\hat{\Gamma}$.990	[.961, .999]	.965	[.927, .987]	.936	[.888, .964]
$\hat{\Gamma}_{adj}$.979	[.920, .999]	.933	[.853, .976]	.889	[.799, .942]
Mc	.970	[.869, .999]	.895	[.764, .977]	.813	[.641, .936]
CFI	.991	[.965, .999]	.969	[.934, .989]	.940	[.884, .977]
TLI	.990	[.947, 1.011]	.950	[.874, .988]	.913	[.817, .969]
NFI	.958	[.859, .998]	.934	[.833, .981]	.905	[.793, .969]

Note. These results marginalize over sample-size and model-type conditions. 90% CI = the average lower and upper bounds. PPP χ^2 = chi-square based posterior predictive p -value; RMSEA = root mean square error of approximation; $\hat{\Gamma}$ = gamma-hat; $\hat{\Gamma}_{adj}$ = adjusted gamma-hat; Mc = McDonald's centrality index; CFI = comparative fit index; TLI = Tucker-Lewis index; NFI = normed fit index; CI = credible interval.

Table 6
PPP χ^2 Across Misspecification and Sample Size Conditions

Misspecification levels	Sample size	Mean	90% CI	% PPP χ^2 < .01
Level 0	75	.453	[.105, .799]	.125
	100	.447	[.105, .795]	.100
	250	.466	[.120, .804]	.075
	500	.487	[.135, .818]	.075
	1,000	.494	[.161, .814]	.025
Level 1	75	.249	[.011, .596]	4.475
	100	.192	[.006, .537]	7.425
	250	.034	[.000, .172]	57.650
	500	.002	[.000, .008]	95.300
	1,000	.000	[.000, .000]	100
Level 2	75	.091	[.000, .353]	28.575
	100	.050	[.000, .227]	44.700
	250	.001	[.000, .002]	98.575
	500	.000	[.000, .000]	100
	1,000	.000	[.000, .000]	100

Note. These results marginalize over model-type conditions. 90% CI = the average lower and upper bounds. PPP χ^2 = chi-square based posterior predictive p -value.

data–model correspondence (e.g., higher BRMSEA and lower BCFI). This follows from priors being part of a Bayesian model, and sufficiently informative priors place restrictions on the parameter space sampled during MCMC estimation, so invalid restrictions should be seen as model misspecifications that yield poorer fit measures.

- We expect a model with small-variance priors for cross-loadings hypothesized to be approximately zero (Muthén & Asparouhov, 2012) to yield better fit than a model that fixes those parameters to zero. This is becoming a common practice in BSEM, raising some concerns about making poor-fitting models appear acceptable (Stromeyer, Miller, Sriramachandramurthy, & DeMartino, 2015), prompting Asparouhov, Muthén, and Morin (2015) to advocate sensitivity analyses in combination with small-variance priors.
- Because blavaan marginalizes over imputed missing values, we expect missing data to have negligible impact on the effective number of parameters pD , but we expect $D(\hat{\theta})$ to be smaller with greater proportions of missing data, as observed by Davey (2005). This phenomenon also held for the independence model (Davey, 2005), making it more difficult to predict the behavior of incremental fit indices.

Noninformative Priors

To verify our first assumption about the effects of priors—that different specifications of noninformative priors yield approximately the same results—we compared the following noninformative priors:

- Default priors in blavaan (Merkle & Rosseel, 2018): Factor loadings and indicator intercepts were distributed as $\sim \mathcal{N}(\mu = 0, \sigma^2 = 100)$; indicator residual standard deviations were distributed as \sim half-Cauchy($\mu = 0, \sigma^2 = 2.5$); and factor correlations were distributed as $\sim U(-1, 1)$.

- Alternative noninformative priors: Factor loadings and indicator intercepts were distributed as $\sim \mathcal{N}(\mu = 0, \sigma^2 = 1,000,000)$; indicator residual standard deviations were distributed as $\sim U(0.01, 1000)$; and factor correlations were distributed as $\sim U(-1, 1)$. These priors are even less informative, to a degree similar to the default priors in Mplus (Muthén & Muthén, 1998–2017). However, we could not exactly replicate all the default priors of Mplus because they place priors on factor covariances, whereas blavaan places priors on factor correlations, which are rescaled to covariances implied by the estimated correlations and (fixed or free) variances (Merkle & Rosseel, 2018).

The default priors yielded $pD_{\text{LOO}} = 32.263$ (similar to 30 estimated parameters under MLE) and $p^* - pD = 21.737$ (similar to $df = 24$ under MLE). The alternative priors (the “Wide” column in Table 9) yielded nearly identical results: $pD_{\text{LOO}} = 32.451$ and $p^* - pD = 21.549$.

We calculated Bayesian fit indices using both the “DevM” formulation proposed here (columns labeled MCMC^{DevM} in Table 8) and the “PPMC” formulation (columns labeled MCMC^{PPMC} in Table 8) based on the BRMSEA proposed by Hoofs et al. (2018). Each index’s distribution is summarized in Table 8 by its mean, standard deviation, and percentiles corresponding to a 90% credible interval. The means using default priors can be compared with their MLE counterparts (also presented in Table 8), and the upper and lower bounds of BRMSEA can be compared with the 90% CI of RMSEA.

Posterior means of MCMC^{DevM}-based fit indices were close to the MLE fit indices, whereas the MCMC^{PPMC}-based fit indices indicated better data–model fit. This is due to the difference in how model discrepancy is rescaled between the “DevM” and “PPMC” formulations (compare Equation 19 with Equation 16), as shown in the row of Table 8 labeled “ χ^2 .” The MLE chi square and MCMC^{DevM} chi-square analog were 85.31 and 83.56, respectively, whereas the average of the difference ($D^{\text{obs}} - D^{\text{rep}}$) used by MCMC^{PPMC} was 61.28. Chi-square-based approximate fit indices using the “PPMC” formulation therefore indicated better data–model fit. If one were to test a H_0 of approximate fit using the guideline RMSEA < 0.08 (MacCallum et al., 1996, 2006), this model would have been rejected using both MLE and

Table 7
Correlations Between PPP χ^2 and Bayesian Approximate Fit Indices

Index	PPP χ^2
RMSEA	-.668
$\hat{\Gamma}$.695
$\hat{\Gamma}_{adj}$.587
Mc	.642
CFI	.611
TLI	.565
NFI	.284

Note. PPP χ^2 = chi-square based posterior predictive p -value; RMSEA = root mean square error of approximation; $\hat{\Gamma}$ = gamma-hat; $\hat{\Gamma}_{adj}$ = adjusted gamma-hat; Mc = McDonald’s centrality index; CFI = comparative fit index; TLI = Tucker-Lewis index; NFI = normed fit index.

Table 8
Holzinger and Swineford (1939) CFA Fit Indices

Index	MLE	MCMC ^{DevM}		MCMC ^{PPMC}	
		Mean (SD)	90% CI	Mean (SD)	90% CI
χ^2	85.306 ^a	83.560 ^b (8.003)	[70.639, 96.251]	61.278 (13.233)	[38.958, 82.242]
RMSEA	.092 [.071, .114]	.097 (.006)	[.086, .107]	.076 (.014)	[.055, .099]
$\hat{\Gamma}$.957	.956 (.005)	[.948, .965]	.972 (.009)	[.956, .986]
$\hat{\Gamma}_{adj}$.919	.892 (.013)	[.870, .913]	.930 (.023)	[.891, .966]
Mc	.903	.902 (.012)	[.884, .922]	.937 (.021)	[.902, .970]
CFI	.931	.930 (.009)	[.916, .945]	.953 (.016)	[.928, .979]
TLI	.896	.887 (.015)	[.864, .911]	.924 (.025)	[.884, .966]
NFI	.907	.909 (.009)	[.895, .923]	.931 (.015)	[.906, .955]

Note. Row names under the Index column should be understood as representing their Bayesian analogs (e.g., $D(\bar{\theta})$ in place of χ^2 , BRMSEA in place of RMSEA, and so on). Under the MCMC^{PPMC} formulation, the mean of $(D_i^{bs} - D_i^{cp})$ is reported in place of χ^2 . 90% CI = the average lower and upper bounds. For MLE, only RMSEA is accompanied by its 90% confidence bounds. χ^2 = chi-square; SD = standard deviation; CI = credible interval; PPP χ^2 = chi-square based posterior predictive p -value; RMSEA = root mean square error of approximation; $\hat{\Gamma}$ = gamma-hat; $\hat{\Gamma}_{adj}$ = adjusted gamma-hat; Mc = McDonald's centrality index; CFI = comparative fit index; TLI = Tucker-Lewis index; NFI = normed fit index.

^a $p < .001$, with $df = 24$ and $q = 30$ estimated parameters. ^b PPP $\chi^2 < .001$, with $pD_{LOO} = 31.768$ estimated parameters.

BRMSEA^{DevM}, yet accepted using BRMSEA^{PPMC}. Using the 90% CI, a H_0 of poor fit (RMSEA > .10) would be rejected using BRMSEA^{PPMC} but not by RMSEA or BRMSEA^{DevM}, so the latter should prompt researchers to further investigate whether their model fails in specific ways. These results with our nine-indicator model were consistent with the simulation results in Hoofs et al. (2018), which showed that in small to moderate samples, BRMSEA^{PPMC} was similar to RMSEA in six-indicator models but was much smaller than RMSEA in 12-indicator models.

Specifying alternative noninformative priors yielded nearly identical results. These examples further support that MCMC^{DevM} fit indices are reasonable approximations of their MLE counterparts, but with narrower intervals (see Table 8, Figure 5). In contrast, MCMC^{PPMC} fit indices indicate a more positive (perhaps misleading) impression of data-model fit.

Small-Variance Priors for Cross-Loadings

Holzinger and Swineford (1939) posited a bifactor solution from their exploratory factor analysis results, in which the three factors

of our example model would be interpreted as orthogonal method factors, and all indicators would also load on a single general-intelligence factor. However, a hypothetical researcher might be interested in retaining approximately simple structure by specifying a Bayesian model that more flexibly represents theoretical expectations. Following Muthén and Asparouhov (2012), we specified the same model with default priors but also added all possible cross-loadings with the constraint that they were approximately zero (i.e., specifying a normal prior with $\mu = 0$ and $\sigma^2 = 0.01$). The standard deviations of the observed variables ranged from 1.01 to 1.29, so (similar to Muthén & Asparouhov, 2012) these priors represents a hypothesized 95% probability that approximately standardized cross-loadings are within ± 0.2 of zero.

As expected, adding small-variance priors for cross-loadings hypothesized to be approximately zero yielded fit measures that indicated better data-model correspondence. Comparing the "Cross" column in Table 9 with the "Wide" column (the latter of which is essentially equivalent to the MCMC^{DevM} results in Table 8), PPP $\chi^2 = 0.17$ (compared with $<.001$) and BRMSEA^{DevM} =

Table 9
Holzinger and Swineford (1939) CFA Fit Indices With Different Priors

Index	Wide		Cross		Orth		Strict	
	Mean	90% CI	Mean	90% CI	Mean	90% CI	Mean	90% CI
PPP χ^2	<.001		.169		<.001		<.001	
pD_{LOO}	32.451		40.835		29.607		22.837	
$D(\bar{\theta})$	83.319	[70.865, 96.048]	27.322	[11.634, 42.833]	122.651	[107.211, 138.635]	163.988	[141.420, 186.359]
BRMSEA	.097	[.087, .107]	.056	[.018, .094]	.116	[.106, .125]	.119	[.109, .129]
$\hat{\Gamma}$.956	[.948, .965]	.990	[.981, 1.000]	.932	[.922, .942]	.911	[.897, .924]
$\hat{\Gamma}_{adj}$.891	[.869, .912]	.957	[.920, 1.000]	.850	[.828, .872]	.845	[.821, .869]
BMc	.903	[.884, .921]	.977	[.956, 1.000]	.850	[.827, .871]	.802	[.772, .832]
BCFI	.930	[.916, .944]	.984	[.970, 1.000]	.889	[.871, .907]	.850	[.824, .875]
BTLI	.885	[.862, .909]	.957	[.909, 1.004]	.839	[.813, .865]	.829	[.801, .859]
BNFI	.909	[.896, .923]	.970	[.953, .987]	.866	[.849, .884]	.821	[.797, .846]

Note. Wide = wider-than-default noninformative priors; Cross = strongly informative priors for near-zero cross-loadings; Orth = strongly informative priors for near-zero factor correlations; Strict = strongly informative priors for estimated parameters; 90% CI = the average lower and upper bounds; PPP χ^2 = chi-square based posterior predictive p -value; BRMSEA = Bayesian root mean square error of approximation; $\hat{\Gamma}$ = Bayesian gamma-hat; $\hat{\Gamma}_{adj}$ = Bayesian adjusted gamma-hat; BMc = Bayesian McDonald's centrality index; BCFI = Bayesian comparative fit index; BTLI = Bayesian Tucker-Lewis index; BNFI = Bayesian normed fit index; pD_{LOO} = leave-one-out based effective number of parameters; $D(\bar{\theta})$ = deviance.

0.056 (compared with 0.097) indicated good fit. However, adding 18 cross-loadings to the model—even though constrained to approximately zero—increased the effective number of parameters from $pD \approx 32$ to $pD \approx 41$ (so each cross-loading effectively added approximately half an estimated parameter). A consequence of estimating more parameters is increased variability of the resulting posterior distribution as well as the fit indices derived from it. The width of $\text{BRMSEA}^{\text{DevM}}$'s interval increased from $0.107 - 0.086 = 0.021$ to $0.094 - 0.018 = 0.076$.

Strongly Informative Priors for Orthogonal Correlations

The estimated factor correlations were in the approximate range of .30 to .50 (medium to large; Cohen, 1992), so constraining them to zero would be a gross model misspecification. Thus, if we specified strongly informative priors that constrained the factor correlations to be approximately (but not exactly) zero, those priors would be an aspect of the model that is misspecified, which should be reflected by fit measures. We would expect (a) pD to decrease by no more than three parameters, and (b) PPP_{χ^2} and approximate fit indices to show worse model fit than an otherwise equivalent model with uninformative priors for factor correlations (i.e., the results from Table 8).

To verify this expectation, we fit a model with default priors for all parameters except factor correlations, for which we specified informative priors following a rescaled Beta(200, 200) distribution. The standard boundaries of the Beta distribution from 0 to 1 are rescaled by multiplying by 2 and subtracting 1, effectively setting new boundaries from -1 to 1 (same as correlations). These priors reflect the belief that these factor correlations have a 95% chance of being within ± 0.1 of 0.

The “Orth” column of Table 9 shows the results for the model with orthogonality constraints. As expected, pD decreases from 31.77 with default priors to 29.61 (about 0.72 for each of three constrained correlations). All approximate fit indices indicated worse fit after constraining correlations, for example, $\text{BRMSEA}^{\text{DevM}} = 0.116$ for approximate orthogonality compared with $\text{BRMSEA}^{\text{DevM}} = 0.097$ with default priors. Their intervals were not noticeably more precise after specifying informative priors for factor correlations, so misspecified priors do not appear to translate to increased precision for fit indices in this case, but future simulation research could test whether and under what conditions this could be expected.

Strongly Informative Priors for Reproducing Results

The previous examples investigated informative priors for nuisance parameters and for a substantive hypothesis that was not consistent with the data. We also consider the case of informative priors based on previously obtained estimates. To imitate a situation in which the priors are maximally consistent with the observed data, we specified priors based on the posterior distribution from the model with default priors. Priors for factor loadings and item intercepts were $\sim \mathcal{N}(\mu = \bar{\theta}, \sigma = 0.1)$, and for residual standard deviations were $\sim \text{Log-}\mathcal{N}(\mu = \bar{\theta}, \sigma = 0.1)$, where $\bar{\theta}$ is the estimated posterior mean from the model with default priors, and the standard deviations specified similar precision as for the model with near-zero cross-loadings. Thus, the model was specified with

high certainty that the parameters would be consistent with estimates from the model with default priors (which we treated as results from a “previous” sample). In practice, this is what Gelman et al. (2017) would consider “cheating” because the $\mu = \bar{\theta}$ was specified after seeing the data instead of reflecting *prior* theoretical/probabilistic beliefs about the expected distribution.

The column “Strict” in Table 9 presents the results. As expected, pD decreases substantially, from 31.7 (default priors) to 22.8. Note that this decrease results solely from estimating all the same parameters but with greater precision (i.e., prior variances $\sigma^2 = 0.01$ vs. $\sigma^2 = 100$). On average, pD decreased by 0.33 per estimated parameter. Contrary to expectation, all fit indices presented worse model fit. So despite the priors being set around the previously estimated posterior means, $D(\bar{\theta})$ was also affected, possibly because the priors did not allow the MCMC sampler to sufficiently explore the parameter space. However, the posterior distributions of model parameters estimated with default and strict priors still overlapped substantially: between 85.7% and 99.5% across parameters, with an average overlap of 95.2% ($SD = 3.5\%$). It is also notable that contrary to expectation, the strict priors did not yield narrower intervals for fit indices than noninformative priors did.

We are unsure whether these results would be observed in a model that fit the data well; given the mediocre fit of this model, placing great certainty in the wrong parameters might merely exacerbate evidence against data–model fit. This would be consistent with the expected behavior of a marginal likelihood with constrained priors, which yield an inappropriate marginal likelihood. This is reflected in a marginal likelihood sensitive to aspects of the prior distribution that have minimal effect on the posterior inferences (Gelman et al., 2017), which could imply the approximate fit indices have the potential to diagnose if prior distributions are affecting the marginal likelihood in such a way that would make it unreliable.

The effects of informative priors should be even more pronounced when they have a greater relative influence on the posterior, as would be the case with small samples. We drew a random sample of $N = 75$ from the full data set and fit the model with default priors, wide priors, and strict priors to the subsample. Consistent with our simulation results, Table 10 shows wider intervals for all models (reflecting less information from data), and the different noninformative priors yielded effectively the same posterior means (consistent with MLE across sample sizes). Again, the strict priors yielded worse fit, but the effect of informative priors appears even more extreme than with the full sample. Unfortunately, the data–model correspondence already appears much poorer for this subsample than the full $N = 301$ in Table 9, which might be due (at least in part) to small-sample bias (Jiang & Yuan, 2017; Nevitt & Hancock, 2004). Future simulation studies could verify whether our original expectation (same average fit with smaller intervals) holds when the model accurately represents the population.

Incomplete Data

We used the R package `simsem` (Jorgensen, Pornprasertmanit, Miller, & Schoemann, 2018) to set 20% and 50% of values on all the indicators in the Holzinger and Swineford (1939) data to be missing completely at random (Enders, 2010). We fit the model

Table 10

CFA Fit Indices for Random Subset of $N = 75$ From *Holzinger and Swineford (1939)* Data

Index	Default		Wide		Strict	
	Mean	90% CI	Mean	90% CI	Mean	90% CI
PPP χ^2	.003		.003		<.001	
pD_{LOO}	32.382		31.623		16.604	
$D(\theta)$	63.640	[50.410, 77.144]	64.223	[51.411, 77.690]	137.783	[122.228, 152.542]
BRMSEA	.160	[.134, .186]	.157	[.132, .182]	.189	[.175, .203]
$\hat{\Gamma}$.890	[.859, .921]	.890	[.859, .921]	.771	[.744, .797]
$\hat{\Gamma}_{\text{adj}}$.724	[.647, .804]	.735	[.660, .809]	.669	[.630, .707]
BMc	.757	[.689, .824]	.758	[.692, .824]	.513	[.461, .564]
BCFI	.822	[.765, .879]	.823	[.767, .879]	.575	[.509, .642]
BTLI	.714	[.623, .806]	.725	[.639, .812]	.605	[.544, .668]
BNFI	.765	[.714, .814]	.763	[.714, .812]	.491	[.432, .550]

Note. Default = default noninformative priors in blavaan; Wide = wider-than-default noninformative priors; Strict = strongly informative priors for estimated parameters; 90% CI = the average lower and upper bounds; PPP χ^2 = chi-square based posterior predictive p -value; BRMSEA = Bayesian root mean square error of approximation; $\hat{\Gamma}$ = Bayesian gamma-hat; $\hat{\Gamma}_{\text{adj}}$ = Bayesian adjusted gamma-hat; BMc = Bayesian McDonald's centrality index; BCFI = Bayesian comparative fit index; BTLI = Bayesian Tucker-Lewis index; BNFI = Bayesian normed fit index; pD_{LOO} = leave-one-out based effective number of parameters; $D(\theta)$ = deviance.

using FIML and MCMC using default noninformative priors. Results are presented in Table 11, which repeats the complete-data MCMC results from Table 8 to ease comparison. As expected, missing data have minimal impact on pD given that imputed values are integrated out to yield a marginal likelihood (Merkle & Rosseel, 2018; Muthén & Muthén, 1998–2017). Consistent with Davey (2005), absolute fit measures indicated better fit with greater proportions of missing data.

Contrary to Davey (2005), we observed that incremental fit indices indicated better fit for 20% missing than for complete data yet worse fit for 50% missing than for 20% missing. Our model was similar to the simulated models of Davey's Monte Carlo study, but we imposed missing values on all indicators rather than only three, and our model fit the Holzinger and Swineford (1939) data worse than most of the conditions that Davey simulated. Table 11 includes χ^2_0 for the independence model M_0 to help make sense

of the results for incremental fit indices. With greater proportions of missing data, the fit for M_0 improves proportionally more so than the fit for the hypothesized model, so the hypothesized model's fit *relative* to M_0 is not as high with 50% missing data.

Discussion

We proposed how chi-square-based SEM fit indices can be calculated in BSEM using a Bayesian analog of the chi-square statistic: $D(\theta)$. Because $D(\theta) \approx \chi^2$, it can be used to calculate measures of approximate fit that are commonly used in SEM to complement the chi-square test of exact fit of a model to data. We compared these BSEM fit indices with their frequentist counterparts through a Monte Carlo simulation study, which verified our expectation that MCMC with noninformative priors yields similar results to MLE across levels of misspecification, sample sizes, and

Table 11

Holzinger and Swineford (1939) CFA Fit Indices With Missing Data

MCAR:	0%		20%	MCMC	50%		MCMC	90%	
Index	Mean	90% CI	FIML	Mean	90% CI	FIML	Mean	90% CI	
p value	<.001		.001	.020		.021	.160		
pD_{LOO}	31.768		$q = 30$	32.662		$q = 30$	32.965		
χ^2	83.560	[70.639, 96.251]	51.380	49.440	[35.931, 61.635]	39.981	37.555	[24.769, 49.327]	
χ^2_0	918.302	[908.713, 927.773]	621.833	620.797	[610.834, 630.120]	296.164	294.886	[284.615, 303.880]	
RMSEA	.097	[.086, .107]	.062	.065	[.050, .081]	.047	.049	[.030, .070]	
$\hat{\Gamma}$.956	[.948, .965]	.980	.980	[.971, .989]	.988	.988	[.979, .997]	
$\hat{\Gamma}_{\text{adj}}$.892	[.870, .913]	.963	.949	[.927, .973]	.978	.969	[.947, .992]	
Mc	.902	[.884, .922]	.956	.954	[.935, .976]	.974	.973	[.953, .993]	
CFI	.930	[.916, .945]	.953	.952	[.931, .975]	.939	.936	[.889, .984]	
TLI	.887	[.864, .911]	.930	.922	[.888, .959]	.908	.896	[.818, .973]	
NFI	.909	[.895, .923]	.917	.920	[.900, .941]	.865	.873	[.832, .916]	

Note. Row names under the Index column should be understood as representing their Bayesian analogs (e.g., $D(\theta)$ in place of χ^2 , PPP χ^2 in place of the p value, BRMSEA in place of RMSEA, and so on). 90% CI = the average lower and upper bounds; MCAR = percentage of data points missing completely at random; MCMC = "DevM"-based Bayesian fit indices; FIML = full-information maximum likelihood based fit indices; χ^2 = chi-square; χ^2_0 = null model chi-square; RMSEA = root mean square error of approximation; $\hat{\Gamma}$ = gamma-hat; $\hat{\Gamma}_{\text{adj}}$ = adjusted gamma-hat; Mc = McDonald's centrality index; CFI = comparative fit index; TLI = Tucker-Lewis index; NFI = normed fit index; pD_{LOO} = leave-one-out based effective number of parameters; q = MLE number of parameters.

model types. Because we apply the same formulas to analogous Bayesian and frequentist quantities, we opine that traditional guidelines proposed for interpreting the magnitude of SEM fit indices based on intuition and experience (Bentler & Bonett, 1980; Browne & Cudeck, 1992) would be no less valid to apply to BSEM fit indices applied using the “DevM” formulation, even when informative priors are specified.

An advantage of BSEM fit indices over most of their frequentist counterparts is that the posterior distribution allows uncertainty to be quantified for any fit index, although the $\text{BRMSEA}^{\text{DevM}}$ intervals were narrower than the CIs of RMSEA under MLE. So researchers who insist on interpreting fit indices from a hypothesis-testing perspective (MacCallum et al., 1996, 2006) should not expect previous simulation results that provided guidelines based on Type I and II error rates (Hu & Bentler, 1998, 1999; and for invariance testing: Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008) to generalize to the BSEM fit indices proposed here. Recall, though, that much research has already revealed that fixed cutoffs do not generalize well in frequentist SEM (Beauducel & Wittmann, 2005; Davey, 2005; Fan & Sivo, 2005; Heene et al., 2011, 2012; Jorgensen, Kite, et al., 2018; Marsh et al., 2004; Pornprasertmanit, 2014; Pornprasertmanit et al., 2013; Sass et al., 2014), so we do not advocate their use (see below for discussion of alternatives).

Consistent with previous research (Fan & Sivo, 2007), $\hat{\Gamma}$ and CFI were substantially affected ($\eta^2 > 10\%$) only by the level of misspecification, whereas other indices were also affected by model type (and NFI was also affected by sample size). Based on these simulation results, we would therefore recommend the use of CFI and $\hat{\Gamma}$ in applied research as well as their continued investigation in simulation research. Also, we recommend applied researchers use either pD_{LOO} or pD_{WAIC} to calculate BSEM fit indices because pD_{DIC} appeared sensitive to model type.

Future Directions

Our illustrative examples provide preliminary support for some expectations based on intuition and past results. Different noninformative priors yielded similar results. Informative priors for nuisance parameters (Asparouhov et al., 2015; Muthén & Asparouhov, 2012; Stromeier et al., 2015) yielded fit indices that indicated better fit. Informative priors that were inconsistent with the data yielded fit indices that indicated worse fit. However, informative priors that were consistent with the data also yielded fit indices that indicated worse fit, which might be due to the lack of data–model correspondence in our Holzinger and Swineford (1939) example or a diagnosis of the priors provoking the marginal likelihood to become inappropriate due to the strong constraints (Gelman et al., 2017). Finally, missing data yielded absolute fit indices that indicated better fit, but incremental fit indices only indicated better fit when the majority of data were observed (20% missing) because the M_0 ’s fit seemed to improve proportionally more than M_H ’s fit when 50% of the values were set to missing (a result not observed with FIML; Davey, 2005). Before more general guidelines can be provided for their use in evaluating a wider array of BSEM models, future Monte Carlo research must be designed to investigate whether these assumptions hold under various condi-

tions and to further probe the causes of some unexpected results.

Gelman et al. (2017) classified the relevance and use of priors accordingly: minimalist (or “noninformative”), reference, structural, weakly informative, and regularizing. The *minimalist* and *reference* types do little more than fulfill the requirement of specifying a prior for Bayesian inference, but with no intention of priors providing information about the model parameters. *Structural* priors set a structural form for the related parameters, without guiding the expected values. *Regularizing* priors represent a strong assumption about the model, where the researcher intends to yield smoother and more stable inferences. Between *structural* and *regularizing* is *weakly informative*, which provides information that applies to a general type of problems without taking full advantage of problem-specific knowledge. In BSEM, the priors are usually in the categories of *minimalist*, *reference*, or *structural*, the latter of which can yield unexpected results regarding model comparison¹⁵ when priors add no information about the model, only varying levels of uncertainty about them (e.g., σ^2). The priors are included in the calculation of the marginal likelihood, which can be affected by priors without affecting the posterior distribution inferences—this is a known issue with multivariate models (Gelman et al., 2017). More research is needed to determine the effects of different types of priors and ways to quantify the effect of priors on the marginal likelihood and the posterior distribution.

We considered chi-square-based fit indices only to assess the practical fit of the model (consistent with their use in MLE). Because their true sampling variability is underestimated by the interval derived from the posterior percentiles, using 90% intervals to test a H_0 about the model would yield inflated Type I errors. Instead, researchers interested in using a fit index to test the model could use it as the discrepancy function for the observed data in a PPMC framework. Because Hoofs et al. (2018) proposed calculating $\text{BRMSEA}_i^{\text{PPMC}}$ using the posterior predictive distribution of D_i^{pp} , one could not subsequently use $\text{BRMSEA}_i^{\text{PPMC}}$ as the discrepancy function in a posterior predictive model check (e.g., the way Levy, 2011, used SRMR as a discrepancy function). On the other hand, $\text{BRMSEA}_i^{\text{DevM}}$ (or other “DevM” fit indices Equations 20–25) could serve as a discrepancy function in a PPMC. For example, the posterior distribution of $\text{BRMSEA}^{\text{DevM}}$ could be approximated using their realized values by plugging D_i^{obs} into Equation 17, whereas plugging D_i^{pp} into Equation 17 in place of D_i^{obs} would approximate its posterior predictive distribution (i.e., expected values given the model parameters). However, because $\text{BRMSEA}_i^{\text{DevM}}$ is simply a function of $D_{H,i}^{\text{obs}}$ and other quantities that do not differ for the observed (or replicated) data (i.e., N , df , p^* , and pD), we can expect PPMC to yield the same conclusions when using $\text{BRMSEA}_i^{\text{DevM}}$ (or any indices based solely on D_H^{obs}) as when using χ_H^2 itself. So we would not expect applying PPMC with a fit index based only on χ_H^2 to add any extra information beyond what PPP_{χ^2} provides.

¹⁵ Recall that an SEM’s chi-square statistic compares M_H with a saturated M_S .

In contrast, incremental fit indices are based not only on χ^2_H but also on χ^2_0 , so a PPMC using Equation 25 could provide distinct information from PPMC using χ^2_H (i.e., How well does M_H fit relative to a meaningfully specified M_0 ; Bentler & Bonett, 1980). Beyond the complication of specifying informative priors for a null model M_0 that is known not to be true (Hoofs et al., 2018), a second complication involves estimating a posterior predictive distribution of incremental fit indices. Calculating a posterior distribution of realized values (e.g., of $BCFI^{DevM}_I$) would require χ^2_0 , which only requires fitting an appropriate M_0 (e.g., an independence model) to the observed data. PPMC would also require fitting the hypothesized model M_H to both observed data and replicated data generated from the sampled estimates (at that iteration) of the model being fitted to the observed data. However, whereas M_0 is also fitted to observed data, M_0 should be fitted to replicated data generated from M_H , so that the model would be tested assuming M_H is true, not M_0 . We expect a deeper exploration of how to conduct PPMC with incremental fit indices (e.g., whether to use the same replicated data generated when fitting M_H) to be a valuable extension of the current proposal.

We conclude by reminding readers that fit indices were designed merely to be descriptive measures meant to help researchers evaluate the degree to which their model fails to reproduce the observed data; they were not originally developed to be test statistics (although they have been applied as such). Global fit indices are meant to complement, not replace, informative tests of the model. Researchers who find values of their fit indices questionable should complement these global measures with more informative local measures of model (mis)fit, such as described in Levy (2011). We are hopeful that fit indices can be conscientiously applied in (B)SEM in conjunction with other model-evaluation tools that summarize different dimensions of data–model correspondence (or lack thereof).

References

- Arbuckle, J. L. (2012). *IBM SPSS Amos 21 user's guide* [Computer software manual]. Chicago, IL: IBM.
- Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeier et al. *Journal of Management*, 41, 1561–1577. <http://dx.doi.org/10.1177/0149206315591075>
- Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, 12, 41–75. http://dx.doi.org/10.1207/s15328007sem1201_3
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246. <http://dx.doi.org/10.1037/0033-2909.107.2.238>
- Bentler, P. M. (2006). *EQS 6 structural equations program manual* [Computer software manual]. Encino, CA: Multivariate Software.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606. <http://dx.doi.org/10.1037/0033-2909.88.3.588>
- Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455. <http://dx.doi.org/10.1080/10618600.1998.10474787>
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83. <http://dx.doi.org/10.1111/j.2044-8317.1984.tb00789.x>
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21, 230–258. <http://dx.doi.org/10.1177/0049124192021002005>
- Carpenter, B., Gelman, A., Hoffman, D. M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 1–32. <http://dx.doi.org/10.18637/jss.v076.i01>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464–504. <http://dx.doi.org/10.1080/10705510701301834>
- Cheng, C., & Wu, H. (2017). Confidence intervals of fit indexes by inverting a bootstrap test. *Structural Equation Modeling*, 24, 870–880. <http://dx.doi.org/10.1080/10705511.2017.1333432>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159. <http://dx.doi.org/10.1037/0033-2909.112.1.155>
- Congdon, P. (2009). Modelling the impact of socioeconomic structure on spatial health outcomes. *Computational Statistics & Data Analysis*, 53, 3047–3056. <http://dx.doi.org/10.1016/j.csda.2007.10.021>
- Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields specified minimizer and specified minimum discrepancy function value. *Psychometrika*, 57, 357–369.
- Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research*, 32, 208–252. <http://dx.doi.org/10.1177/0049124103256130>
- Davey, A. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling*, 12, 578–597. http://dx.doi.org/10.1207/s15328007sem1204_4
- Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, 40, 1–5. <http://dx.doi.org/10.2307/2683105>
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, 12, 343–367. http://dx.doi.org/10.1207/s15328007sem1203_1
- Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, 42, 509–529. <http://dx.doi.org/10.1080/00273170701382864>
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). *Bayesian data analysis* (3rd ed.). London, UK: Chapman and Hall/CRC.
- Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–760.
- Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472. <http://dx.doi.org/10.1214/ss/1177011136>
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66, 8–38. <http://dx.doi.org/10.1111/j.2044-8317.2011.02037.x>
- Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19, 555. <http://dx.doi.org/10.3390/e19100555>
- Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in Mplus and SAS. *Structural Equation Modeling*, 16, 676–701. <http://dx.doi.org/10.1080/10705510903206055>

- Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, 16, 319–336. <http://dx.doi.org/10.1037/a0024917>
- Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling*, 19, 36–50. <http://dx.doi.org/10.1080/10705511.2012.634710>
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1593–1623. Retrieved from <http://jmlr.org/papers/v15/hoffman14a.html>
- Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bifactor solution*. Chicago, IL: University of Chicago Press.
- 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. <http://dx.doi.org/10.1177/0013164417709314>
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453. <http://dx.doi.org/10.1037/1082-989X.3.4.424>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. <http://dx.doi.org/10.1080/10705519909540118>
- Jiang, G., & Yuan, K.-H. (2017). Four new corrected statistics for SEM with small samples and nonnormally distributed data. *Structural Equation Modeling*, 24, 479–494. <http://dx.doi.org/10.1080/10705511.2016.1277726>
- Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.8 for Windows*. Skokie, IL: Scientific Software International.
- Jorgensen, T. D., Kite, B. A., Chen, P.-Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods*, 23, 708–728. <http://dx.doi.org/10.1037/met0000152>
- Jorgensen, T. D., Pornprasertmanit, S., Miller, P., & Schoemann, A. (2018). *simsem: SIMulated structural equation modeling* [Computer software manual]. R package Version 0.5–14. Retrieved from <https://CRAN.R-project.org/package=simsem>
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44, 486–507. <http://dx.doi.org/10.1177/0049124114543236>
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65, 457–474. <http://dx.doi.org/10.1007/BF02296338>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.
- Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14, 293–300. <http://dx.doi.org/10.1016/j.tics.2010.05.001>
- Lai, M. H., & Yoon, M. (2015). A modified comparative fit index for factorial invariance studies. *Structural Equation Modeling*, 22, 236–248. <http://dx.doi.org/10.1080/10705511.2014.935928>
- Levy, R. (2011). Bayesian data–model fit assessment for structural equation modeling. *Structural Equation Modeling*, 18, 663–685. <http://dx.doi.org/10.1080/10705511.2011.607723>
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.
- MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, 38, 113–139. http://dx.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. <http://dx.doi.org/10.1037/1082-989X.11.1.19>
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149. <http://dx.doi.org/10.1037/1082-989X.1.2.130>
- Maiti, S. S., & Mukherjee, B. N. (1990). A note on distributional properties of the Jöreskog–Sörbom fit indices. *Psychometrika*, 55, 721–726. <http://dx.doi.org/10.1007/BF02294619>
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2
- Matthews, R. A. (2001). Why should clinicians care about Bayesian methods? *Journal of Statistical Planning and Inference*, 94, 43–58. [http://dx.doi.org/10.1016/S0378-3758\(00\)00232-9](http://dx.doi.org/10.1016/S0378-3758(00)00232-9)
- McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, 6, 97–103. <http://dx.doi.org/10.1007/BF01908590>
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93, 568–592. <http://dx.doi.org/10.1037/0021-9010.93.3.568>
- Meredith, M., & Ridout, M. (2017). overlap: Estimates of coefficient of overlapping for animal activity patterns [Computer software manual] (R package version 0.3.0). Retrieved from <https://cran.r-project.org/web/packages/overlap/index.html>
- Merkle, E. C. (2011). A comparison of imputation methods for Bayesian factor analysis models. *Journal of Educational and Behavioral Statistics*, 36, 257–276. <http://dx.doi.org/10.3102/1076998610375833>
- Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (n.d.). *Bayesian model assessment: Use of conditional vs marginal likelihoods*. Retrieved from <https://arxiv.org/abs/1802.04452>
- Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85, 1–30. <http://dx.doi.org/10.18637/jss.v085.i04>
- Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, 42, 875–881. <http://dx.doi.org/10.1016/j.paid.2006.09.021>
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23, 103–123. <http://dx.doi.org/10.3758/s13423-015-0947-8>
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313–335. <http://dx.doi.org/10.1037/a0026802>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.) [Computer software manual]. Los Angeles, CA: Author.
- Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, 39, 439–478. http://dx.doi.org/10.1207/S15327906MBR3903_3
- Plummer, M. (2017). *JAGS 4.3.0 User manual*. Retrieved from <https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/>
- Pornprasertmanit, S. (2014). *The unified approach for model evaluation in structural equation modeling*. Unpublished doctoral dissertation, University of Kansas, Lawrence, KS. Retrieved from <http://hdl.handle.net/1808/16828>

- Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). A Monte Carlo approach for nested model comparisons in structural equation modeling. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th annual psychometric society meeting* (pp. 187–197). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-9348-8_12
- Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, 41, 227–259. http://dx.doi.org/10.1207/s15327906mbr4103_1
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Rigdon, E. E. (1998). The equal correlation baseline model for comparative fit assessment in structural equation modeling. *Structural Equation Modeling*, 5, 63–77. <http://dx.doi.org/10.1080/10705519809540089>
- Rossee, Y. (2012). lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, 48, 1–36. <http://dx.doi.org/10.18637/jss.v048.i02>
- Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling*, 21, 167–180. <http://dx.doi.org/10.1080/10705511.2014.882658>
- Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, 21, 149–160. <http://dx.doi.org/10.1080/10705511.2013.824793>
- Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, 64, 37–52.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464. <http://dx.doi.org/10.1214/aos/1176344136>
- Song, X.-Y., & Lee, S.-Y. (2006). Bayesian analysis of structural equation models with nonlinear covariates and latent variables. *Multivariate Behavioral Research*, 41, 337–365. http://dx.doi.org/10.1207/s15327906mbr4103_4
- Song, X.-Y., & Lee, S.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. West Sussex, UK: Wiley.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 64, 583–639. <http://dx.doi.org/10.1111/1467-9868.00353>
- Stan Development Team. (2018). *RStan: The R interface to Stan*. R package Version 2.18.2 [Computer Software]. Retrieved from <http://mc-stan.org/>
- Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYSGRAPH*. Evanston, IL: SYSTAT.
- Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society (IMPS), Iowa City, IA.
- Stromeyer, W. R., Miller, J. W., Sriramachandramurthy, R., & DeMartino, R. (2015). The prowess and pitfalls of Bayesian structural equation modeling: Important considerations for management research. *Journal of Management*, 41, 491–520. <http://dx.doi.org/10.1177/0149206314551962>
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10. <http://dx.doi.org/10.1007/BF02291170>
- van der Lans, R., van den Bergh, B., & Dieleman, E. (2014). Partner selection in brand alliances: An empirical investigation of the drivers of brand fit. *Marketing Science*, 33, 551–566. <http://dx.doi.org/10.1287/mksc.2014.0859>
- 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. <http://dx.doi.org/10.1037/met0000100>
- 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. <http://dx.doi.org/10.1007/s11222-016-9696-4>
- Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY: Springer http://dx.doi.org/10.1007/978-0-387-09612-4_9
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford Press.
- Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8, 16–37. <http://dx.doi.org/10.1037/1082-989X.8.1.16>
- Wu, H., & Browne, M. W. (2015). Quantifying adventitious error in a covariance structure as a random effect. *Psychometrika*, 80, 571–600. <http://dx.doi.org/10.1007/s11336-015-9451-3>
- Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling*, 23, 319–330. <http://dx.doi.org/10.1080/10705511.2015.1065414>
- Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, 14, 301–322. <http://dx.doi.org/10.1037/a0016972>
- Zhang, X., & Savalei, V. (2016). Bootstrapping confidence intervals for fit indexes in structural equation modeling. *Structural Equation Modeling*, 23, 392–408. <http://dx.doi.org/10.1080/10705511.2015.1118692>

(Appendix follows)

Appendix

Formulas for Additional Bayesian Fit Indices

These formulas use the same quantities defined for Equation 19 and are expected (when using noninformative priors) to be reasonable approximations of fit indices under MLE. We designate these with a superscript “DevM” to indicate the observed deviance at iteration i in the Markov chain is rescaled using pD to make its expectation equal to the deviance evaluated at the posterior mean. Note, however, that because the mean of a nonlinear function of X does not generally equal the same nonlinear function of \bar{X} , the recentering of D_i^{obs} does not imply that the posterior mean of any of these indices will equal the values of the indices calculated using $D(\bar{\theta})$. The simulation study shows that they are nonetheless reasonable approximations of their ML counterparts.

For all incremental fit indices, quantities for the hypothesized model (M_H) have an “H” subscript, whereas a “0” indicates the same quantity from the null model (M_0).

$$\text{BMc}_i^{\text{DevM}} = e^{-\frac{1}{2N}[(D_i^{\text{obs}} - pD) - (p^* - pD)]} = e^{-\frac{1}{2N}(D_i^{\text{obs}} - p^*)}. \quad (20)$$

$$\begin{aligned} \text{B-}\hat{\Gamma}_i^{\text{DevM}} &= \frac{p}{p + \frac{2}{N}[(D_i^{\text{obs}} - pD) - (p^* - pD)]} \\ &= \frac{p}{p + \frac{2}{N}(D_i^{\text{obs}} - p^*)}. \end{aligned} \quad (21)$$

$$\text{B-}\hat{\Gamma}_{adj,i}^{\text{DevM}} = 1 - \frac{p^*}{p^* - pD}(1 - \text{B-}\hat{\Gamma}_i^{\text{DevM}}). \quad (22)$$

$$\begin{aligned} \text{BTLI}_i^{\text{DevM}} = \text{BNNFI}_i^{\text{DevM}} &= \frac{\frac{D_{0,i}^{\text{obs}} - pD_0}{p^* - pD_0} - \frac{D_{H,i}^{\text{obs}} - pD_H}{p^* - pD_H}}{\frac{D_{0,i}^{\text{obs}} - pD_0}{p^* - pD_0} - 1}. \end{aligned} \quad (23)$$

$$\text{BNFI}_i^{\text{DevM}} = \frac{(D_{0,i}^{\text{obs}} - pD_0) - (D_{H,i}^{\text{obs}} - pD_H)}{D_{0,i}^{\text{obs}} - pD_0}. \quad (24)$$

$$\text{BCFI}_i^{\text{DevM}} = 1 - \frac{(D_{H,i}^{\text{obs}} - pD_H) - (p^* - pD_H)}{(D_{0,i}^{\text{obs}} - pD_0) - (p^* - pD_0)} = 1 - \frac{D_{H,i}^{\text{obs}} - p^*}{D_{0,i}^{\text{obs}} - p^*}. \quad (25)$$

Following from Hoofs et al. (2018), indices can be derived using similar principles, rescaling D_i^{obs} not by pD but by D_i^{rep} . We designate these with a superscript “PPMC” to indicate the observed deviance at iteration i in the Markov chain is rescaled using posterior predictive model checks.

$$\text{BMc}_i^{\text{PPMC}} = e^{-\frac{1}{2N}[(D_i^{\text{obs}} - D_i^{\text{rep}}) - (p^* - pD)]}. \quad (26)$$

$$\text{B-}\hat{\Gamma}_i^{\text{PPMC}} = \frac{p}{p + \frac{2}{N}[(D_i^{\text{obs}} - D_i^{\text{rep}}) - (p^* - pD)]}. \quad (27)$$

$$\text{B-}\hat{\Gamma}_{adj,i}^{\text{PPMC}} = 1 - \frac{p^*}{p^* - pD}(1 - \text{B-}\hat{\Gamma}_i^{\text{PPMC}}). \quad (28)$$

$$\begin{aligned} \text{BTLI}_i^{\text{PPMC}} = \text{BNNFI}_i^{\text{PPMC}} &= \frac{\frac{D_{0,i}^{\text{obs}} - D_{0,i}^{\text{rep}}}{p^* - pD_0} - \frac{D_{H,i}^{\text{obs}} - D_{H,i}^{\text{rep}}}{p^* - pD_H}}{\frac{D_{0,i}^{\text{obs}} - D_{0,i}^{\text{rep}}}{p^* - pD_0} - 1}. \end{aligned} \quad (29)$$

$$\text{BNFI}_i^{\text{PPMC}} = \frac{(D_{0,i}^{\text{obs}} - D_{0,i}^{\text{rep}}) - (D_{H,i}^{\text{obs}} - D_{H,i}^{\text{rep}})}{D_{0,i}^{\text{obs}} - D_{0,i}^{\text{rep}}}. \quad (30)$$

$$\text{BCFI}_i^{\text{PPMC}} = 1 - \frac{(D_{H,i}^{\text{obs}} - D_{H,i}^{\text{rep}}) - (p^* - pD_H)}{(D_{0,i}^{\text{obs}} - D_{0,i}^{\text{rep}}) - (p^* - pD_0)}. \quad (31)$$

Received April 23, 2018

Revision received March 25, 2019

Accepted March 28, 2019 ■