Psychological Test and Assessment Modeling, Volume 64, 2022 (1), 22-47

Using BIC and aBIC to develop Bayesian posterior probabilities for latent variable models

W. Holmes Finch[1]

Abstract

Use of structural equation models (SEMs) to investigate relationships between latent variables has become increasingly widespread over the last 20 years. The use of SEM involves fitting multiple plausible models and selecting the one that provides optimal fit. A common approach for determining the optimal model involves use of information indices, which combine model misfit with a penalty for model complexity, with a minimum value being optimal. One concern regarding this strategy is that it does not acknowledge uncertainty inherent in the process. Wu, et al. (2020) described a Bayesian approach to quantifying this uncertainty and demonstrated its utility in the context of observed variable path models. This simulation study extends the work of Wu, et al. to the case of latent variable SEMs. Results demonstrate that the Bayesian approach does work well in the latent variable context. Implications for practice are discussed.

Keywords: Structural equation modeling; model selection; Bayesian information criterion

[1] *Correspondence concerning this article should be addressed to:* Holmes Finch, Department of Educational Psychology, Muncie, IN, 47306. USA, whfinch@bsu.edu

Structural equation models (SEMs) are widely used to investigate relationships among latent constructs such as mood, executive functioning, cognitive ability, and personality in disciplines across the social sciences (Steegh, Hoffler, Hoft, & Parchmann, 2020; Ulper, Cetinkaya, & Dikici, 2018; Aguado, et al., 2015; Aveh, 2015). These models allow researchers to link latent variables with one another in complex ways, thereby assessing the plausibility of theories about human psychology, behavior, and social structures. One of the key issues faced by researchers using SEM to address their research questions is the selection of the optimal model given a set of data. Such model selection can be done using measures of both absolute and relative model fit. In particular, data analysts often use one or more from the family of information indices, which combine a measure of model misfit for a given set of data (i.e., the log-likelihood), and add to it a penalty for model complexity. These indices are designed to favor relatively parsimonious models, unless additional the additional parameters in more complex models reduce model misfit by a sufficiently large degree so as to counteract the complexity penalty. In practice, researchers typically fit several models to a sample of data and then select the one with the lowest information index value as being optimal (Raftery, 1995). They would then assess whether the selected model fits the data well by using multiple absolute fit indices, such as CFI, TLI, RMSEA, and SRMR (Kline, 2016).

Recently, Wu, Cheung, & Leung (2020) described an alternative method for using one particular information index, the Bayesian Information Criterion (BIC), to create a set of posterior probabilities for a group of plausible models given a particular dataset. In this context, each model that is fit to the data is assigned a posterior probability based on the data and a prior probability. The researcher using this approach has information about the likelihood of each model, and can then make decisions regarding which one(s) should be explored more fully. An advantage of this Bayesian approach over the method for selecting a single optimal model as described above is that the uncertainty inherent in model building is reflected in the posterior probabilities, as opposed to the selection of a single model without consideration of how likely it is when compared to the alternatives. The purpose of the current study was to extend the work of Wu, et al., who tested this approach in an observed variables path analysis framework. In this simulation study, latent variable SEM models were considered, and the Bayesian approach to using information indices, which is described in more detail below, was applied under a variety of conditions.

BIC and sample size adjusted BIC

There exist a wide array of fit statistics for use in the context of SEM. Some of the more common of these, including the comparative fit index (CFI), the Tucker-Lewis Index (TLI), and the root mean squared error of approximation (RMSEA) are used to assess the overall fit of a model with respect to the data. In terms of comparing the fit of models with one another, statistics such as the Akaike Information Criterion (AIC; Akaike, 1974), BIC (Schwartz, 1978), and sample size adjusted BIC (aBIC; Sclove, 1987) can be employed by researchers. Each of these statistics is built upon the log-

likelihood of the estimate model along with a penalty for model complexity. More specifically, they are calculated as:

$$AIC = -2LL + 2q \quad (1)$$

$$BIC = -2LL + qln(n) \quad (2)$$

$$aBIC = -2LL + qln\left[\frac{n+2}{24}\right] \quad (3)$$

Where

$LL =$ Log-likelihood of the fitted model

$q =$ Number of free model parameters

$n =$ Sample size

These information indices can be used by data analysts to select the optimal model, which is the one with the smallest value. In practice, researchers fit multiple models to a set of data and then select the one with the lowest information index value to be optimal.

A limitation of this model selection approach is that it ignores the uncertainty associated with sampling variability that is inherent in the model selection process. In other words, the values of fit statistics such as the BIC will differ from sample to sample drawn from the same population, given a constant population value. For many statistics (e.g., regression model coefficients) this uncertainty is measured by the standard error. However, information indices such as the BIC do not have associated standard errors and thus quantifying this uncertainty and using it in comparing the relative fit of models is not possible. For example, if the BIC for model 1 is 200, and the BIC for model 2 is 100, then the researcher would select model 2 as providing the better fit to the data. Likewise, if BIC for model 1 is 200, and the BIC for model 2 is 199 the researcher would also select model 2 as being better fitting. However, it would be tempting to conclude that if the sample sizes and number of free parameters are identical in the two examples, then there would be greater certainty with the first comparison (where BIC values differed by 100) as opposed to the second (where BIC values differed by 1). However, the standard approach to applying information indices in the model selection process do not allow for any quantification of such model selection uncertainty. However, as Wu, et al. (2020) described, it is possible to use the BIC in the calculation of posterior probabilities for all models in the model space. We will discuss this approach next.

BIC posterior model probabilities

Wu, et al. (2020) discussed in some detail how the commonly used BIC statistic can be used to express the marginal likelihood of a SEM. They drew on earlier work by Raftery (1993), Konishi, Ando, and Imoto (2004), and Preacher and Merkle (2012), who refer to this quantity as the predictive probability of the model. It reflects the degree to which the model accurately predicts the observed data. In the population, the marginal likelihood of the model can be expressed as:

$$Pred(D|M) = \int f(D|\theta, M)p(\theta|M)d\theta \quad (4)$$

Where

D =Observed data

θ =Parameters for model M

$f(D|\theta, M)$ =Density function for the data given the parameters from the model

$p(\theta|M)$ =Prior density of model parameters

Maximum likelihood is often used to estimate the parameters (θ) in equation (4), but can be computationally intractable in many applied situations involving many model parameters of varying distributions (Beck & Yuen, 2004). Raftery (1993) and Konishi, et al. (2004) showed that if BIC is calculated as in equation (2), then for a given set of data (*D*) with a proposed model (*M*), this predictive probability takes the form:

$$Pred(D|M) \propto e^{\left(-\frac{BIC_M}{2}\right)} \quad (5)$$

Where

BIC_M =BIC value for the target model.

If the likelihood in equation (5) is calculated for two competing models given the same set of data, the ratio of those values is the Bayes Factor, and can be used in model selection. In this context, the Bayes Factor is calculated as:

$$BF = \frac{Pred(D|M_1)}{Pred(D|M_2)} \quad (6)$$

Where

$Pred(D|M_1)$ =Predictive probability of the data given model 1

$Pred(D|M_2)$ =Predictive probability of the data given model 2

Jeffreys (1961) proposed a set of rules for using this rule for the purpose of model selection. Essentially, when $BF \geq 1$ we have support for Model 1 and when $BF < 0.316$ there is substantial evidence against Model 1. Jeffreys suggested that BF values between 0.316 and 1 indicate minimal evidence against Model 1. Of course, given the relationships between $Pred(D|M)$ and BIC, the model selected as optimal based on the minimum BIC criterion will also be selected using the Bayes Factor. Perhaps more interestingly, the probability expressed in (4) can be used to calculate a

probability space for an entire set of models, as opposed to allowing only for comparisons of model pairs.

A brief description of the model posterior probabilities appears in the following text. The interested reader who would like more details is encouraged to read the Wu, et al. (2020) manuscript, where these issues are discussed in much greater detail. In the context of *K* candidate models, the posterior probability of model *i* can be calculated using the BIC (Wu, et al.).

$$posterior(M_i) = \frac{e^{-0.5(BIC_i - BIC_0)}}{\sum_{i=0}^{k} e^{-0.5(BIC_i - BIC_0)}} \quad (7)$$

Where

BIC_0 =BIC for the null model

BIC_i =BIC for candidate model *i*

Because in many cases BIC is a very large number, its exponentiated value will be extremely small. Therefore, in order to scale these values to a more useful format, the BIC for the null model is subtracted from that of the candidate model *i*. The null model is defined by setting the covariances among the observed variables to be 0 (Kline, 2016; Brown, 2015). It is important to note that prior probability for each model is assumed to be 1/*K*. As is discussed below, alternative prior probabilities can be employed by the researcher.

Once the posterior probabilities for the candidate models are calculated, they can be used to assess the relative likelihood of each model being optimal given the data at hand. More specifically, each model has an associated probability, which can be ordered from largest to smallest. The models within the credibility set are those where the sum of posterior probabilities is 0.95. This 95% model credibility interval can be used by the researcher to determine which should be explored in more depth. Consider an example involving 4 candidate models with the following ordered posterior probabilities for model 1 to model 4: 0..54, 0.42, 0.03, 0.01. In order to determine set of credibility models, we would sum the probabilities until achieving a value of 0.95. In this case the posterior probabilities for the first two models sum to 0.96 (0.54+0.42), meaning that they fall in the credibility set. Therefore, given that they are in the credibility set, the researcher would conclude that models 1 and 2 are the most likely to be optimal for the population from which the sample was drawn, and are therefore investigated more closely. Because they are not in the credibility set, models 3 and 4 are discarded as being unlikely to be appropriate representations of the relationships within the population.

This approach to model exploration for a given research problem presents a distinct alternative to the standard use of information indices such as BIC, in which the model with the smallest value is selected as being optimal, and little to no consideration is given regarding the relative fit of several models, or the uncertainty in the model selection process.

Bayesian model averaging

A closely related model selection and combination approach to the BIC based method described above is Bayesian model averaging (BMA). An early description of BMA was provided by Jeffreys (1939) and later expanded on by various authors (e.g., Hoeting, et al., 1999; Raftery, et al., 1997; Madigan & Raftery, 1994; Leamer, 1978). The basic idea behind the use of BMA is that it provides a framework for both model selection through the use of model probabilities, as well as obtaining more accurate predictions of one or more outcomes using a combination of multiple possible models weighted by their probabilities. In both cases, BMA provides the researcher with a better understanding of the uncertainty underlying the possible models, given theory and the data, than does the selection of a single model based upon information indices or hypothesis tests comparing model fit (Hoeting, et al., 1999).

The posterior probability of a particular model can be expressed as:

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{m=1}^{K} p(y|M_m)p(M_m)} \quad (8)$$

Where

$p(y|M_k)$ =Marginal likelihood of the data y given model M_k

$p(M_k)$ =Prior probability of model M_k

The value $p(M_k|y)$ represents the probability of model M_k based upon its fit to the data and its prior probability.

As described in Hinne, et al. (2020) BMA has several advantages for researchers in practice, including the estimation of model uncertainty, optimal predictions for an outcome variable using multiple models weighted by their probabilities, reduction of the impact of outlying observations, and robustness to model misspecification. Given these advantages, BMA is especially useful for situations in which prediction of the dependent variable values and parameter estimation are of most importance (Hinne, et al.). Conversely, the individual models themselves are typically of less importance in the context of BMA than in more traditional model selection contexts. Finally, as noted by previous authors (e.g., Hinne, et al., 2020; Hoeting, 1999) BMA can be relatively difficult to carry out in some context because of the complexity of parameter estimation and specification of $p(M_k|y)$, particularly when the number of candidate models is large. Given that the current study is focused on model selection, rather than prediction or parameter estimation, and the models being considered are relatively complex, BMA was not included in the current study.

Prior research examining BIC posterior probabilities

Wu, et al. (2020) conducted a series of simulation studies to investigate the performance of the technique described above for developing a credibility set of models based on the BIC. These simulations featured mediated and moderated models

involving observed variables. The first of the three simulation studies examined the performance of the BIC posterior probability approach when all 16 possible models based on three observed variables were used for both data generation and data analysis. Relationships among three observed variables were classified as small (0.14), medium (0.36), and large (0.51) based on Cohen's (1988) guidelines, and the sample size conditions were 100 and 1000. The outcome variables were the posterior probabilities for the candidate models. The goal of this first simulation study was to assess the accuracy of the posterior probabilities described in equation (5) in terms of identifying the most likely candidate model given the data generating model. In a second simulation study, the same basic data structure described above was used in data generation and analysis, but with only the small magnitude coefficients linking the variables and a wider array of sample sizes between 50 and 1000. This second study assessed the performance of the proposed approach for model poster determination in the more challenging conditions (small samples and weak relationships among variables). The third simulation study presented in Wu, et al. involved the use of priors other than the uniform. Only four models were considered in this last simulation, with the standard uniform prior (0.25 for each model), as well as correct strong priors (0.7 for the partial mediation model and 0.1 for the other models).

Based on the results of the three simulation studies described above, Wu, et al., (2020) reached several conclusions regarding the utility of the posterior probability approach for characterizing relative model fit. First, across studies the BIC posterior probabilities were accurate in terms of correctly identifying the data generating model with the largest posterior probability except for samples of 100 or fewer and weak relationships among the observed variables. In those cases, the posterior probabilities for the simpler models tended to be larger, even when they were not the data generating model. Second, the use of informative priors was associated with higher posterior probabilities for the model with the largest value. When these informative priors were incorrect the posterior probabilities for the incorrect models were somewhat lower in the context of larger samples and stronger coefficients linking the variables. However, for weak relationships and samples of 100 or fewer, incorrect informative priors were associated with larger values for the posterior probability of the incorrect model. Based on their findings, the authors concluded that the posterior probabilities are effective tools for identifying the most plausible model, and that they provide useful information about a full set of potentially credible models. They also suggested that researchers may consider whether the best fitting model has a posterior probability greater than a predetermined cut-off such as 0.5. If no posterior probability reaches this threshold, it may be that there is not a clearly defined best model for the data, regardless of which one has the lowest BIC value. Among recommendations coming from the study is that future research should consider a wider variety of models, including SEMs involving latent variables, which is the focus of the current work.

Methods

A simulation study approach was used to address the goals of this study. For each combination of study conditions, which are described below, a total of 1000 replications was used. Data generation and analysis were carried out using Mplus version 8 (Muthèn & Muthèn, 2018) and R version 4.02 (R Core Team, 2020). A variety of conditions were manipulated in order to investigate the performance of the posterior probability approach that is described above. The observed indicator variables were generated from the standard normal ($\mu = 0, \sigma = 1$) distribution. There were three observed indicators for each of the latent variables. For each indicator variable, the sum of the squared factor loading and the error variance was 1 for all simulations.

Data generation models

Three separate models were used for data generation, and are based on the research questions associated with the empirical example that is also a part of this study. These models include a fully mediated model (Figure 8), a partially mediated model (Figure 9), and moderated partially mediated model (Figure 10). For all combinations of the following conditions, each of these models was used to generate the data, and then each was fit to the resulting datasets. Thus, data were generated from the fully mediated, partially mediated, and moderated partially mediated models, and each of these was then fit to each of the datasets. These models were selected because they correspond to the hypothesized models in an actual research scenario represented in the empirical example.

Structural coefficient magnitudes

The structural coefficients used in this study are based on those presented in Wu, et al. (2020), and correspond to small (0.14), medium (0.36), and large (0.51) relationships based upon guidelines presented in Cohen (1988). In keeping with that earlier work, three structure coefficient conditions were used including all structure coefficients are 0.14 (small), some are 0.15 (medium), and large (all are larger than 0.14).

Factor loadings magnitudes

The factor loadings were manipulated to be 0.5, 0.7, and 0.9, representing weak to strong factor structure. The same loading values were used for all indicators for a give condition. As an example, in the 0.5 loading condition, all 12 factor loadings (3 indicators for each of 4 factors) were 0.5.

Sample size

Sample sizes were selected to reflect relatively small samples for latent variable SEM to large samples. These values were 100, 200, 300, 500, and 1000. These values also cover the range of those used by Wu, et al. (2020). Given that those authors found that sample size was an important factor in terms of the performance of the posterior probability approach, it was determined that this factor should also be considered in the current study.

Prior probabilities

Wu, et al showed that the choice of prior probability can have a marked impact on the resulting posterior probabilities. All data for this portion of the study were generated using the partial mediation model, and three levels of prior probability were used: (1) Uniform, (2) Correct informative, and (3) Incorrect informative. In the uniform prior condition each model had a 1/3 prior probability. For the correct informative prior condition, the partial mediation model was assigned a prior probability of 0.7, with the other two models having a prior probability of 0.15 each. In the incorrect informative prior condition, the fully mediated model had a prior probability of 0.7, which was incorrect given that the data were generated using the partial mediation model. The partial mediation and moderated partial mediation models each had a prior probability of 0.15 in the incorrect priors condition.

Study outcomes

The outcome variables of interest for this study were the posterior probabilities for each model based upon both BIC and aBIC. In order to determine which of the manipulated factors were related to the outcome variable, analysis of variance (ANOVA) was used, and both statistical significance and the η^2 effect size were used. A full factorial ANOVA model including the type of model, coefficient magnitude, factor loading magnitude, and sample size was fit to the outcome data for the BIC based results. For the portion of the study examining the use of informative priors, a separate ANOVA was applied using the same sets of variables listed above, with the addition of type of prior (uniform, informative correct, and informative incorrect).

Results

Partial mediation model

The ANOVA identified the interactions of the coefficient magnitude by fitted model by sample size ($F_{8,16} = 8.21, p < 0.001, \eta^2 = 0.80$), and factor loading value by fitted model by sample size ($F_{8,16} = 20.27, p < 0.001, \eta^2 = 0.84$) as being statistically significantly related to the posterior distribution of the BIC when the partial mediation model underlay the data. The posterior probabilities of the models by coefficient magnitude, sample size, and fitted model appear in Figure 1.

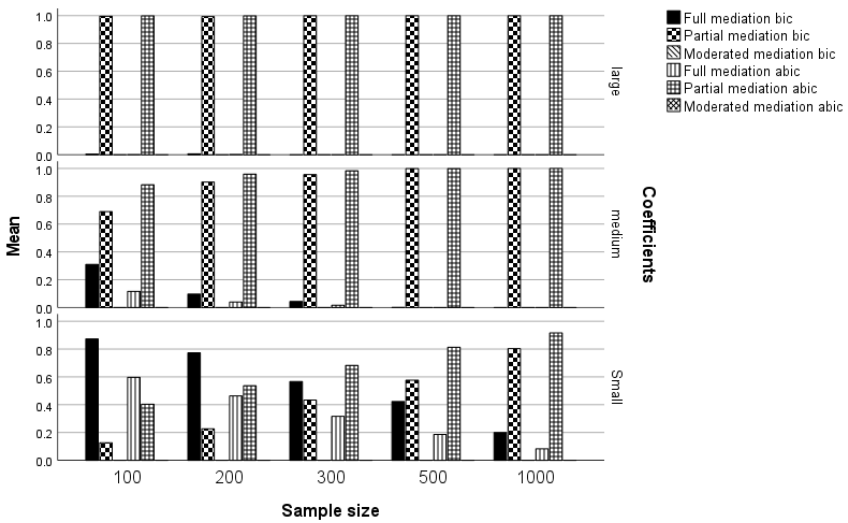


Figure 1: Posterior probabilities by coefficient magnitudes, sample size, and fitted model: Underlying partial mediation model

When the data generating structural coefficients were large, the posterior probabilities based on both BIC and aBIC for the partial mediation model were at or near 1.0 across sample sizes. Furthermore, when the coefficients were of moderate magnitude, the posterior probabilities based on both information indices were above 0.9 for samples of 200 or more. When the sample size was 100, the posterior probability for the partial mediation model remained largest for the partial mediation model, but for BIC it fell to 0.68, as compared to 0.32 for the fully mediation model. With regard to the aBIC, the posterior probability for the partial mediation model was 0.89, with the fully mediated model having a posterior value of 0.11.

Finally, when the coefficients were small, the posterior probabilities based on the aBIC for the partial mediation model was larger than that of the fully mediated model with samples of 200 or more. However, as the sample size decreased in value the posteriors of the two models converged, with that of the correct partially mediated model declining to 0.55 for a sample size of 200. When N=100, the fully mediated model posterior based on the aBIC was less than that of the full mediation model. When the BIC served as the basis for the model posteriors, the full mediation model had larger values for samples of 300 or smaller. In addition, the partial mediation model posterior probability never exceeded 0.8 in the small coefficient condition.

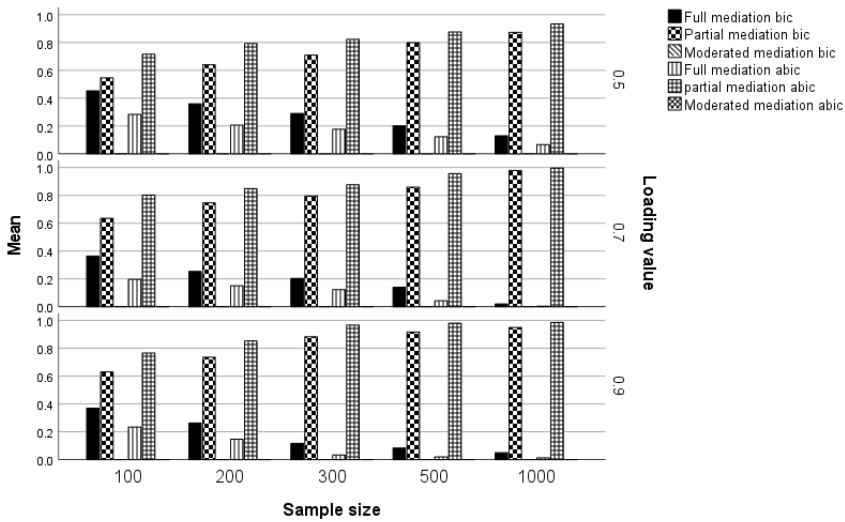


Figure 2: Posterior probabilities by factor loading value, sample size, and fitted model: Underlying partial mediation model

Figure 2 includes the posterior probabilities by factor loading, sample size, and the fitted model. Across sample sizes and factor loading values the partial mediation posterior probability was largest for both BIC and aBIC. However, with smaller samples and lower factor loading values, the difference between the posterior probability of the partial mediation and full mediation models was smaller; i.e., the posterior of the fully mediated model increased in value vis-à-vis larger samples and larger factor loadings. When the sample size was 300 or more, the posterior probability of the partial mediation model was greater than 0.8 except when the factor loading was 0.5. For samples smaller than 300, the partial mediation model posterior was never greater than 0.8, regardless of the factor loading magnitude.

Full mediation model

The ANOVA identified the interactions of structural coefficients by factor loading values by fitted model ($F_{4,16} = 12.06, p < 0.001, \eta^2 = 0.75$), and sample size by fitted model ($F_{4,16} = 81.76, p < 0.001, \eta^2 = 0.95$) as being statistically significantly related to the posterior probabilities when the fully mediated model was used to generate the data.

Table 1: Posterior probabilities by sample size and fitted model: Underlying fully mediated model

Sample size	Full mediation BIC	Partial mediation BIC	Moderated mediation BIC	Full mediation aBIC	Partial mediation aBIC	Moderated mediation aBIC
100	0.85	0.15	0	0.57	0.43	0
200	0.89	0.11	0	0.66	0.34	0
300	0.92	0.08	0	0.72	0.28	0
500	0.94	0.06	0	0.77	0.23	0
1000	0.96	0.04	0	0.84	0.16	0

Table 1 includes the mean posterior probabilities by sample size and fitted model. From these results, it is clear that when the data were generated from the fully mediated model, the posterior probabilities for the BIC were higher for the fully mediated model than were those based on the aBIC, across sample sizes. In addition, as the sample increased in value, the probability of the fully mediation model increased for both methods.

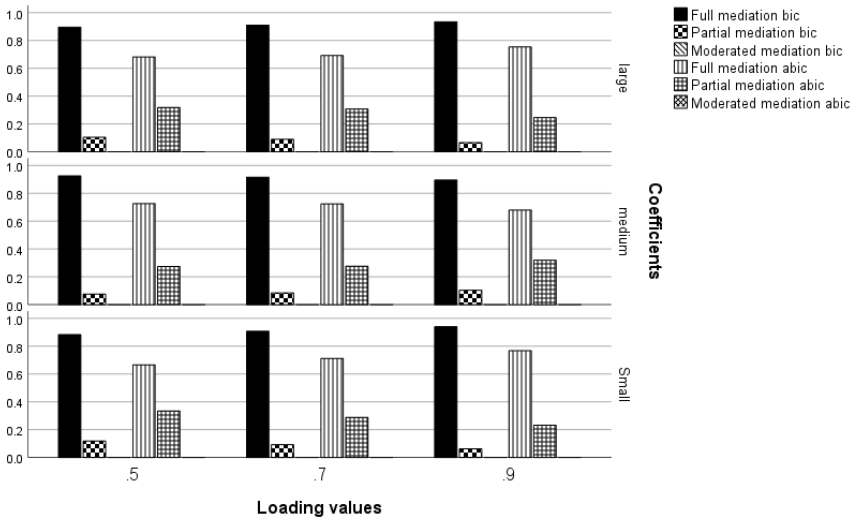


Figure 3: Posterior probabilities by structural coefficient magnitude, factor loading value, and fitted model: Underlying full mediation model

Figure 3 includes the posterior probabilities by factor loading values, structural coefficient magnitudes, and the fitted models. As was true in Table 1, the posterior probabilities for the full mediation model was largest across all study conditions. In addition, the probabilities for the correct full mediation model were larger when based on the BIC as compared to the aBIC. In addition, for both the BIC and aBIC based posteriors, the value for the full mediation model increased in value with increases in the loading values and coefficient magnitudes. Finally, the impact of loading and coefficient magnitude on the posterior probabilities for the full mediation model was stronger for the aBIC based probabilities.

Moderated mediation model

The results of the ANOVA identified the interactions of sample size by fitted model ($F_{4,16} = 10.25, p < 0.001, \eta^2 = 0.72$) and factor loading values by coefficient magnitude by fitted model ($F_{4,16} = 10.43, p < 0.001, \eta^2 = 0.72$) as statistically significantly associated with the posterior probabilities when the underlying model involved moderated mediation. Across sample sizes and fitted models, the posterior probabilities for the moderated mediation model were the largest regardless of whether the BIC or aBIC was used to calculate them (Table 2).

Table 2: Posterior probabilities by sample size and fitted model: Underlying moderated mediation model

Sample size	Full mediation BIC	Partial mediation BIC	Moderated mediation BIC	Full mediation aBIC	Partial mediation aBIC	Moderated mediation aBIC
100	0.03	0.10	0.87	0.04	0.11	0.85
200	0.03	0.07	0.90	0.03	0.08	0.89
300	0.02	0.04	0.94	0.03	0.08	0.92
500	0.02	0.02	0.96	0.02	0.03	0.95
1000	0	0	1.00	0	0	1.00

In addition, the posteriors for the correct model were slightly larger for the BIC. Finally, the posterior of the moderated mediation model increased in value concomitantly with increases in sample size.

The posterior probabilities based on BIC and aBIC by fitted model, factor loadings, and structural coefficients appear in Figure 4.

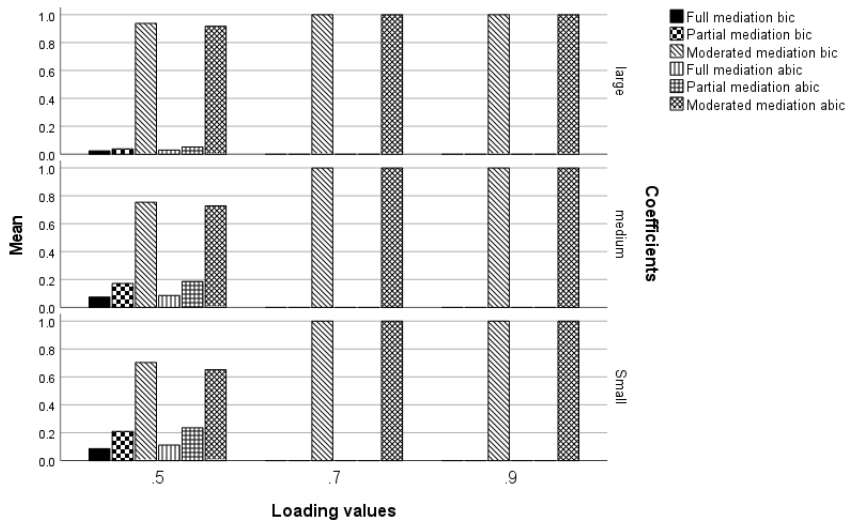


Figure 4: Posterior probabilities by structural coefficient magnitude, factor loading value, and fitted model: Underlying moderated mediation model

These results reinforce the finding from Table 2 that across conditions the posterior probabilities for the (correct) moderated mediation model were largest. When the factor loadings were 0.7 or larger, the posteriors for this model were at or near 1.0, regardless of the structural coefficients. Furthermore, when the coefficients were of large magnitude, the posterior for the moderated mediation model was above 0.9 when the structure coefficients were large. When the coefficients were medium or small, the moderated mediation model posteriors were below 0.8. In these cases, the partial mediation model exhibited larger posterior probabilities than did the full mediation model.

Partial mediation model with informative priors

Results of the ANOVA revealed that the interactions of prior probability type, structural coefficient magnitudes, sample size, and fitted model (), and prior probability type, factor loading value, and fitted model () were significantly associated with the posterior probabilities. Based on the results in Figure 5, it appears that when the structural coefficients were large, the posterior probabilities for the (correct) partial mediation model were the largest for both BIC and aBIC, across sample sizes. In addition, when informative priors were used, the posterior probabilities for the partial mediation model were greater than 0.8 across sample sizes in the medium structural coefficients case. Indeed, when informative priors were used, the posterior probabilities based on the aBIC were 0.8 or larger across all study conditions. In contrast, when the structural coefficients were small and the sample was less than 1000, the posterior probabilities based on the BIC for the partial mediation model were less than 0.8.

When incorrect informative priors were applied, the BIC based posterior probabilities for the (incorrect) full mediation model was largest for samples of 500 or fewer when the structural coefficients were small, and 300 or fewer for the aBIC posterior probabilities. In addition, When the structural coefficients were of medium magnitude, the posteriors for the partial mediation model were the largest across sample sizes when based on the aBIC. For the posterior probabilities calculated using the BIC, the posteriors for the partial mediation model was largest for samples of 200 or more. For the small structural coefficient condition the posterior probabilities for the (correct) partial mediation model based on the aBIC were larger than those for the BIC.

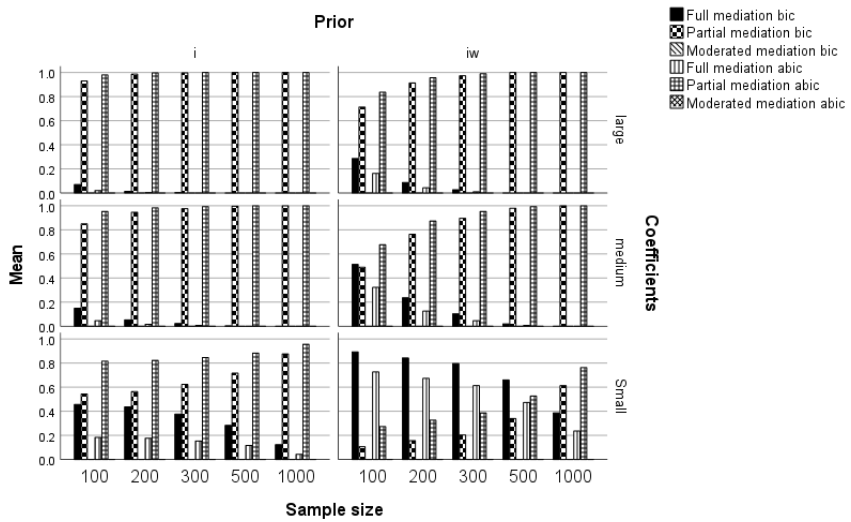


Figure 5: Posterior probabilities by prior type, structural coefficient magnitude, sample size, and fitted model: Underlying partial mediation model

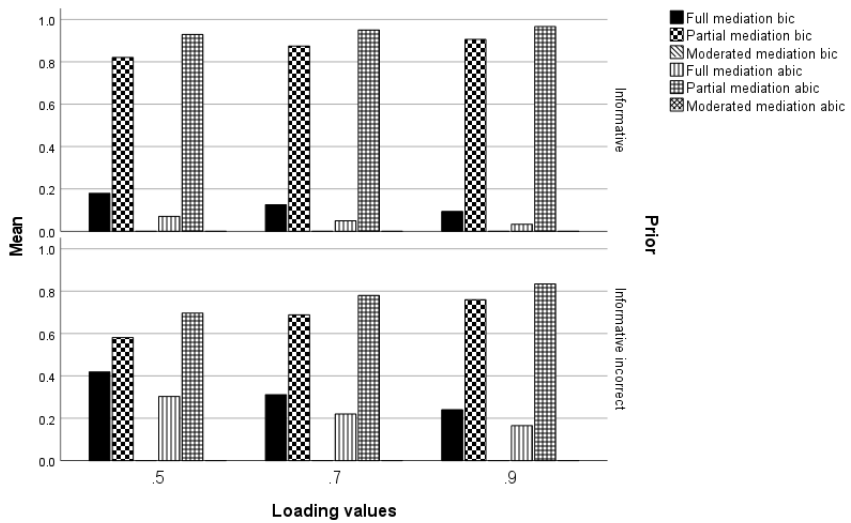


Figure 6: Posterior probabilities by prior type, factor loading value, and fitted model: Underlying partial mediation model

Figure 6 displays the posterior probabilities by prior type, factor loading value and the fitted model. As was evident in Figure 5, the posterior probabilities based on the aBIC were uniformly larger than those based on the BIC. In addition, for both approaches, the posteriors for the partial mediation model were larger in the correct informative prior condition. For both approaches, the posterior probabilities for the partial mediation model were larger for models with larger factor loadings. The posterior probabilities for the full mediation model were second largest across conditions.

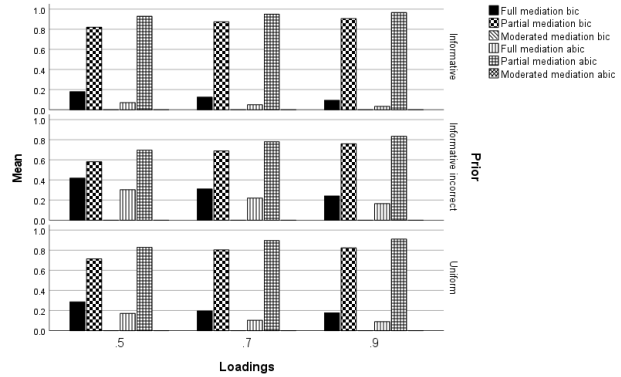
Comparison of posterior probabilities by prior for partial mediation model

The three panels of Figure 7 display the posterior probabilities for each model based on the BIC and aBIC by the prior distribution, and factor loadings, sample size, and structural coefficient magnitude.

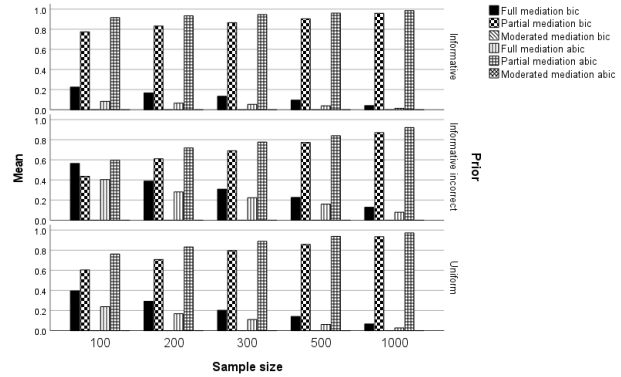
The goal of this examination was to compare the performance of the fitted models under different prior probability conditions. With respect to factor loading values (Panel 1), the posterior probabilities for (correct) partial mediation model were largest in the informative prior case and smallest in the incorrect informative prior. These differences were magnified for smaller factor loadings, for which the use of informative priors was particularly effective in terms of the posterior probabilities being largest for the partial mediation model.

Results in Panel 2 of Figure 7 reveal a similar overall pattern of results, in which the use of an informative prior yielded higher posterior probabilities for the partial mediation model than was the case for the other two prior types, particularly the incorrect informative priors. These differences were most magnified for smaller samples, whereas for sample sizes of 500 or more the results for the informative and uniform priors were similar to one another. On the other hand, the informative prior consistently yielded the lowest posterior probabilities for the partial mediation model across sample sizes. Finally, Panel 3 of Figure 7 shows that when the structure coefficients were large, the posterior probabilities for the partial mediation model were nearly 1.00 for both the informative and uniform priors, and somewhat lower for the incorrect informative priors. When the structure coefficients were of medium magnitude, the posteriors for the partial mediation model in both the informative and uniform prior conditions were above 0.9, but were just above 0.8 for the incorrect informative priors. Finally, when the coefficients were small, the BIC based posterior probabilities were largest for the partial mediation model only for the informative prior condition. When considering the aBIC posteriors, the partial mediation model had the largest posteriors for both the informative and uniform priors. However, the partial mediation posterior probability in the informative prior case were above 0.8 in the small coefficient case, whereas for the uniform priors the posteriors of the partial mediation model were approximately 0.7 in the small coefficient case.

Factor loadings



Sample size



Structure coefficient magnitude

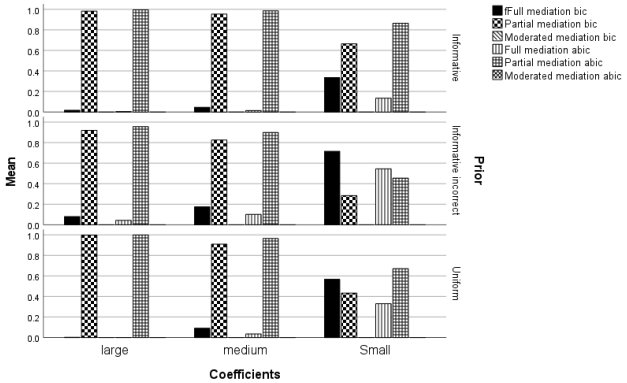


Figure 7: Posterior probabilities by prior type, factor loading value, sample size, structural coefficient magnitude, and fitted model: Underlying partial mediation model

Empirical example

In order to demonstrate the application of the Bayesian approach to interpreting model fit based on the BIC and aBIC, a series of SEMs were used with a dataset collected from a sample of 417 college students. Among other measures, the participants completed items on the Adult Temperament Questionnaire (ATQ; Evans & Rothbart, 2007), which includes subscales measuring extraversion, negative affect, effortful control, and orienting sensitivity. A series of SEMs were fit around these variables, and the Bayesian BIC approach to obtaining model posterior probabilities in order to identify those that are most probable, given the data. A total of four models were fit to the data, and for each the posterior probabilities was calculated based upon the BIC and aBIC, respectively. The models appear in Figures 8-10, and reflect partial and full mediation, and a partially mediated model with moderation. The models were fit using Mplus, version 8 (Muthèn & Muthèn, 2018), and reflect the set of theoretically justified models for these data.

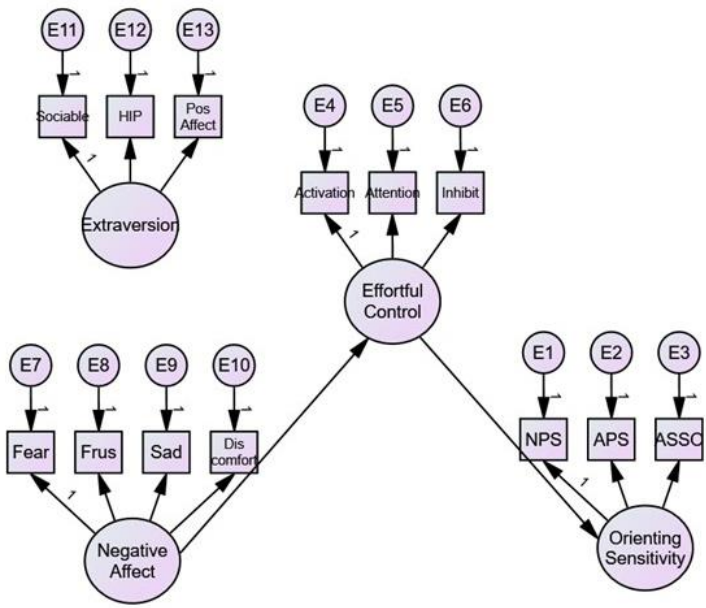


Figure 8: Full mediation model for empirical example

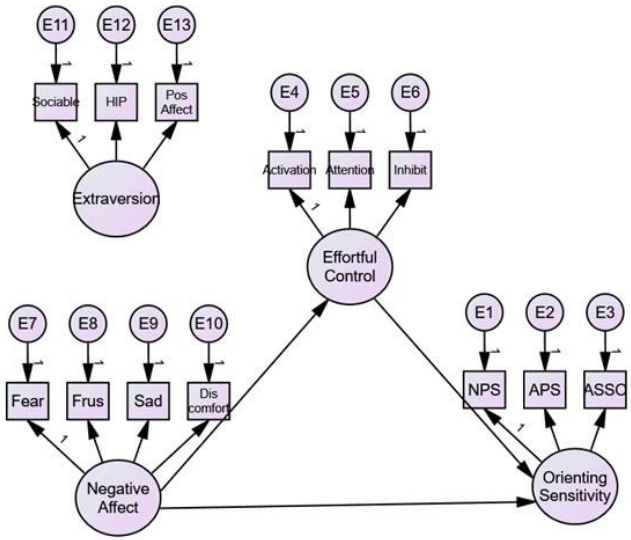


Figure 9: Partial mediation model for empirical example

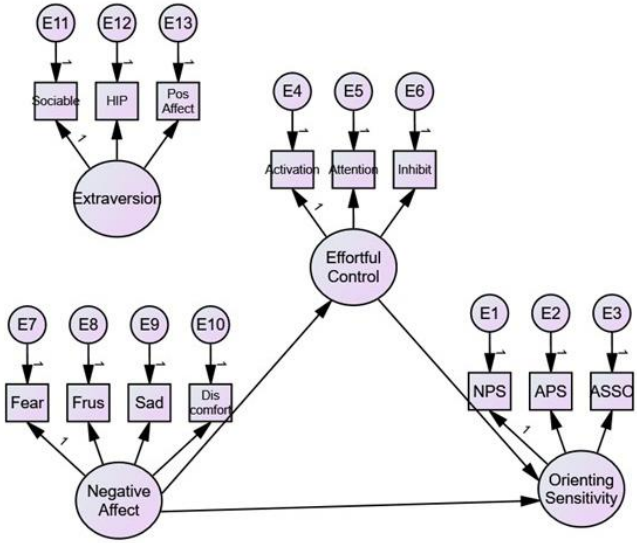


Figure 10: Moderated partial mediation model for empirical example

Table 3 includes the BIC, aBIC, and posterior probabilities for each model.

Table 3: BIC, aBIC, and posterior probabilities for each model in the empirical example

Model	BIC	aBIC	BIC Posterior	aBIC posterior
Full mediation	10189.241	10097.216	0	0
Partial mediation	10163.083	10067.885	0.85	0.54
Moderated partial mediation	10166.544	10068.173	0.15	0.46

If we were to use the BIC and aBIC only, we would conclude that the partial mediation model yielded the best fit and would move on to describe the model coefficients, with no further examination of the competing models. In contrast, by relying on the posterior probabilities, we gain a greater understanding of the nuances among the various models. First considering the BIC posterior probabilities only, we would conclude that it is very likely that the correct model involves partial, as opposed to full, mediation. In addition, it seems unlikely that there is a moderation effect present in the model, though we could not rule it out entirely. However, we would be inclined to primarily focus on the partial mediation model. If we were to rely on the posterior distributions derived from the aBIC statistic, we would also conclude that partial mediation is very likely to be present. However, in contrast to what we found based on the BIC only, the aBIC posterior probabilities for the partial and moderated partial mediation models were relatively close to one another. Thus, although the partial mediation model has a higher posterior than does the moderated partial mediation model, both appear to be plausible given the data, and therefore deserve continued consideration.

The structural coefficients and direct/indirect effects for the partial mediation and moderated partial mediation models appear in Table 4. An examination of these results reveals that for either model, the direct relationships between effortful control and negative affect with orienting sensitivity were statistically significant, based on the 95% confidence intervals. In both cases, the relationships were positive indicating that higher levels of effortful control and negative affect were associated with higher levels of orienting sensitivity. The indirect effect between negative affect and effortful control was not statistically significant, though the total effect was. Finally, the moderated effect due to extraversion was not statistically significant. Taken together, these results show that orienting sensitivity was positively associated with both effortful control and negative affect, and that these relationships were direct in nature. The posterior probabilities reflected that the partial mediation model was the most likely given the data, and the coefficient results support this finding in that the moderator effect was not statistically significant.

Table 4: Parameter estimates, standard errors, and 95% confidence intervals for structural coefficients for the partial mediation and moderated partial mediation models

Term	Coefficient	Standard error	95% confidence interval
Partial mediation model			
Effortful control direct effect	0.21	0.11	-0.01, 0.42
Negative affect direct effect	0.32	0.12	0.08, 0.56
Negative affect->Effortful control	-0.13	0.19	-0.49, 0.24
Negative Affect indirect effect	-0.03	0.03	-0.09, 0.03
Negative Affect total effect	0.29	0.12	0.06, 0.53
Moderated partial mediation model			
Effortful control direct effect	0.21	0.11	-0.004, 0.42
Negative affect direct effect	0.32	0.12	0.09, 0.55
Negative affect->Effortful control	-0.13	0.16	-0.45, 0.19
Negative Affect indirect effect	-0.03	0.03	-0.09, 0.03
Negative Affect total effect	0.29	0.11	0.07, 0.51
Extraversion X Negative affect	0.01	0.07	-0.12, 0.14

Discussion

The goal of this study was to extend upon the work of Wu, et al., (2020) by applying BIC based posterior probability approach to model comparison to the case of latent variable SEM. In addition, the utility of posterior probabilities based on aBIC was also investigated. The results of the simulations described above yielded several findings of interest. First, when the coefficients relating the latent variables to one another were medium or large, posterior probabilities based on both BIC and aBIC were largest for the data generating model. However, when the coefficients were small in magnitude and the sample size was 300 or fewer, the posterior probabilities for the simpler full mediation model tended to be larger than for the more complex models, even when those more complex models were used to generate the data. This result is similar to findings for observed variable models that was reported in Wu, et. al. A second major finding of the current study was that cases with smaller factor loadings, coupled with smaller sample sizes led to larger posterior probability values for the full mediation model when the data were generated from the partial mediation model. However, unlike with small structural coefficients, the full mediation model never had the highest posterior probability even for samples as small as 100 and factor loadings of 0.5, when either the partial or moderated partial mediation models were used to generate the data. In other words, it appears that, although factor loading magnitude does

influence model posterior probabilities, this impact is not as strong as for coefficient magnitude.

Third, the use of aBIC to calculate model posterior probabilities yielded larger values for the partial mediation model when it was the data generating model than did BIC, but such was not the case for the full mediation model, where the BIC posterior probabilities for the correct (full mediation model) were largest. Finally, the use of informative priors can have a positive impact for researchers when they are applied correctly. This is particularly the case for smaller sample sizes and weaker factor loadings, where the use of correct informative priors led to higher posterior probabilities for the correct model than was the case for the uniform priors. However, when informative priors are incorrect (i.e., larger priors are associated with an incorrect model), the incorrect model will have relatively large incorrect priors when the sample size is less than 300, the structure coefficients are small, or the factor loadings are 0.5. In those cases, the posterior probabilities for the incorrect model with the larger prior probability was itself relatively large, which could mislead researchers into assuming that there was greater model uncertainty than is actually the case.

Implications for practice

The results of this study provide several implications for practice. First, it appears that the use BIC to calculate Bayesian posterior probabilities for a set of models that was described by Wu, et al. (2020) in the context of observed variable path models does translate well to latent variables as well. Generally speaking, the data generating models had the largest posterior probabilities in the current study. However, one major caveat that researchers using this approach need to keep in mind is that the strength of relationships among the latent variables, and the factor structure itself will impact the utility of these posterior probabilities. When the factor structure is relatively weak and/or the relationships among the variables is small, the simpler models will have higher posterior probabilities than would be reflected in the data generating processes. Thus researchers need to consider the posterior probabilities in light of the factor loading and structure coefficient magnitudes. Similarly, sample size is also an important consideration given the results presented above. When it is below 300, the posterior probabilities appear to favor simpler models, particularly when comparisons are between partial and full mediation. Thus, researchers should keep in mind that, in the context of latent variable modeling, smaller sample sizes will be associated with relatively larger probabilities for simpler models even when these are not actually correct. Indeed, it may be the case that for very small samples (e.g., 100) the use of these Bayesian posterior probabilities may not be reasonable.

A third, related implication from this research is that the use of informative priors can ameliorate the deleterious impact of small samples and weak models on the posterior probabilities. When the informative priors are correct (i.e., the data generating model has the larger prior probability) the resulting posterior probabilities for these correct models will generally be larger, regardless of the sample size, an factor loading and

structural coefficient magnitudes. However, the converse is also true. Namely, when the largest informative prior is applied to the incorrect model, and the sample size is small or the model is weak, the posterior probability for that incorrect model will be large, thereby leading to potential confusion and/or incorrect conclusions regarding model uncertainty. Considering these latter two implications together, if researchers have relatively high confidence that one of the proposed models is more likely to underlie the observed data, then the use of informative priors can be helpful, particularly in traditionally difficult situations for latent variable SEM; i.e., with small samples and weak models. However, in those very same situations, the researcher should probably avoid using informative priors if their prior information about the correctness of one of the models is relatively low.

A final implication of this study is that the number of candidate models will very likely impact the likelihoods of the individual models through the size of the denominator. The more candidate models that are considered, the larger the denominator and potentially the lower the marginal likelihood of any single model. Wu, et al. (2020) acknowledge this fact in their work, and argue that researchers should therefore only consider theoretically meaningful and plausible models. They also note that even when a relatively large number of models are considered, the posterior likelihoods should still reflect the most important models relative to one another. Nonetheless, researchers should be aware that the number of candidate models will have an impact on the individual model likelihoods.

In summary, the use of the posterior probability approach to describing and comparing model fit that was described in Wu, et al. (2020) and extended here to the context of latent variable modeling appears to be a viable option for researchers. It is particularly useful when the factor loadings are 0.7 or larger, and for samples of 300 or more. In addition, when researchers have good prior information regarding the relative likelihood of the models being considered, this can be applied in the form of prior probabilities, and would be particularly useful in traditionally difficult modeling situations. Care must be taken in using such priors however, particularly in those difficult cases.

Directions for future research

The results of the current study point to several directions for additional research. First, additional latent structure situations should be considered, including more complex structural relationships among the latent variables, more latent variables, and different factor loading conditions. Second, future work should consider the use of categorical indicator variables. In many situations, scales consisting of ordinal items are used in social science research. Therefore, future work should examine how the posterior probability approach performs when the indicators are ordinal and dichotomous in nature. Additional research should also continue to investigate the use of aBIC for calculating the posterior probabilities. Results of the current work are indeterminate with regard to the use of the aBIC for this purpose. In some scenarios it appears to yield more accurate posteriors, but this was not true across all conditions. Thus, more

work needs to examine this issue. Future research should compare the method that was the focus of this work with BMA. One hurdle that such work would need to overcome is the potential difficulty in model parameter estimation for the SEMs. However, it would be useful to researchers if traditional BMA could be compared to the BIC method considered here. Finally, future work should also examine the performance of the posterior probabilities for models with relationships among a mix of observed and latent variables, such as the MIMIC model.

References

Aguado, J., Luciano, J.V., Cebolla, A., Serrano-Blanco, A., Soler, J., & Garcia Campayo, J. (2015). Bifactor analysis and construct validity of the Five Facet Mindfulness Questionnaire (FFMQ) in non-clinical Spanish samples. *Frontiers in Psychology, 6*, 1-14.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19 (6)*, 716–723.

Aveh, J.K. (2015). Travellers' acceptance of consumer-generated media: An integrated model of technology. *Computers in Human Behavior, 48*, 173-180.

Brown, T.A. (2015). Confirmatory Factor Analysis for Applied Research. New York: The Guilford Press.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Evans, D.E. & Rothbart, M.K. (2007). Development of a model for adult temperament. *Journal of Research in Personality, 41*, 868-888.

Hinne, M., Gronau, Q.F., van den Bergh, D., & Wagenmakers, E-J. (2020). A conceptual introduction to Bayesian Model Averaging. Advances in Methods and Practices in *Psychological Science, 3(2),* 200-215.

Hoeting, J.A., Madigan, D., Raftery, A.E., & Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*, 382-417.

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.

Kline, R. B. (2016). Principles and practice of structural equation modeling. New York: The Guilford Press.

Konishi, S., Ando, T., & Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika, 9(1),* 27-43.

Leamer, E.E. (1978). *Specification Searches*. New York: Wiley.

Madigan, D. & Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association, 89*, 1535-1546.

Muthèn, L.K. & Muthèn, B.O. (2018). Mplus User's Guide (Eighth Edition). Los Angeles, CA: Muthèn & Muthèn.

Preacher, K.J. & Merkle, E.C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17(1),* 1-14.

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

Raftery, A.E., Madigan, D., & Hoeting, J.A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92,* 179-191.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163.

Raftery, A.E. (1993). Bayesian model selection in structural equation models. In K.A. Bollen & J.S. Long (Eds.), Testing Structural Equation Models (pp. 163-180). Thousand Oaks, CA: Sage publications.

Schwarz, Gideon E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6 (2),* 461–464.

Sclove SL. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika. 52(3),* 333–343.

Steegh, A., Hoffler, T., Hoft, L., & Parchmann, I. (2020). First steps toward gender equity in the chemistry Olympiad: Understanding the role of implicit gender-science stereotypes. *Journal of Research in Science Teaching, 58(1),* 40-68.

Ulper, H., Cetinkaya, G., & Dikici, A. (2018). Examination of factors affecting students' reading comprehension achievement with structural equation modeling. *International Journal of Assessment Tools in Education, 5(3),* 428-442.

Wu, H., Cheung, S.F., & Leung, S.O. (2020). Simple use of BIC to assess model selection uncertainty: An illustration using mediation and moderation models. Multivariate Behavioral Research, 55(1), 1-16.