Structural Equation Modeling: A Multidisciplinary Journal, 26: 39–50, 2019
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2018.1490648

Fit for a Bayesian: An Evaluation of PPP and DIC for Structural Equation Modeling

Meghan K. Cain[1] and Zhiyong Zhang[2]
[1]*University of Texas at San Antonio*
[2]*University of Notre Dame*

Despite its importance to structural equation modeling, model evaluation remains underdeveloped in the Bayesian SEM framework. Posterior predictive p-values (PPP) and deviance information criteria (DIC) are now available in popular software for Bayesian model evaluation, but they remain underutilized. This is largely due to the lack of recommendations for their use. To address this problem, PPP and DIC were evaluated in a series of Monte Carlo simulation studies. The results show that both PPP and DIC are influenced by severity of model misspecification, sample size, model size, and choice of prior. The cutoffs PPP < 0.10 and ΔDIC > 7 work best in the conditions and models tested here to maintain low false detection rates and misspecified model selection rates, respectively. The recommendations provided in this study will help researchers evaluate their models in a Bayesian SEM analysis and set the stage for future development and evaluation of Bayesian SEM fit indices.

Keywords: Bayesian, deviance information criteria, model fit, posterior predictive p-values, structural equation modeling

The Bayesian framework offers a flexible approach to structural equation modeling (SEM; Kaplan & Depaoli, 2012; Lee, 2007; Palomo, Dunson, & Bollen, 2007; Raftery, 1993). Incorporation of prior knowledge allows estimation of under-identified models, a natural means of constraining parameters, and better small-sample performance (Scheines, Hoijtink, & Boomsma, 1999). Priors are combined with the current sample through Bayes' Theorem, often using Markov chain Monte Carlo (MCMC) sampling and data augmentation. Although MCMC tends to be more computationally demanding for simple models, highly complex problems can be less computationally demanding through MCMC than through numerical methods (Berger, 2006). Data augmentation naturally handles issues such as missingness, nonlinearity, multilevel structure, and others (Lee, 2007). Lastly, because Bayesian SEM provides full posterior distributions for each parameter and latent variable, more can be learned about the model as a whole.

Bayesian SEM has many applications within the social sciences (Rupp, Dey, & Zumbo, 2004), but its utility continues to be limited in practice largely due to the lack of guidelines and recommendations for model evaluation. Jordan (2011) has cited the lack of "off-the-shelf" methods for model selection the number one open problem in Bayesian statistics. Within the linear modeling framework, Bayes factor >3 is a common criterion and has been found to correspond highly with the traditional $\alpha < 0.01$ criterion (Jeon & De Boeck, 2017). Within the SEM framework, however, a systematic evaluation of Bayesian model evaluation has yet to be conducted. This is particularly challenging because the traditional fit indices, such as the chi-squared goodness-of-fit statistic (T_{ML}), RMSEA (Steiger & Lind, 1980), or CFI (Bentler, 1990), are not available or well defined when performing Bayesian SEM.

Posterior predictive p-value (PPP; Gelman, Meng, & Stern, 1996; Meng, 1994) and deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Linde, 2014; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) are Bayesian methods of model evaluation available in popular software. Currently, DIC is the only measure of model fit available in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), and PPP and DIC are available in Mplus (Muthén & Muthén, 2012). Due to their

Correspondence should be addressed to Meghan K. Cain, Department of Psychology, University of Texas at San Antonio, San Antonio, TX 78249. E-mail: meghan.cain@utsa.edu

availability, it is important for users to know what analysis features can affect PPP and DIC and how to interpret their values.

PPP

PPP can be thought of as a Bayesian-motivated generalization of T_{ML}. It is a natural byproduct of the MCMC approximation, calculated using posterior predictive distributions of the same sample size and of the same likelihood as the original data. At each MCMC iteration j, a new set of data $\boldsymbol{Y}^j$ is generated based on updated parameter estimates, $\boldsymbol{\theta}^j$. A discrepancy statistic, such as T_{ML}, is calculated for each generated posterior predictive distribution, resulting in as many T_{ML} statistics as there are samples in the posterior. A T_{ML} statistic is also calculated for the sample data, $\boldsymbol{X}$, using each updated parameter estimate, $\boldsymbol{\theta}^j$. PPP is the proportion of posterior predictive discrepancy statistics that are greater than the discrepancy statistics of the current data,

$$\mathrm{PPP} = p\left(T_{\mathrm{ML}}\left(\boldsymbol{X}, \boldsymbol{\theta}^j\right) < T_{\mathrm{ML}}\left(\boldsymbol{Y}^j, \boldsymbol{\theta}^j\right)\right). \tag{1}$$

An excellent-fitting model is expected to have a PPP value around 0.5, and an extreme value indicates misspecification. In Mplus, T_{ML} is used as the discrepancy statistic and only a low PPP indicates that the model is not appropriate for these data (Asparouhov & Muthén, 2010b). Within the item response theory (IRT) framework, PPP can also be used for model comparison by comparing the number of items or item pairs with extreme PPP values across models (e.g., Zhu & Stone, 2012).

In practice, it is still largely unknown whether any cutoffs can be reliably used with PPP to detect misspecification. Cutoffs are useful because they can provide a dichotomous indicator of model fit: A PPP below a specified cutoff would indicate that the model does not fit the data, and a PPP above the cutoff would indicate that the model does fit the data. Because PPP is not uniformly distributed, it has no theoretical cutoff to maintain Type I error at 0.05 like p-values do (Hjort, Dahl, & Steinbakk, 2006). Cutoffs of 0.01, 0.05, and 0.10 have been proposed (Asparouhov & Muthén, 2010b; Gelman et al., 1996; Muthén & Asparouhov, 2012).

The 0.05 cutoff has been used most often to maintain consistency with traditional p-values. Using this cutoff, it has been shown that PPP is a useful tool for model evaluation that is less sensitive to misspecification than T_{ML} and better maintains Type I error rate (Asparouhov & Muthén, 2010a; Sinharay, 2006). More research is needed to investigate the other cutoffs.

DIC

DIC is a generalization of AIC, in which the model complexity penalty is determined using the deviance of the hypothesized model (Spiegelhalter et al., 2002). Operationally, at each MCMC iteration j, the deviance is calculated using the updated parameter estimates, $\boldsymbol{\theta}^j$, and the current data, $\boldsymbol{X}$. The mean of these posterior deviances, $\bar{D}$, is compared to the deviance of the posterior mean, $D(\bar{\boldsymbol{\theta}})$, to obtain a calculation of model complexity,

$$p_D = \bar{D} - D(\bar{\boldsymbol{\theta}}). \tag{2}$$

DIC is then formulated the same way as AIC, with p_D replacing the number of parameters p:

$$\mathrm{DIC} = -2\log\{p(\boldsymbol{X}|\boldsymbol{\theta})\} + 2p_D. \tag{3}$$

DIC was developed for models with hierarchical structure and in Bayesian analysis when using informative priors, because the effective number of parameters is no longer straightforward (Spiegelhalter et al., 2002). In linear models with noninformative priors, AIC and DIC are expected to be equal (Ellison, 2004). Like AIC, the target model of DIC is not the true model. Rather, DIC tries to find the simplest model that fits the current data well (Plummer, 2006). Both AIC and DIC tend to prefer models that overfit the data in small samples (Ando, 2011; Plummer, 2008; Van der linde, 2005, 2012).

Although DIC has not been extensively tested through simulation, there are some reports of it working rather well. For example, Asparouhov, Muthén, and Morin (2015) found that DIC outperforms BIC in models with informative priors. Zhang, Lai, Lu, and Tong (2013) used DIC for model selection in a Bayesian growth curve model with good performance when the true model was the more complex model. In the same paper, however, DIC did not perform as well when the true model was the less complex model. DIC has been shown to prefer more complex testlet models within the IRT framework, as well (Li, Bolt, & Fu, 2006).

Like other information criteria, DIC does not follow a specified distribution, thus there is no formal test to compare two models. A minimum difference in DIC between 3 and 7 has been proposed to show sufficient evidence that the model with the smaller DIC fits better than the alternative model (Lee & Song, 2012; Spiegelhalter et al., 2002), but these have not been fully evaluated.

SIMULATION STUDIES

The objective of the current study is to evaluate the performance of PPP and DIC and to provide recommendations for their use. Specifically, it would benefit Bayesian SEM users to know which properties of their data analysis influence PPP and DIC so that they can appropriately interpret their values. Through two Monte Carlo simulation studies, this report will evaluate the impacts of model misspecification, sample size, model size, and choice of prior.

A model is said to be misspecified when one or more parameters are estimated whose population values are zero (*over-parameterization*), one or more parameters are fixed to zero whose population values are nonzero (*under-parameterization*), or both (Hu & Bentler, 1998). In addition to being sensitive to model misspecification, it is also desirable to have a fit index that is not sensitive to other features. There has been some previous research to show that PPP is sensitive to sample size (Asparouhov & Muthén, 2010a) and that DIC performance improves with sample size (Zhu & Stone, 2012), but little else is known about what other features of the data or model can affect PPP and DIC. Without this information, it is difficult to interpret PPP and DIC in a data analysis setting, especially if they contradict.

We therefore explore factors that might be related to PPP and DIC through two simulation studies. The first simulation study will evaluate the effects of sample size, model size, and model misspecification on PPP and DIC using different cutoff values. The second simulation study will evaluate the effect of prior choice. Using the results from both of these studies, the authors will provide recommendations and guidelines for the interpretation of PPP and DIC in a practical Bayesian SEM analysis.

Simulation design

Each simulation study uses one true model from which the data are generated. These data are then fit to the true model and five misspecified models, each of which is missing one additional parameter. Because the consequences of under-parameterization are more severe than over-parameterization (e.g., Maxwell & Delaney, 2004), this study will focus on the case of under-parameterized model misspecification. PPP's performance is evaluated by its ability to correctly detect misspecification in a misspecified analysis model and to not falsely detect misspecification in a true analysis model. DIC's performance is evaluated by its ability to select the true model over a misspecified model in a model comparison. As a benchmark, PPP's performance is shown alongside T_{ML}'s, and DIC's performance is shown alongside the likelihood ratio test (LRT). T_{ML} is the original fit statistic, testing the difference between the sample and model-implied covariance matrix (Hu & Bentler, 1999). The LRT tests whether the more complex model significantly improves the fit of the simpler model given the change in degrees of freedom (Satorra & Saris, 1985). The population models used to generate data for this simulation study were chosen based on work by Paxton, Curran, Bollen, Kirby, and Chen (2001) and have since been used by Bollen, Harden, Ray, and Zavisca (2014); Chen, Curran, Bollen, Kirby, and Paxton (2008), and others in SEM simulation studies. Paxton et al. searched the literature for applications of SEM in psychology and sociology journals to find what they describe as the models most "commonly encountered in applied research" (p. 292). The path diagrams of the smaller and larger versions of this model appear in Figure 1, along with their population parameter values. The smaller model has 9 manifest variables (9MV model) and the larger has 15 manifest variables (15MV model).

Population values for parameters were chosen to provide specific population RMSEA values. The authors began by using the values found by Paxton et al. and adjusted them so that the same misspecifications would have the same population RMSEA in both the 9MV and the 15MV models. This makes it easier to compare across model sizes. The variances of the error terms were varied to provide total unit variance for each latent variable and each manifest variable. Communalities of manifest variables without cross-loadings are 0.40. All data were simulated in R (R Core Team, 2016).

Along with fitting the true model to the simulated data, five misspecified models were fit to the data. A summary of the analysis models is in Table 1. Each subsequent model is missing an additional parameter, increasing its population RMSEA and degrees of freedom. The first misspecified model, Model 2, represents only slight misspecification while Model 6 represents severe misspecification. Each of these models was fit to data generated under Model 1 with sample sizes of 75, 150, 250, 500, and 1,000. Mplus (Muthén & Muthén, 2012) was used to fit all models because it can estimate both ML-SEM and Bayesian SEM and it is user friendly. The syntax used to fit the models is in the supplementary material. All simulation study conditions are listed in Table 2.

In the results, terms such as detection rates and model selection rates will be used in lieu of the traditional terms power and Type I error rates. This is because neither PPP nor DIC has significance tests where they can be categorized as being statistically significant or nonsignificant. Rather, PPP detection rates refer to the proportion of samples for which PPP was below a chosen cutoff, i.e., 0.10. If the model is truly misspecified, PPP < 0.10 is a *correct detection*; if the model is the true model, PPP < 0.10 is a *false detection*. False detection rates are comparable to T_{ML} Type I error rates and correct detection rates are comparable to T_{ML} power rates.

To compare two models, their difference in DIC is calculated,

$$\Delta\mathrm{DIC} = \mathrm{DIC}_{\mathrm{m}} - \mathrm{DIC}_{1}, \tag{4}$$

where $\mathrm{DIC}_{\mathrm{m}}$ refers to the DIC of the misspecified model and DIC_1 refers to the DIC of Model 1, the data generating model. DIC model selection rates refer to the proportion of samples for which ΔDIC is larger than a chosen cutoff, i.e., 7. *True model selection* occurs when $\Delta\mathrm{DIC} > 7$; *misspecified model selection* occurs when $\Delta\mathrm{DIC} < -7$. When DIC does not select a model, $-7 < \Delta\mathrm{DIC} < 7$, it is up to the researcher to choose the more substantively meaningful model or the less complex model based on the rule of parsimony. Note that while DIC has three options (true model selection, misspecified model selection, no selection), LRT has only two options (reject simpler model, fail

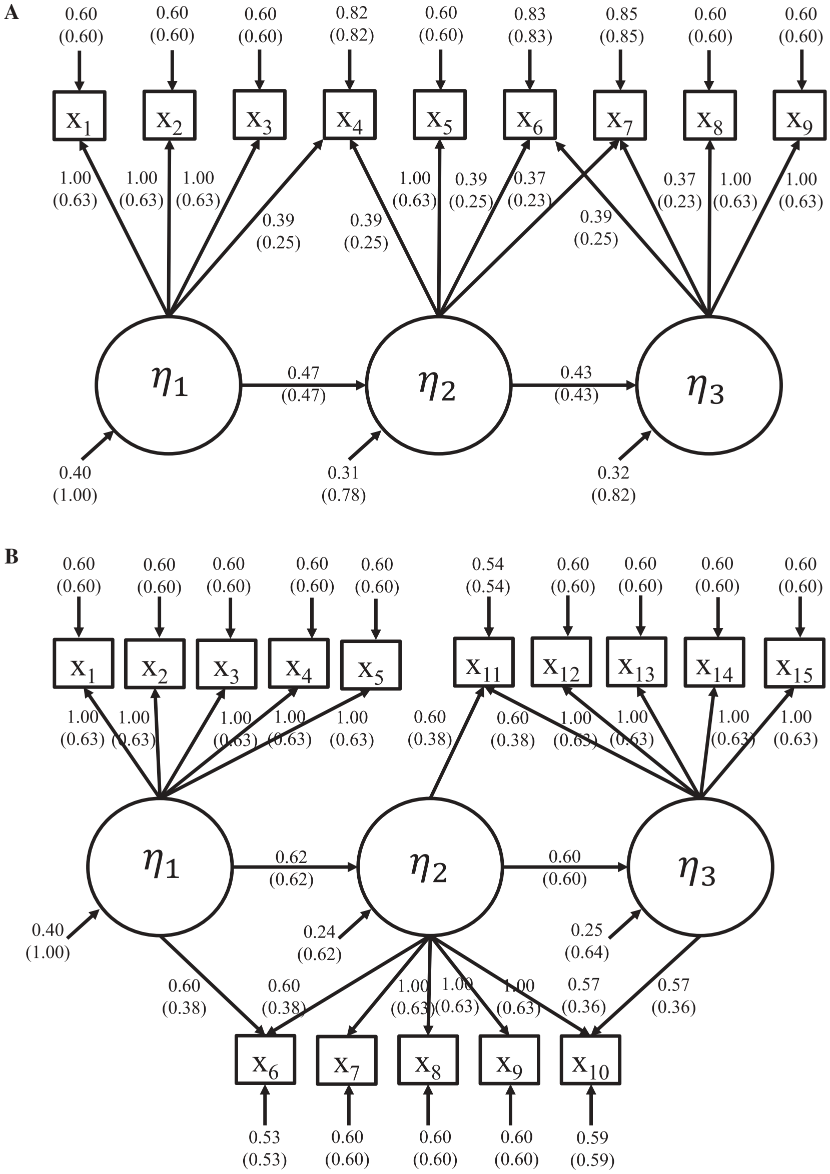

FIGURE 1 The 9MV (a) and 15MV (b) data generating models and their unstandardized (standardized) population parameter values.

to reject simpler model). For this simulation, the true model is always the more complex model. Consequently, LRT power rates are comparable to DIC true model selection rates. There is no LRT equivalent to DIC misspecified model selection rates.

Because PPP and DIC results are both represented as proportions of replications above or below some cutoff, standard errors can be computed to interpret their results. Any rate that is 2 standard errors (0.032) away from any given rate can be considered different.[1]

Simulation I: Establishment of cutoffs

The purpose of the first simulation is to evaluate the impact of model misspecification, sample size, and model size on the performance of PPP and DIC in order to establish cutoffs and other rules of thumb for their use. Mplus default priors were used for all parameters; these are provided in Table 3 for the reader's convenience. The PPP cutoffs of 0.05, 0.10, and 0.15 are shown alongside T_{ML} power and type I error rates, and DIC cutoffs of 3, 5, and 7 are shown alongside LRT power rates. PPP < 0.01 was found to be too conservative, and so these results are not shown.

[1] Standard error (SE) $= \sqrt{(p(1-p))/r}$, where r is the number of replications and p is the proportion. Using a proportion of 0.50 yields a standard error of 0.016, the largest possible standard error given $r = 1000$.

TABLE 1
True and Misspecified Models, Their RMSEA, and Degrees of Freedom

Model	*Description of misspecification*	*Pop. RMSEA*	*9MV model df*	*15MV model df*
1	True model	0.00	22	85
2	Missing one cross-loading	0.03	23	86
3	Missing two cross-loadings	0.04	24	87
4	Missing three cross-loadings	0.05	25	88
5	Missing three cross-loadings and one regression pathway	0.08	26	89
6	Missing three cross-loadings and two regression pathways	0.10	27	90

Note. The path diagram for the 9MV and 15MV models are shown in Figure 1. Pop. RMSEA $= \sqrt{F_{\mathrm{ML}}/\mathrm{df}}$, where F_{ML} is calculated as in Jöreskog (1967).

TABLE 2
Simulation Study Conditions

Factor	*Levels*
Sample size	75, 150, 250, 500, 1,000
Model size (no. of MVs)	9, 15
Model misspecification (RMSEA)[a]	0 (True), 0.028, 0.038, 0.050, 0.080, 0.100
Priors for λs	$1: N(0,\infty),\ 2:\ N(0.43, 0.04),$ $3: N(0.43, 0.01),\ 4: N(0.43, 0.005)$
PPP cutoffs	0.05, 0.10, 0.15
ΔDIC cutoffs	3, 5, 7
No. of replications/ condition	1,000

[a]The analysis models are listed in Table 1. MVs: Manifest variables.

TABLE 3
Prior Distributions in Mplus

Parameter type	*Prior distributions available*	*Default prior*
λ	Normal	$N\ (0, \infty)$
β	Normal	$N\ (0, \infty)$
ϵ	Inverse gamma	IG $(-1, 0)$
ζ	Inverse wishart	IW$(0, -p-1)$

Note. The only parameters listed here are those used in this study. For available distributions and default settings of priors on other types of parameters, see the Mplus User's Manual (Muthén & Muthén, 2012).

Results

A total of 1,000 converged replications were used to compute results for each condition. ML convergence rates were lowest (79–89%) at $n = 75$ and >93% at $n = 150$ for the 9MV models.

All convergence rates for the 15MV models were >99%. All Bayesian models converged; however, replications in which the calculation of p_D was negative were thrown out (<4% of replications). Calculation of p_D improved with sample size.

PPPs

PPP detection rates for each cutoff and T_{ML} significance rates for all models appear in Table 4. PPP false detection rates decrease with sample size, increase with model size, and increase with larger cutoffs. All PPP false detection rates are ≤0.05 for the 9MV models. For the 15MV models, PPP false detection rates are ≤0.02 using PPP < 0.05, ≤0.06 using PPP < 0.10, and ≤0.11 using PPP < 0.15. All T_{ML} Type I error rates are ≤0.06 with the 9MV model, but as high as 0.21 with the 15MV model at $n = 75$. These results show that the PPP < 0.15 cutoff may be inappropriate for the larger model; furthermore, T_{ML} may be inappropriate for the larger model with sample sizes less than 500.

PPP correct detection rates increase with sample size, model size, model misspecification, and larger cutoffs. In general, its behavior is similar to T_{ML}. For both the 9MV and 15MV models, PPP < 0.15 has correct detection rates closest to T_{ML} power rates. However, given the increased false detection rates with PPP < 0.15 for the 15MV model, it is recommended that cutoffs be decreased as model size increases.

DICs

A comparison of DIC model selection rates for each cutoff and LRT power rates for all models are in Table 5. DIC misspecified model selection rates decrease with sample size, model size, increased comparison model misspecification, and larger cutoffs. Performance in evaluating the smaller models is inconsistent at small sample sizes. Elimination of replications with negative p_Ds improved performance but did not entirely correct it. For the 9MV models, ΔDIC > 7 is the only cutoff with all misspecified model selection rates ≤0.05. All ΔDIC > 5 misspecified model selection rates are ≤0.05 with $n \geq 150$, and all ΔDIC > 3 misspecified model selection rates are ≤0.05 with $n \geq 250$. For the 15MV models, all DIC misspecified model selection rates are ≤0.05.

DIC correct model selection rates increase with sample size, model size, and comparison model misspecification and decrease with larger cutoffs. For both the 9MV and 15MV models, ΔDIC > 3 has true model selection rates closest to LRT power rates. However, this cutoff should only be used with larger sample and/or model sizes given its increased misspecified model selection rates with smaller samples sizes and models.

TABLE 4
PPP Detection Rates at Different Cutoffs

		9MV models				*15MV models*			
Model	n	*PPP < 0.05*	*PPP < 0.10*	*PPP < 0.15*	T_{ML}	*PPP < 0.05*	*PPP < 0.10*	*PPP < 0.15*	T_{ML}
1	75	0.01	0.02	0.05	0.06	0.02	0.06	0.11	0.21
	150	0.01	0.02	0.04	0.05	0.02	0.05	0.10	0.09
	250	0.01	0.01	0.04	0.05	0.01	0.03	0.07	0.07
	500	0.00	0.01	0.03	0.05	0.02	0.04	0.09	0.06
	1,000	0.00	0.02	0.05	0.06	0.02	0.03	0.07	0.06
2	75	0.01	0.04	0.08	0.09	0.05	0.10	0.19	0.30
	150	0.02	0.05	0.09	0.11	0.09	0.18	0.27	0.28
	250	0.03	0.08	0.13	0.16	0.15	0.29	0.44	0.43
	500	0.10	0.20	0.30	0.32	0.54	0.70	0.80	0.77
	1,000	0.33	0.50	0.65	0.67	0.96	0.98	0.99	0.99
3	75	0.02	0.06	0.11	0.12	0.08	0.16	0.25	0.40
	150	0.04	0.09	0.15	0.18	0.21	0.34	0.47	0.48
	250	0.07	0.16	0.27	0.31	0.48	0.62	0.73	0.72
	500	0.34	0.49	0.60	0.62	0.94	0.97	0.99	0.99
	1,000	0.83	0.92	0.95	0.96	1.00	1.00	1.00	1.00
4	75	0.03	0.09	0.14	0.18	0.17	0.28	0.40	0.53
	150	0.11	0.20	0.29	0.32	0.48	0.65	0.75	0.76
	250	0.22	0.39	0.51	0.58	0.82	0.91	0.95	0.94
	500	0.75	0.87	0.93	0.93	1.00	1.00	1.00	1.00
	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	75	0.16	0.31	0.42	0.51	0.62	0.77	0.85	0.92
	150	0.58	0.76	0.85	0.86	0.99	1.00	1.00	1.00
	250	0.92	0.96	0.98	0.98	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	75	0.42	0.61	0.73	0.78	0.95	0.98	0.99	1.00
	150	0.92	0.96	0.98	0.99	1.00	1.00	1.00	1.00
	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. Descriptions of these models are in Table 1. PPP results for Model 1 are false detection rates and results for the remaining models are correct detection rates. All results are based on 1000 converged replications. MVs: Manifest variables.

Conclusions

The results from these simulations indicate that PPP and DIC are both heavily impacted by sample size, model size, and model misspecification. PPP is better able to detect misspecification and DIC is better able to choose the correct model as sample size increases and as model size increases. Furthermore, DIC's ability to not select the misspecified model improved as sample size increased and was only consistently maintained using $\Delta\text{DIC} > 7$. PPP's false detection rates were all lower than 0.05 with cutoffs ≤ 0.10. As with T_{ML}, in larger samples, PPP will always reject a model even with minimal misspecification. In practical data analysis, the true model will likely not be evaluated. Therefore, PPP may not be useful in large samples unless the true model is among the candidate models. Alternatively, DIC showed inconsistent performance with $n < 250$ and should not be used in small sample sizes. As sample size increases, DIC's performance improves.

Larger PPP cutoffs corresponded most similarly to T_{ML} performance, and smaller DIC cutoffs corresponded most similarly to LRT performance. However, these are also the cutoffs that had too high false detection rates and misspecified model selection rates, respectively. In practical data analysis, sample size and model size should be taken into account when evaluating PPP and DIC. For the particular models tested here, PPP < 0.15 is recommended for the 9MV models and PPP < 0.10 is recommended for the 15MV models; $\Delta\text{DIC} > 7$ is recommended for the 9MV models unless sample size is large, and $\Delta\text{DIC} > 3$ is recommended for the 15MV model. In the second simulation, only the 9MV model is used. Therefore, PPP < 0.15 and $\Delta\text{DIC} > 7$ are used to calculate the results appearing in the remainder of this document.

Simulation II: Influence of priors

The purpose of this simulation is to evaluate the impact of prior choice on the performances of PPP and DIC. Specifically, this simulation study will assess how priors on factor loadings will affect PPP detection rates and DIC model selection rates. It is

TABLE 5
DIC Model Selection Rates at Different Cutoffs

		Model 2		*Model 3*		*Model 4*		*Model 5*		*Model 6*	
n	*Criterion*	*M1*	*M2*	*M1*	*M3*	*M1*	*M4*	*M1*	*M5*	*M1*	*M6*
9MV models											
75	ΔDIC > 3	0.28	0.10	0.30	0.20	0.38	0.20	0.77	0.04	0.90	0.01
	ΔDIC > 5	0.19	0.03	0.20	0.06	0.27	0.08	0.68	0.01	0.84	0.01
	ΔDIC > 7	0.14	0.02	0.14	0.03	0.21	0.02	0.58	0.00	0.79	0.00
	LRT	0.24		0.28		0.41		0.85		0.97	
150	ΔDIC > 3	0.31	0.08	0.38	0.11	0.60	0.09	0.96	0.01	1.00	0.00
	ΔDIC > 5	0.20	0.03	0.26	0.01	0.48	0.03	0.92	0.01	1.00	0.00
	ΔDIC > 7	0.14	0.03	0.20	0.00	0.37	0.00	0.89	0.01	1.00	0.00
	LRT	0.37		0.52		0.71		0.99		1.00	
250	ΔDIC > 3	0.45	0.02	0.62	0.03	0.85	0.01	1.00	0.00	1.00	0.00
	ΔDIC > 5	0.31	0.01	0.46	0.01	0.76	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 7	0.19	0.01	0.35	0.00	0.68	0.00	1.00	0.00	1.00	0.00
	LRT	0.55		0.74		0.90		1.00		1.00	
500	ΔDIC > 3	0.76	0.00	0.93	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 5	0.63	0.00	0.85	0.00	0.99	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 7	0.50	0.00	0.77	0.00	0.98	0.00	1.00	0.00	1.00	0.00
	LRT	0.81		0.96		1.00		1.00		1.00	
1,000	ΔDIC > 3	0.96	0.00	1.00	0.00	1.00	0.00	0.99	0.01	1.00	0.00
	ΔDIC > 5	0.92	0.00	1.00	0.00	1.00	0.00	0.99	0.01	1.00	0.00
	ΔDIC > 7	0.87	0.00	1.00	0.00	1.00	0.00	0.99	0.01	1.00	0.00
	LRT	0.98		1.00		1.00		1.00		1.00	
15MV models											
75	ΔDIC > 3	0.44	0.00	0.61	0.05	0.86	0.01	1.00	0.00	1.00	0.00
	ΔDIC > 5	0.32	0.00	0.49	0.00	0.78	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 7	0.22	0.00	0.39	0.00	0.70	0.00	0.99	0.00	1.00	0.00
	LRT	0.57		0.74		0.90		1.00		1.00	
150	ΔDIC > 3	0.77	0.00	0.93	0.00	0.99	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 5	0.65	0.00	0.87	0.00	0.99	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 7	0.53	0.00	0.80	0.00	0.97	0.00	1.00	0.00	1.00	0.00
	LRT	0.88		0.97		1.00		1.00		1.00	
250	ΔDIC > 3	0.97	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 5	0.91	0.00	0.99	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 7	0.86	0.00	0.99	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	LRT	0.98		1.00		1.00		1.00		1.00	
500	ΔDIC > 3	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 5	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 7	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	LRT	1.00		1.00		1.00		1.00		1.00	
1,000	ΔDIC > 3	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 5	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	ΔDIC > 7	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	LRT	1.00		1.00		1.00		1.00		1.00	

Note. Descriptions of each model are in Table 1. ΔDIC is defined in Equation (2). Each pair of columns shows true model selection rates under "M1" and misspecified model selection rates under the alternative model column heading.

well known that prior choice can affect parameter estimates and substantive conclusions (Gelman, 2006; Gelman & Shalizi, 2013; Johnson, 2013; Seaman, Seaman, & Stamey, 2012; van de Schoot & Depaoli, 2014), but it is less clear how sensitive PPP and DIC are to prior choice. Some simulation studies have shown PPP to be prior-dependent (Asparouhov & Muthén, 2010a), while theoretical work suggests that PPP is robust to small modification on the prior (De la Horra & Teresa Rodriguez-Bernal, 2003; Gelman et al., 1996). In contrast, it is believed that DIC is strongly sensitive to prior choice (Spiegelhalter et al., 2014; Ward, 2008).

The aim of this simulation is to assess the sensitivity of PPP and DIC to changes in prior accuracy in a Bayesian SEM analysis. For the purposes of demonstration, only the factor loadings will be given an informative prior, while the other parameters will keep the same default priors used in the previous study. This approach was chosen because researchers often have interest in only a subset of parameters, and it is likely that previous studies will provide some information about the factor loadings. Because Mplus currently only allows specification of normal priors for factor loadings, only the hyperparameters will be changed.

Three priors, Prior 2: $N(0.43, 0.040)$, Prior 3: $N(0.43, 0.010)$, and Prior 4: $N(0.43, 0.005)$, are compared to the default prior, Prior 1: $N(0, \infty)$. The population values of the factor loadings range between 0.23 and 0.63. Therefore, Priors 2–4 cover the population parameter range within 1 standard deviation (SD), 2 SDs, and 3 SDs, respectively. It is predicted that Prior 2 would have the best performance while Prior 4 would have the worst. The hypothetical setup for this experiment could be that a researcher has found several previous studies using these variables that provide some knowledge of how each latent variable is measured. These studies have shown mean standardized factor loadings to be around 0.43; however they're unsure of how informative to make the priors.

TABLE 6
PPP Detection Rates for Each Prior

Model	n	*Prior 1*	*Prior 2*	*Prior 3*	*Prior 4*
1	75	0.05	0.05	0.27	0.55
	150	0.04	0.04	0.29	0.78
	250	0.04	0.03	0.17	0.76
	500	0.03	0.02	0.09	0.60
	1,000	0.05	0.04	0.07	0.27
2	75	0.08	0.08	0.26	0.49
	150	0.09	0.08	0.24	0.64
	250	0.13	0.10	0.24	0.66
	500	0.30	0.27	0.38	0.75
	1,000	0.65	0.64	0.71	0.87
3	75	0.11	0.08	0.19	0.33
	150	0.15	0.14	0.27	0.54
	250	0.27	0.27	0.37	0.63
	500	0.60	0.57	0.65	0.83
	1,000	0.95	0.95	0.95	0.97
4	75	0.14	0.11	0.19	0.27
	150	0.29	0.29	0.37	0.56
	250	0.51	0.48	0.57	0.71
	500	0.93	0.92	0.93	0.97
	1,000	1.00	1.00	1.00	1.00
5	75	0.42	0.37	0.47	0.57
	150	0.85	0.84	0.88	0.92
	250	0.98	0.98	0.98	0.99
	500	1.00	1.00	1.00	1.00
	1,000	1.00	1.00	1.00	1.00
6	75	0.73	0.67	0.76	0.82
	150	0.98	0.98	0.99	1.00
	250	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00
	1,000	1.00	1.00	1.00	1.00

Note. Descriptions of each model are in Table 1. PPP results for Model 1 are false detection rates and results for the remaining models are correct detection rates. Priors are described in Table 2.

Results

All replications converged for each condition; however, replications with a negative calculation of p_D were thrown out (<4% replications). Only conditions using the noninformative default prior, Prior 1, had any instances of a negative calculation of p_D.

PPPs

PPP detection rates for each prior are shown in Table 6; all rates are the proportion of samples with PPP < 0.15. As expected, PPP false detection rates increase with the increasingly inaccurate priors. Only Priors 1 and 2's false detection rates are all ≤ 0.05, demonstrating that the chosen prior distribution must cover the population parameter range within 1 SD to obtain accurate results. Correct detection rates are lower for Prior 2 than Prior 1 at $n = 75$ for all models. Correct detection rates are higher for Priors 3 and 4 in detecting minor model misspecifications, but this effect lessens as model misspecification increases. As expected, the effect of prior choice lessens as sample size increases.

DICs

DIC model selection rates for each prior are in Table 7; all rates are the proportion of samples with $\Delta\text{DIC} > 7$. As expected, Prior 2 has the best performance and Prior 4 has the worst. When comparing Models 1 (true) and 2 (misspecified), Model 2 is selected in up to 52% of replications when $n = 150$ using Prior 4; Model 2 is never selected when using Prior 2. In fact, the highest misspecified model selection rate using Prior 2 is 3% across conditions. Across all priors, misspecified model selection rates are all ≤ 0.05 with $n \geq 500$. True model selection rates decrease with increasingly inaccurate priors, though this effect dissipates as sample size increases. One of the larger gaps appears when comparing Models 1 (true) and 5 (misspecified) at $n = 75$. Model 1 is selected in 69% of replications using Prior 2; Model 1 is selected in only 33% of replications when using Prior 4.

Conclusions

The aim of this simulation was to assess the impact of prior choice. It was shown that using inaccurate priors negatively impacted both PPP and DIC in terms of both selecting the true model and detecting a misspecified model. The difference in performance between the default prior, Prior 1, and Prior 2 was generally larger for DIC than for PPP, suggesting that DIC is more sensitive to prior selection than is PPP.

TABLE 7
DIC Model Selection Rates for Each Prior

		Model 2		*Model 3*		*Model 4*		*Model 5*		*Model 6*	
n	*Prior*	*M1*	*M2*	*M1*	*M3*	*M1*	*M4*	*M1*	*M5*	*M1*	*M6*
75	Prior 1	0.19	0.03	0.20	0.06	0.27	0.08	0.68	0.01	0.84	0.01
	Prior 2	0.12	0.00	0.15	0.00	0.22	0.03	0.69	0.00	0.87	0.00
	Prior 3	0.10	0.04	0.09	0.30	0.11	0.42	0.45	0.13	0.72	0.05
	Prior 4	0.11	0.11	0.08	0.45	0.08	0.59	0.33	0.26	0.62	0.09
150	Prior 1	0.20	0.03	0.26	0.01	0.48	0.03	0.92	0.01	1.00	0.00
	Prior 2	0.12	0.00	0.29	0.00	0.52	0.01	0.95	0.00	1.00	0.00
	Prior 3	0.06	0.29	0.14	0.28	0.26	0.25	0.81	0.03	0.98	0.00
	Prior 4	0.05	0.52	0.06	0.64	0.11	0.60	0.58	0.13	0.91	0.02
250	Prior 1	0.31	0.01	0.46	0.01	0.76	0.00	1.00	0.00	1.00	0.00
	Prior 2	0.24	0.00	0.56	0.00	0.80	0.00	1.00	0.00	1.00	0.00
	Prior 3	0.14	0.12	0.32	0.09	0.58	0.06	0.99	0.00	1.00	0.00
	Prior 4	0.06	0.45	0.09	0.52	0.21	0.42	0.90	0.01	1.00	0.00
500	Prior 1	0.63	0.00	0.85	0.00	0.99	0.00	1.00	0.00	1.00	0.00
	Prior 2	0.64	0.00	0.88	0.00	0.99	0.00	1.00	0.00	1.00	0.00
	Prior 3	0.58	0.00	0.81	0.00	0.98	0.00	1.00	0.00	1.00	0.00
	Prior 4	0.41	0.05	0.53	0.08	0.84	0.02	1.00	0.00	1.00	0.00
1,000	Prior 1	0.92	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	Prior 2	0.92	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	Prior 3	0.89	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	Prior 4	0.84	0.00	0.97	0.00	1.00	0.00	1.00	0.00	1.00	0.00

Note. Descriptions of each model are in Table 1. Each pair of columns shows true model selection rates under "M1" and misspecified model selection rates under the alternative model column heading. The priors are described in Table 2.

DISCUSSION

The Bayesian framework offers a flexible approach to SEM estimation, but one major challenge that continues to limit its utility is the lack of guidelines for evaluating model fit and model comparison. PPP and DIC are now available through Mplus and WinBUGS, but their performance has not been widely evaluated and guidelines for their use have not been provided. Without practical guidelines such as cutoff values, Bayesian SEM users cannot appropriately use and interpret PPP and DIC in their own data analysis.

The broad goals of this project were to evaluate PPP and DIC to identify the conditions of a data analysis that may affect their performance in order to provide guidelines for their use in practical analysis. Specifically, the simulation studies in this report examined the impacts of sample size, model size, model misspecification, and prior choice on PPP detection rates and DIC model selection rates using different cutoff values. Model size was defined by the number of manifest variables, model misspecification was defined by RMSEA, and prior accuracy was defined by prior coverage of the population parameter space. The choice of RMSEA and of specific priors were solely for demonstration purposes to show how the performances of PPP and DIC are impacted as model misspecification or prior accuracy gets worse. The definitions of "worse" for either were not important, because the trend in results were more of interest here than the rates themselves. Similar trends in results are expected with other definitions.

The results from the simulation studies showed that both PPP detection rates and DIC true model selection rates increased with sample size, model size, and model misspecification. PPP and DIC were found to be less powerful than their ML counterparts, T_{ML} and LRT, respectively, but were in general comparable. PPP false detection rates were lower than T_{ML} Type I error rates in most conditions. PPP < .15 had correct detection rates most comparable to T_{ML} and maintained false detection rates below 5% when evaluating the true model among the smaller analysis models, but a lower cutoff is required for larger models to maintain a low false detection rate. $\Delta\mathrm{DIC} > 7$ had the lowest true model selection rates but maintained misspecified model selection rates below 5% for the smaller models; a smaller cutoff could be used for larger samples and/or larger models.

In evaluating the impact of prior choice, it was found that the performances of PPP and DIC both suffered greatly when using an inaccurate prior. Unless population values were within 1 SD of the prior, PPP false detection rates increased, DIC misspecified model selection rates increased, and DIC true model selection rates decreased. For PPP, the performances of the default prior and the accurate informative prior were similar for false detection rates, but the accurate informative prior had slightly lower correct detection rates. In other words, for PPP, the default prior had the best performance. For DIC, the accurate informative prior outperformed the default prior. In general, prior influence of all priors decreased as sample size increased, although there were still small effects even at $n = 1,000$.

Because the calculation of DIC was inconsistent in samples smaller than 250, it is recommended to only use DIC in larger sample sizes. Alternatively, PPP may not be useful in large samples because it will always detect misspecification even when the model is minimally misspecified. In smaller samples, PPP had high rates of detecting misspecification in the true model only when an informative prior was inappropriate. These results show that PPP may be used for prior selection in practical data analysis. Future studies would be required, however, before any practical guidelines could be established for this use of PPP. These comparisons were not done with DIC in the current analysis and therefore are out of the scope of the current project.

This paper provides the first large-scale simulation study to evaluate the performances of PPP and DIC and provide guidelines for their use. As for any simulation study on fit statistics and fit indices, there are many limitations to the current work, the most severe being generalizability. Because the purpose of this study was to give an overview of the performances of PPP and DIC and to set a foundation for future work, more in-depth research in a particular area would be required before any recommendations or guidelines could be made for specific circumstances not tested in the current analysis. Future studies should evaluate PPP and DIC in alternative models, misspecifications, priors, and data distributions. There also needs to be future work in using PPP and DIC on categorical data, multilevel data, missing data, and in models with mean structure. Some specific limitations that warrant future research are discussed below.

First, only one model type was evaluated in this study. This model was chosen for being the most commonly used in social science SEM applications, but the use of one model severely limits the scope of the recommendations provided here. In addition, only one type of misspecification, under-parameterization, was used here. Second, in evaluating the performance of DIC, misspecified models were only compared to the true model. In practice, the true model would not be among the competing models. Future research should examine DIC's performance in selecting among misspecified models. Third, only four prior configurations were tested here. These were sufficient to show the sensitivity of PPP and DIC to prior choice, but a much larger simulation study should be conducted to broadly evaluate the impact that prior has on each. There also needs to be future work to establish guidelines on whether and how they could be used in a sensitivity analysis or in prior selection, for example.

Fourth, only normally distributed data were evaluated here and were only evaluated using the normal likelihood-based model. Future studies should examine the impact of nonnormality and examine the performances of PPP and DIC in a robust Bayesian SEM analysis.

In addition to these limitations, it also seems necessary at this point to remind readers of the dangers of using specific cutoffs for fit indices in general. This has been discussed in many places (Chen et al., 2008; Fan & Sivo, 2007; Kenny & McCoach, 2003; Marsh, Hau, & Wen, 2004) and so will not be reproduced here, except to say that if the model being analyzed is not reasonably close in characteristic to that analyzed here that the performance of PPP and DIC may differ substantially. For added insurance, the two-index presentation strategy (Hu & Bentler, 1999) should be employed in Bayesian SEM as it is in ML-SEM. Because PPP and DIC provide different information, reporting both would provide a more complete picture of the models being tested. It is also important to keep in mind that PPP and DIC are merely additional tools to help guide a researcher in any particular study. The traditional methods of cross-validation and replication should always be applied to assess a given model.

Until future research can be conducted, the results of this study show that PPP and DIC can be used for model evaluation in Bayesian SEM if the models and conditions of the data analysis are similar to those investigated here. Based on the results from the simulation studies, a summary of the recommended cutoffs for these models is shown in Table 8. Through these results and recommendations, Bayesian SEM can become a more accessible option for social science researchers to have more flexibility in their SEM analysis.

TABLE 8
Summary of Recommendations

Sample size	*Model size*	*Recommended cutoffs*	
75	9MV	PPP < 0.15	ΔDIC > 7
	15MV	PPP < 0.10	ΔDIC > 3
150	9MV	PPP < 0.15	ΔDIC > 5
	15MV	PPP < 0.10	ΔDIC > 3
≥250	9MV	PPP < 0.15	ΔDIC > 3
	15MV	PPP < 0.10	ΔDIC > 3

SUPPLEMENTAL DATA

Supplemental data for this article can be accessed here.

REFERENCES

Ando, T. (2011). Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences*, *31*(1–2), 13–38. doi:10.1080/01966324.2011.10737798

Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis of latent variable models using Mplus*. Retrieved June, 17, 2014, from http://www.statmodel2.com/download/BayesAdvantages18.pdf

Asparouhov, T., & Muthén, B. O. (2010b). Bayesian analysis using Mplus: Technical implementation. *Manuscript Submitted for Publication*. Retrieved from https://www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., Muthén, B. O., & Morin, A. J. (2015). *Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al*. Los Angeles, CA: Sage Publications Sage CA.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. doi:10.1037/0033-2909.107.2.238

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. doi:10.1214/06-BA115

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling*, *21*(1), 1–19. doi:10.1080/10705511.2014.856691

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*(4), 462–494. doi:10.1177/0049124108314720

De la Horra, J., & Teresa Rodriguez-Bernal, M. (2003). Bayesian robustness of the posterior predictive p-value. *Communications in Statistics-Theory and Methods*, *32*(8), 1493–1503. doi:10.1081/STA-120022241

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, *7*(6), 509–520. doi:10.1111/ele.2004.7.issue-6

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*(3), 509–529. doi:10.1080/00273170701382864

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. doi:10.1214/06-BA117A

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. doi:10.1111/j.2044-8317.2011.02037.x

Hjort, N. L., Dahl, F. A., & Steinbakk, G. H. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, *101*(475), 1157–1174. doi:10.1198/016214505000001393

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

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

Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychological Methods*, *22*(2), 340–360. doi:10.1037/met0000140

Johnson, V. E. (2013). Uniformly most powerful Bayesian tests. *Annals of Statistics*, *41*(4), 1716–1741. doi:10.1214/13-AOS1123

Jordan, M. (2011). What are the open problems in Bayesian statistics. *The ISBA Bulletin*, *18*, 1–4.

Jöreskog, K. G. (1967). A general approach to confirmatory maximum likelihood factor analysis. *ETS Research Bulletin Series*, *(1967*(2), 183–202. doi:10.1002/j.2333-8504.1967.tb00991.x

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). New York, NY, US: Guilford Press.

Kenny, D. A., & McCoach, B. D. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, *10*(3), 333–351. doi:10.1207/S15328007SEM1003_1

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Chichester, England: John Wiley & Sons.

Lee, S.-Y., & Song, X.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. Chichester, England: John Wiley & Sons.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21. doi:10.1177/0146621605275414

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS — A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337. doi:10.1023/A:1008929526011

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: A Multidisciplinary Journal*, *11*(3), 320–341. doi:10.1207/s15328007sem1103_2

Maxwell, S., & Delaney, H. (2004). A brief primer of principles of formulating and comparing models. In *Designing experiments and analyzing data: A model comparison perspective* (2nd ed., pp. B-26–B-36). Mahwah, NJ: Lawrence Erlbaum Associates.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*, 1142–1160. doi:10.1214/aos/1176325622

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. doi:10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Palomo, J., Dunson, D. B., & Bollen, K. (2007). Bayesian structural equation modeling. *Handbook of Latent Variable and Related Models*, *1*, 163–188.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312. doi:10.1207/S15328007SEM0802_7

Plummer, M. (2006). Comment on article by Celeux et al. *Bayesian Analysis*, *1*(4), 681–686. doi:10.1214/06-BA122C

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, *9*(3), 523–539. doi:10.1093/biostatistics/kxm049

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen and J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 163–180), Beverly Hills, CA: Sage..

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or Not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling*, *11*(3), 424–451. doi:10.1207/s15328007sem1103_7

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*(1), 83–90. doi:10.1007/BF02294150

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*(1), 37–52. doi:10.1007/BF02294318

Seaman, J. W., III, Seaman, J. W., Jr, & Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, *66*(2), 77–84. doi:10.1080/00031305.2012.695938

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 429–449. doi:10.1348/000711005X66888

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 485–493. doi:10.1111/rssb.2014.76.issue-3

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*(4), 583–639. doi:10.1111/rssb.2002.64.issue-4

Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. In *Annual meeting of the Psychometric Society*, Iowa City, IA (Vol.758, pp. 424–453).

van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. Marta Marques & Kyra Hamilton, 36, 75

Van Der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica*, *59*(1), 45–56. doi:10.1111/stan.2005.59.issue-1

Van Der Linde, A. (2012). A Bayesian view of model complexity. *Statistica Neerlandica*, *66*(3), 253–271. doi:10.1111/j.1467-9574.2011.00518.x

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modelling*, *211*(1–2), 1–10. doi:10.1016/j.ecolmodel.2007.10.030

Zhang, Z., Lai, K., Lu, Z., & Tong, X. (2013). Bayesian inference and application of Robust growth curve models using student's *t* distribution. *Structural Equation Modeling*, *20*(1), 47–78. doi:10.1080/10705511.2013.742382

Zhu, X., & Stone, C. A. (2012). Bayesian comparison of alternative graded response models for performance assessment applications. *Educational and Psychological Measurement*, *72*(5), 774–799. doi:10.1177/0013164411434638