Structural Equation Modeling: A Multidisciplinary Journal, 00: 1–11, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2015.1092088

Bayesian Model Averaging Over Directed Acyclic Graphs With Implications for the Predictive Performance of Structural Equation Models

David Kaplan and Chansoon Lee
University of Wisconsin–Madison

This article examines Bayesian model averaging as a means of addressing predictive performance in Bayesian structural equation models. The current approach to addressing the problem of model uncertainty lies in the method of Bayesian model averaging. We expand the work of Madigan and his colleagues by considering a structural equation model as a special case of a directed acyclic graph. We then provide an algorithm that searches the model space for submodels and obtains a weighted average of the submodels using posterior model probabilities as weights. Our simulation study provides a frequentist evaluation of our Bayesian model averaging approach and indicates that when the true model is known, Bayesian model averaging does not yield necessarily better predictive performance compared to nonaveraged models. However, our case study using data from an international large-scale assessment reveals that the model-averaged submodels provide better posterior predictive performance compared to the initially specified model.

Keywords: Bayesian model averaging, Bayesian structural equation modeling, prediction

The distinctive feature that separates Bayesian statistical inference from its frequentist counterpart is its focus on describing and modeling all forms of uncertainty. The primary focus of uncertainty within a Bayesian analysis concerns prior knowledge about model parameters. In the Bayesian framework, all unknowns are described by probability distributions. Parameters constitute the central focus of statistical modeling, and because they are, by definition, unknown, Bayesian inference encodes background knowledge about parameters by means of prior distributions.

Within the Bayesian framework, parameters are not the only unknown elements. In fact, the Bayesian framework recognizes that models themselves possess uncertainty insofar as a particular model is typically chosen based on prior knowledge of the problem at hand and the variables that have been used in previously specified models. This form of uncertainty often goes unnoticed. Quoting Hoeting, Madigan, Raftery, and Volinsky (1999):

> Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are. (p. 382)

The current approach to addressing the problem of model uncertainty from a Bayesian perspective lies in the method of *Bayesian model averaging* (BMA).

BMA has had a long history of theoretical and practical applications. Early work by Learner (1978) laid the foundation for BMA. Fundamental theoretical work on BMA was conducted in the mid-1990s by Madigan and his colleagues (e.g., Hoeting et al., 1999; Madigan & Raftery, 1994; Raftery, Madigan, & Hoeting, 1997). Additional theoretical work was conducted by Clyde (1999). Draper (1995) discussed how model uncertainty can arise even in the context of experimental designs, and Kass and Raftery (1995) provided a review of BMA and the costs of ignoring model uncertainty. A more recent review of the general problem of model uncertainty can be found in Clyde and George (2004).

Correspondence should be addressed to David Kaplan, Department of Educational Psychology, University of Wisconsin–Madison, 1025 West Johnson Street, Madison, WI 53706. E-mail: david.kaplan@wisc.edu

Practical applications of BMA can be found across a wide variety of domains. A perusal of the extant literature shows applications of BMA to economics (e.g., Fernández, Ley, & Steele, 2001), bioinformatics of gene express (e.g., Yeung, Bumbarner, & Raftery, 2005), and weather forecasting (e.g., Sloughter, Gneiting, & Raftery, 2013), to name just a few. Indeed, of relevance to this article is the earlier work of Madigan and Raftery (1994), who applied BMA to so-called recursive causal models. However, our review of the extant literature suggests that BMA applied to structural equation modeling (SEM) has yet to be fully developed or studied under controlled conditions, nor is there readily available software to conduct BMA for SEM models. BMA has been implemented in the R software program "BMA" (Raftery, Hoeting, Volinsky, Painter, & Yeung, 2014) and can be applied to regression models, generalized linear models, and survival models.

The purpose of this article is to develop and apply BMA for SEM. Open source and proprietary software for conducting Bayesian SEM are now widely available, but, quite naturally, these programs focus on quantifying uncertainty in the model parameters and offer flexibility in encoding informative and noninformative priors. Bayesian approaches to structural equation model evaluation are also available in these software programs through such methods as posterior predictive checking (Gelman, Carlin, Stern, & Rubin, 2003). However, these software programs do not account for model uncertainty. We argue that the Bayesian framework for SEM estimation should also account for model uncertainty, and to that end, this article develops and assesses the BMA approach to addressing modeling uncertainty in SEM and provides open source R code to conduct such an analysis.

The organization of this article is as follows. In the next section we briefly describe Bayesian SEM. This is then followed by an outline of the method of BMA with an additional discussion of Occam's window following closely the work of Madigan and his colleagues (Hoeting et al., 1999; Madigan & Raftery, 1994; Raftery et al., 1997). Next, we describe our algorithm for searching the space of possible submodels of a general structural equation model. This is followed by a description and results of our simulation study wherein we describe the outcome of interest, namely improved posterior predictive performance of a Bayesian structural equation model. A case study is provided based on data from the 2009 cycle of the *Program for International Student Assessment* (PISA 2009; Organization for Economic Cooperation and Development [OECD], 2010) Results follow, and the article concludes with a discussion of the implications of our study for the practice of SEM as well as future research directions.

SPECIFICATION OF A BAYESIAN STRUCTURAL EQUATION MODEL

We focus our attention on structural equation models among observed variables. Following Kaplan and Depaoli (2012) and Kaplan (2014), a structural equation model can be specified as follows. Let

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta} \tag{1}$$

where $\mathbf{y}$ is a vector of manifest endogenous variables and $\mathbf{x}$ is a vector of observed exogenous variables with covariance matrix $\boldsymbol{\Phi}$. Further, let $\boldsymbol{\alpha}$ be a vector of structural intercepts, $\mathbf{B}$ is a matrix of structural regression coefficients relating the observed variables $\mathbf{y}$ to other observed endogenous variables, $\boldsymbol{\Gamma}$ is a matrix of structural regression coefficients relating the endogenous variables to observed exogenous variables $\mathbf{x}$, and $\boldsymbol{\zeta}$ is a vector of structural disturbances with covariance matrix $\boldsymbol{\Psi}$ assumed to be diagonal.

Conjugate Priors for SEM Parameters

To specify prior distributions on all model parameters, it is notationally convenient to arrange the model parameters as sets of common conjugate prior distributions. Parameters with the subscript *norm* follow a normal distribution, whereas those with the subscript *IW* follow a inverse Wishart distribution. Let $\boldsymbol{\theta}_{norm} = \{\boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\Gamma}\}$ be the vector of free model parameters that are assumed to follow a normal distribution, and let $\boldsymbol{\theta}_{IW} = \{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}$ be the vector of free model parameters that are assumed to follow the inverse Wishart distribution. Formally, we write

$$\boldsymbol{\theta}_{norm} \sim N(\boldsymbol{\mu},\ \Omega), \tag{2}$$

where $\boldsymbol{\mu}$ and Ω are the mean and variance hyperparameters, respectively, of the normal prior. For blocks of variances and covariances in $\boldsymbol{\Xi}$ and $\boldsymbol{\Psi}$, we assume that the prior distribution is inverse Wishart; that is,[1]

$$\boldsymbol{\theta}_{IW} \sim IW(\mathbf{R}, \delta), \tag{3}$$

where $\mathbf{R}$ is a positive definite matrix, and $\delta > q - 1$, where q is the number of observed variables. Different choices for $\mathbf{R}$ and δ will yield different degrees of "informativeness" for the inverse Wishart distribution.

MCMC Sampling for Bayesian SEM

Without question, the growth of interest in Bayesian methods is due to the availability of powerful statistical software that enables the application of Markov chain Monte Carlo (MCMC) algorithms such as the Metropolis algorithm, the Metropolis–Hastings algorithm, and the Gibbs sampler (see, e.g., Gilks, Richardson, & Spiegelhalter, 1996). For the purposes of this section, we discuss Gibbs sampling for Bayesian SEM and BMA. The Bayesian approach begins by

[1] Note that in the case where there is only one element in the block, the prior distribution is assumed to be inversegamma; that is, $\boldsymbol{\theta}_{IW} \sim IG(a, b)$.

considering $\boldsymbol{\eta}$ as missing data. Then, the observed data $\mathbf{y}$ are augmented with $\boldsymbol{\eta}$ in the posterior analysis. The Gibbs sampler then produces a posterior distribution $[\boldsymbol{\theta}_{norm}, \boldsymbol{\theta}_{IW}, \boldsymbol{\eta}|y]$ via the following algorithm. At the $(s+1)^{th}$ iteration, using current values of $\boldsymbol{\eta}^{(s)}$, $\boldsymbol{\theta}^{(s)}_{norm}$ and $\boldsymbol{\theta}^{(s)}_{IW}$,

1. sample $\boldsymbol{\eta}^{(s+1)}$ from $p\left(\boldsymbol{\eta}|\boldsymbol{\theta}^{(s)}_{norm}, \boldsymbol{\theta}^{(s)}_{IW}, y\right)$ (4)

2. sample $\boldsymbol{\theta}^{(s+1)}_{norm}$ from $p\left(\boldsymbol{\theta}_{norm}|\boldsymbol{\theta}^{(s)}_{IW}, \boldsymbol{\eta}^{(s+1)}, y\right)$ (5)

3. sample $\boldsymbol{\theta}^{(s+1)}_{IW}$ from $p\left(\boldsymbol{\theta}_{IW}|\boldsymbol{\theta}^{(s+1)}_{norm}, \boldsymbol{\eta}^{(s+1)}, y\right)$. (6)

In words, Equations 4 through 6 first require start values for $\boldsymbol{\theta}^{(0)}_{norm}$and $\boldsymbol{\theta}^{(0)}_{IW}$ to begin the MCMC generation. Then, given these current start values and the data $\mathbf{y}$ at iteration s, we generate $\boldsymbol{\eta}$ at iteration $s+1$. Given the latent data and observed data, the algorithm produces the posterior distribution of the measurement model and structural model parameters in Equations 4 through 6, respectively.

Bayesian SEM Model Evaluation and Selection

SEM, by its very nature, involves the specification estimation, and testing of models that purport to represent the underlying structure of the data. As in the case of model evaluation in frequentist SEM, it is important to evaluate the quality of a Bayesian SEM model vis-a-vis the data. One method available to evaluate Bayesian SEM models involves *posterior predictive checking* and the corresponding *Bayesian p-value,* which uses the posterior predictive distribution of replicated data and compares it to the sample data. Any deviation between the model-generated data and actual data suggests possible model misspecification (Gelman et al., 2003; Kaplan, 2014).

An equally important feature of SEM practice is model comparison and selection. The goal of model selection and comparison in the Bayesian domain differs somewhat from the frequentist domain. Specifically, the goal of model comparison and selection in the frequentist domain focuses primarily on the model that best fits the data. In the Bayesian domain, however, the goal of model comparison and selection is to find a model that best predicts the data—and in the Bayesian domain the focus is on posterior prediction. A very simple and intuitive approach to model selection uses so-called *Bayes factors* (Kass & Raftery, 1995; Raftery, 1995). In essence, the Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another and is defined to be the ratio of two integrated likelihoods. Assuming a priori that there is no reason to favor one model over another (equal prior odds), then the posterior odds that the data favor one model over another is equivalent to the Bayes factor. One key benefit of the Bayes factor is that the analyst can incorporate informative prior odds into the Bayes factor if there is reason to believe that one model is more likely to be true than another.[2] Another key benefit of Bayes factors is that models do not have to be nested.

An interesting feature of the Bayes factor is that under conditions where there is little prior information, Raftery (1995) showed that an approximation of the Bayes factor yields the *Bayesian information criterion* (BIC). However, although the BIC is derived from a fundamentally Bayesian perspective, it is often productively used for model comparison in the frequentist domain. An explicitly Bayesian approach to model comparison and selection based on the concept of Bayesian deviance was developed by Spiegelhalter, Best, Carlin, and van der Linde (2002) and referred to as the *deviance information criterion* (DIC). The BIC and DIC are used the same way—namely, for a set of competing models, the model with the lowest value of the BIC or DIC is the model to be chosen from the predictive point of view mentioned earlier.

BAYESIAN MODEL AVERAGING

To begin, consider a quantity of interest such as a future observation. Following the notation given in Madigan and Raftery (1994), we denote this quantity as Υ. Next, consider a set of competing models $M_k, k = 1, 2, \ldots, K$ that are not necessarily nested. The posterior distribution of Υ given data D can be written as

$$p(\Upsilon|D) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|D), \tag{7}$$

where $p(M_k|D)$ is the posterior probability of model M_k written as

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{l=1}^{K} p(D|M_l)p(M_l)}. \tag{8}$$

The interesting feature of Equation 8 is that $p(M_k|D)$ can be different for different models. The term $p(D|M_k)$ can be expressed as an integrated likelihood

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{9}$$

where $p(\theta_k|M_k)$ is the prior density of θ_k under model M_k (Raftery et al., 1997). Thus, BMA provides an approach for combining models specified by researchers. The advantage of BMA was discussed by Madigan and Raftery (1994), who showed that BMA provides better predictive

[2] This is rarely seen in practice, and of course, software packages that produce the Bayes factor will use equal prior odds as a default.

performance than that of any single model. We show that a Bayesian model averaged structural model provides better prediction of the endogenous variable of interest than any single model, including the initially specified model.

Occam's Window

As pointed out by Hoeting et al. (1999), BMA is difficult to implement. In particular, they noted that the number of terms in Equation 7 can be quite large, the corresponding integrals are hard to compute (although possibly less so with the advent of MCMC), the specification of $p(M_k)$ might not be straightforward, and choosing the class of models to average over is also challenging. The problem of reducing the overall number of models that one could incorporate in the summation of Equation 7 has led to a solution based on the notion of *Occam's window* (Madigan & Raftery, 1994).

To motivate the idea behind Occam's window, consider the problem of finding the best subset of predictors in a linear regression model. Following closely the discussion given in Raftery et al. (1997), we consider an initially large number of predictors, but perhaps the goal is to find a subset that provides accurate predictions.[3] As noted in the earlier quote by Hoeting et al. (1999), the concern in drawing inferences from a single "best" model is that the choice of a single set of predictors ignores uncertainty in model selection. Occam's window (Madigan & Raftery, 1994) provides an approach for BMA by reducing the subset of models under consideration.

The Occam's window algorithm proceeds in two steps (Raftery et al., 1997). In the first step, models are eliminated if they predict the data less well than the model that provides the best predictions. Formally, consider a set of models $k = 1 \ldots K,$ and a cutoff value C chosen in advance by the analyst. Then, the set A'

$$A' = \left\{ M_k : \frac{max_l\{p(M_l|D)\}}{p(M_k|D)} \leq C \right\}. \tag{10}$$

We see that Equation 10 compares the model with the largest posterior model probability, $max_l\{p(M_l|D)\},$ to a given model $p(M_k|D)$. If the ratio in Equation 10 is less than or equal to a chosen value *C,* then it is discarded from the set of models to be included in the model averaging.

In the second step, models are discarded from consideration if they receive less support from the data than simpler submodels. Formally, we consider a set B, where

$$B = \left\{ M_k : \exists M_l \in A', M_l \subset M_k, \frac{p(M_l|D)}{p(M_k|D)} > 1 \right\}. \tag{11}$$

Equation 11 states that there exists a model M_l within the set A' and where M_l is simpler than M_k. If the simpler model receives more support from the data than the more complex model, then it is included in the set *B.* Notice that the second step corresponds to the principle of Occam's razor (Madigan & Raftery, 1994).

With Step 1 and Step 2, the problem of BMA is simplified by replacing Equation 7 with

$$p(\Upsilon|D,A) = \sum_{M_k \in A} p(\Upsilon|M_k,D)p(M_k|D,A), \tag{12}$$

where A is the relative complement of A' and *B.* That is, the models under consideration for BMA are those that are in A' but not in *B.*

Madigan and Raftery (1994) then outlined the approach to choosing between two models to be considered for BMA. Specifically, now consider just two models M_1 and M_0, where M_0 is the smaller of the two models. This could be the case where M_0 contains fewer predictors than M_1 in a regression analysis. In terms of log-posterior odds, if the log-posterior odds are positive, indicating support for M_0, then we reject M_1. If the log-posterior odds are large and negative, then we reject M_0 in favor of M_l. Finally, if the log-posterior odds lie in between the preset criterion, then both models are retained.

THE BMA-SEM ALGORITHM

Our approach to BMA for SEM makes use of the relationship between path diagrams commonly encountered in SEM practice and so-called *directed acyclic graphs* (DAGs), the latter having been developed by Pearl (2009). BMA over DAGs was discussed in Madigan and Raftery (1994). In this section, we describe the full algorithm used to conduct BMA for structural equation models.

The general steps of our algorithm are as follows:

1. Specify an initial model of interest recognizing that this might not be the model that generated the data.
2. Starting with the initial model represented as a DAG, implement a search over the DAG to reduce the the space of models to a reasonable size.
3. Obtain the posterior model probabilities for each model.
4. Obtain the weighted average of structural parameters over each model, weighted by the posterior model probabilities.

[3] The notion of best subset regression is controversial in the frequentist framework because of concern over capitalization on chance. However, in the Bayesian framework with its focus on predictive accuracy, finding the best subset of predictors is less of a problem.

5. Compare predictive performance of the BMA-SEM to the initially specified Bayesian SEM by computing the reduced form of the models and calculating the log score or the predictive coverage. Our approach has been programmed in R (R Core Team, 2014) and is available at http://bise.wceruw.org/publications.html.

Model Selection Via the Up and Down Algorithm

We apply the search algorithm suggested by Madigan and Raftery (1994), which we refer to as the up and down algorithm. For a set of models under consideration we first execute the down algorithm, where each model in the set is compared with its submodels. If there is a model with no submodel in the down algorithm, then the model comes under consideration for the up algorithm. Thus, the up algorithm is carried out only when a set of models under consideration for the up algorithm exist after the down algorithm is completed.

The notation for the up and down algorithm is as follows, where initial values are denoted in parentheses:

- A is the set of acceptable models $(A = \phi)$.
- C_D is the set of initial models considered for the down algorithm.
- C_U is the set of models considered for the up algorithm $(C_U = \phi)$.
- M is a model from C_D in the down algorithm or a model from C_U in the up algorithm.
- M_{sub} is a submodel of M in the down algorithm.
- M_{sup} is a supermodel of M in the up algorithm.
- BIC_M is the BIC for model M.
- BIC_{Msub} is the BIC for model M_{sub}.
- BIC_{Msup} is the BIC for model M_{sup}.
- C is the cutoff value in the Occam's window.

We outline the down algorithm and then the up algorithm in the following sections.

Down algorithm

1. Randomly select a model M from C_D.
2. Remove M from C_D and add M into A.
3. If M has no submodel, then remove M from A, add M into C_U, and go to step 8.
4. If M has submodels in C_D, then select a submodel M_{sub} by randomly removing a link from M.
5. Fit M and M_{sub} to the data and compute BIC_M and BIC_{Msub}, respectively.
6. Model comparison
   a. Calculate the difference between the BIC values. $\Delta BIC = BIC_M - BIC_{M_{sub}}$.
   b. If $\Delta BIC > \log(C)$, then add M_{sub} into C_D and remove M from A.
   c. If $\Delta BIC < -\log(C)$, then remove both M_{sub} and all its submodels from C_D.
   d. If $-\log(C) \leq \Delta BIC \leq \log(C)$, then add M_{sub} into C_D.
7. Repeat steps 4 to 6 until there are no remaining submodels of M in C_D to be compared.
8. Go to step 1 until there is no remaining model in C_D.
9. If $C_U \neq \phi$, then execute the up algorithm.

Up algorithm

10. Randomly select a model M from C_U.
11. Remove M from C_U and add M into A.
12. If M has no supermodel, go to step 17.
13. If M has supermodels in C_U, then select a supermodel M_{sup} by randomly adding a link to model M.
14. Fit M and M_{sup} to data and compute BIC_M and $BIC_{M_{sup}}$, respectively.
15. Model comparison
    a. Calculate the difference between the two BIC values, $\Delta BIC = BIC_{M_{\text{sup}}} - BIC_M$
    b. If $\Delta BIC > \log(C)$, then add M into A.
    c. If $\Delta BIC < -\log(C)$, then add M_{sup} into C_U and remove M from A.
    d. If $-\log(C) \leq \Delta BIC \leq \log(C)$, then add M_{sup} into C_U.
16. Repeat steps 13 to 15 until there is no remaining supermodel of M in C_U to be compared.
17. Go to step 10 until there is no remaining model in C_U.

Model Averaging

A set of J possible structural equation models (l = 1,2,…, J) in A are chosen through the up and down algorithm. With this set, the posterior model probabilities (PMPs) are obtained using a BIC approximation as

$$p(M_j|D) = \frac{exp(-.5 \times \Delta BIC_j)}{\sum_{l=1}^{J} exp(-.5 \times \Delta BIC_l)}, \tag{13}$$

where ΔBIC is the difference between the BIC of each model and the maximum of BICs of all the models in the set. The posterior model probabilities are used as weights to obtain posterior means of parameters across all the models in the set. In other words, posterior means of the model parameters are the averaged parameters of the posterior distributions for the set of selected models, weighted by their posterior model probabilities. The posterior mean for, say, parameter θ under model M_j given D, can be written as

$$E(\theta|D, M_j) = \sum_{M_j \in A} \hat{\theta} p(M_j|D). \tag{14}$$

Predictive Performance

This article compares the predictive performance of a BMA-SEM to the predictive performance based on the initially specified Bayesian structural equation model. For this comparison, it is convenient to transform the structural form of the model to its reduced form where the endogenous variables are on the left side of the equation and the exogenous variables are on the right side. The structural form can be rewritten as

$$(\mathbf{I}-\mathbf{B})\mathbf{y}=\boldsymbol{\alpha}+\boldsymbol{\Gamma}\mathbf{x}+\boldsymbol{\zeta}. \tag{15}$$

If $(\mathbf{I}-\mathbf{B})$is nonsingular, then the equation can be written as

$$\mathbf{y}=(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\alpha}+(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{x}+(\mathbf{I}-\mathbf{B})^{-1}\zeta \tag{16}$$

$$=\boldsymbol{\Pi}_0+\boldsymbol{\Pi}_1\mathrm{x}+\zeta^{*}, \tag{17}$$

where $\boldsymbol{\Pi}_0$ is the vector of reduced form intercepts, $\boldsymbol{\Pi}_\mathbf{1}$ is the vector of reduced form slopes, and $\boldsymbol{\zeta}$* is the vector of reduced form disturbances with variance matrix $\boldsymbol{\Psi}$*.

Using the reduced form, we obtain 90% of predictive coverage, which is the percentage of the model for the observation in a new data set or a test set that fall in the 90% prediction interval (Hoeting et al., 1999). The comparison procedure is as follows:

1. Randomly divide the data set into a model-averaging set and a predictive testing set.
2. Fit a single Bayesian structural equation model and BMA-SEM to the model-averaging data.
3. Convert the structural form of the model to its reduced form.
4. Predict the final dependent variable in the reduced form for the predictive testing data with the result of the reduced form of the Bayesian structural equation model and BMA-SEM.
5. Compare their predictive performance based on 90% of predictive coverage.

SIMULATION STUDY: METHODS AND RESULTS

Our simulation study examines two models and two sample size conditions. The two models are specified as shown in Figures 1 and 2 and are the same except that Model 1 has six weaker regression coefficients compared to Model 2. The population covariance matrices for these data sets were generated using M*plus* 7.1 (Muthén & Muthén, 1998–2010). Data under all conditions of the design were replicated 100 times, providing a frequentist evaluation of the estimation methods used in this study.

For each model, two different sample sizes were examined: 200 and 5,000. Each data set was evenly split into the model averaging data set and the predictive testing data set. The model averaging data set $(N=100\,\mathrm{or}\,N=2,500)$ was use to estimate the each model under three different methods: BMA-SEM, Bayesian structural equation model, and conventional frequentist SEM (FSEM) under maximum likelihood estimation, which was fit using the R package "lavaan" (Rosseel, 2012) . The predictive testing data $(N=100\,\mathrm{or}\,N=2,500)$ were used to compare the three methods with respect to predictive Performance.

A Bayesian structural equation model was fit to the model averaging data sets. We used noninformative conjugate priors for all model parameters. To obtain the posterior distributions of the model parameters, we used the "rjags" package (Plummer, 2014). The "coda" package (Plummer, Best, Cowles, & Vines, 2006) was also used for postanalysis processing of the MCMC diagnostics and posterior summaries. For this article, the algorithm was set to produce 5,000

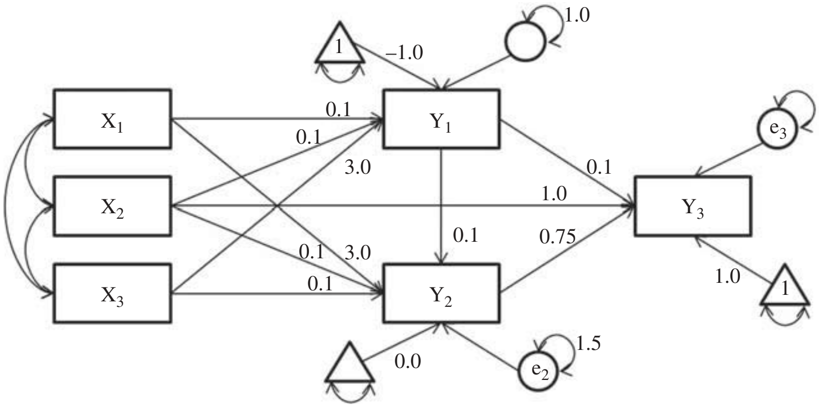

FIGURE 1 Model 1 for the simulation study.

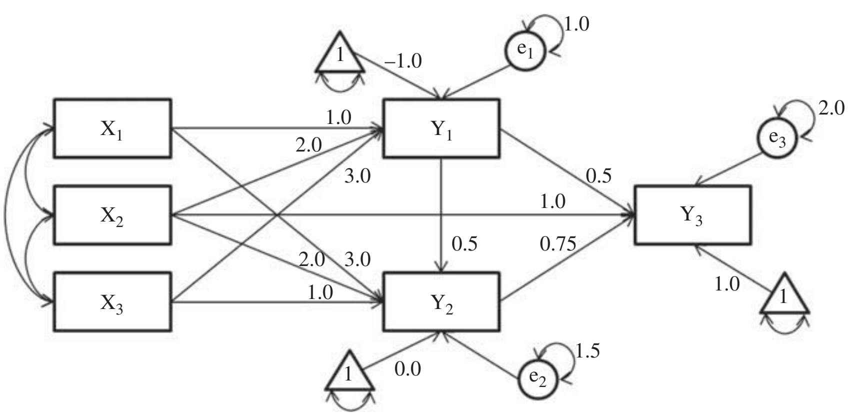

FIGURE 2 Model 2 for the simulation study.

burn-in iterations with 245,000 post-burn-in draws and a thinning interval of 50 from two chains starting at different locations of the posterior distribution.[4]

Our simulation study also implemented BMA-SEM under three different cutoff values (C) of Occam's window–4, 20, and 100— to examine how different cutoff values affect the model averaging and prediction results.

Predictive Performance Comparison

The predictive performance of BMA-SEM with three different cutoff values was compared to the performance of FSEM and the Bayesian structural equation model. Prediction was also examined under true SEM where the true parameters were used for the reduced form in the prediction. The averaged 90% coverage of prediction over 100 replications for the predictive testing data set under the design conditions is illustrated in Table 1. The simulation results show that BMA-SEM performs as well as FSEM when the true model is known, although BMA-SEM shows slightly lower coverage than the Bayesian structural equation model and true SEM with a small sample size. Different cutoff values of the Occam's window for BMA-SEM did not affect prediction under the conditions of this simulation.

CASE STUDY: METHODS AND RESULTS

Having demonstrated that our BMA-SEM approach provides the same predictions we would obtain if the true model were known, we move to our case study, where theory predicts that BMA should provide better predictions than any submodel based on predictive coverage and the log-score rule (e.g., Madigan & Raftery, 1994). For our case study, we used data from PISA 2009 (OECD, 2010) to estimate a model relating reading proficiency to a set of background and reading strategy variables. The sample was collected from PISA-eligible students in the United States, and the sample size was 5,053. The sample was split into a model averaging set (n = 2,526) and a predictive testing set (n = 2,527). The background exogenous variables in the initial structural equation model are Gender $(\text{male} = 0,\ \text{female} = 1)$; immigrant status (Immigr); and a measure of the economic, social, and cultural status of the student (ESCS). Additionally, three measures of student reading strategies were mediating endogenous variables including memorization strategies (MEMO), elaboration strategies (ELAB), and control strategies (CSTRAT). The first plausible value of the PISA 2009 reading assessment (Reading) was used as the final outcome variable. The path diagram is depicted in Figure 3. The Bayesian structural equation model for the case study used 495,000 post-burn-in draws.[5]

Based on the initial model in Figure 3, our BMA-SEM algorithm selected one, one, and three models out of 2^{18} (262,144) total possible models for C = 4, C = 20, and C = 100, respectively. Table 2 presents the chosen three models from the BMA-SEM for C = 100. The first and the best model (M_1) in the BMA-SEM with C = 100 was also the model selected in the BMA-SEM with C = 4 and C = 20. Regressions in the model are marked with a dot if they were included in the model. The three models accounted for 100% of the total posterior model probability. There were seven regressions set to zero in the initial model, including Reading on ESCS, Gender, and Immigr; ELAB on MEMO; CSTRAT on MEMO and ELAB; and MEMO on Immigr. With the exception of Reading on Immigr, the remaining effects appear in all three models, indicating that there is strong evidence for these regressions (D. Wang, Zhang, &

[4] Our use of two chains and this large number of draws was to ensure wide sampling of the posterior distribution to obtain valid comparisons across conditions.

[5] As with the simulation study, use of such a large number of post-burn-in draws was to ensure coverage and convergence of the MCMC sampling.

TABLE 1
Simulation Study Results: 90% Coverage and Log-Score Averaged Over 100 Replications

		$N = 100$		$N = 2,500$	
Model	*Method (C)*	*M (SE)*	*Log-Score*	*M (SE)*	*Log-Score*
Model 1	BMA-SEM (4)	0.91 0.03	–0.09	0.92 0.01	–0.08
	BMA-SEM (20)	0.91 0.03	–0.09	0.92 0.01	–0.08
	BMA-SEM (100)	0.91 0.03	–0.09	0.92 0.01	–0.08
	FSEM	0.91 0.03	–0.09	0.92 0.01	–0.08
	BSEM	0.92 0.03	–0.08	0.92 0.01	–0.08
	True SEM	0.93 0.03	–0.07	0.92 0.01	–0.08
Model 2	BMA-SEM (4)	0.91 0.03	–0.09	0.92 0.01	–0.08
	BMA-SEM (20)	0.91 0.03	–0.09	0.92 0.01	–0.08
	BMA-SEM (100)	0.91 0.03	–0.09	0.92 0.01	–0.08
	FSEM	0.91 0.03	–0.09	0.92 0.01	–0.08
	BSEM	0.92 0.03	–0.08	0.92 0.01	–0.08
	True SEM	0.92 0.03	–0.08	0.92 0.01	–0.08

Note. *C* refers to the cutoff values of the Occam's window; BMA=Bayesian model averaging; BSEM = Bayesian structural equation model; FSEM = frequentist SEM; true SEM = SEM using true parameter values.

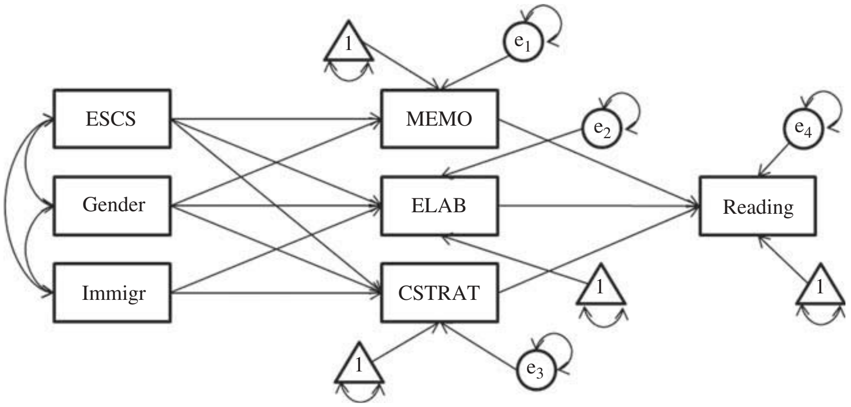

FIGURE 3 SEM model for the case study based on PISA data.

Bakhai, 2004). Nevertheless, the best model accounted for only 72% of the total posterior model probability (PMP) indicating a fair amount of uncertainty remaining in the model selection.

Table 3 presents the results from the BMA-SEM and the Bayesian structural equation model. We observe large differences between two approaches. As we described earlier, the six regressions that were assumed to be zero in the initial model showed strong evidence for their effect on Reading. On the other hand, the regressions of ELAB on Immigr and CSTRAT on Immigr that existed in the initial model indicated weak evidence for their effects in the BMA-SEM (D. Wang et al., 2004).

Predictive Performance Comparison

For our case study using the PISA 2009 data, the predictive performance of BMA-SEM with three different *C* values was compared to the predictive performance of FSEM and the Bayesian structural equation model. The results are presented in Table 4. We find, as theoretically expected, that regardless of the *C* value, BMA-SEM yields better predictive performance based on predictive coverage and the log-score rule. It is important to note that the choice of *C* values could influence predictive performance in other substantive examples.

CONCLUSION

For decades, the focus of attention in SEM has been on goodness-of-fit. This focus has led to a proliferation of alternative fit indexes that are designed to mitigate the known sensitivities of the likelihood ratio test (see, e.g., Kaplan, 2009). This focus on goodness of fit is understandable but has detracted from developing models that could be used beyond the immediate investigation. The question of using a model for some purpose beyond the immediate

TABLE 2
Selected Models by BMA-SEM With the *C* = 100 for the Program for International Student Assessment Data

Parameter	M$_1$	M$_2$	M$_3$
MEMO~ESCS	•	•	
ELAB~ESCS	•	•	•
CSTRAT~ESCS	•	•	•
Reading~ESCS	•	•	•
MEMO~Gender	•	•	•
ELAB~Gender	•	•	•
CSTRAT~Gender	•	•	•
Reading~Gender	•	•	•
MEMO~Immigr	•	•	•
ELAB~Immigr			
CSTRAT~Immigr		•	
Reading~Immigr			
ELAB~MEMO	•	•	•
CSTRAT~MEMO	•	•	•
Reading~MEMO	•	•	•
CSTRAT~ELAB	•	•	•
Reading~ELAB	•	•	•
Reading~CSTRAT	•	•	•
BIC	39461.68	39464.74	39465.15
PMP	0.72	0.15	0.13

Note. ~ refers to regression of left-hand variable onto right-hand variable; BIC = Bayesian information criterion; PMP = posterior model probability.

TABLE 4
90% Coverage and Log-Score for Program for International Student Assessment Example

Method (C)	*90% Coverage*	*Log-Score*
BMA-SEM (4)	0.90	–0.11
BMA-SEM (20)	0.90	–0.11
BMA-SEM (100)	0.90	–0.11
FSEM	0.88	–0.13
B SEM	0.88	–0.13

Note. C refers to the cutoff values of the Occam's window; BMA=Bayesian model averaging; FSEM = frequentist structural equation modeling; BSEM = Bayesian structural equation modeling.

investigation leads us to consider the accuracy of a model's predictions. The issue of predictive accuracy is a central feature of Bayesian statistics—arguably more central than goodness of fit. Indeed, the BIC and DIC are oriented toward choosing models based on considering predictive accuracy. If the goal of model building is one of predictive accuracy, then we should be less concerned about the fit of one's idiosyncratic model and more concerned about finding a model that will predict well.

In the Bayesian domain, BMA is known to yield models that perform better than any given submodel on the criteria of predictive accuracy. This is due to the fact that not all models are equally good as measured by their posterior model probabilities—yet all models contain some important information. By combining models and at the same time accounting for model uncertainty, we obtain a stronger model in terms of predictive accuracy.

It should be noted that model averaging has been studied within the frequentist tradition. A full comparison of

TABLE 3
Comparison of BMA-SEM Versus Bayesian Structural Equation Model (BSEM) Results for the Program for International Student Assessment Data

	BMA-SEM			*BSEM*			
Parameter	*M(β|D)*	*SD (β|D)*	*p (β|d)*	*EAP*	*SD*	*95%*	*PPI*
MEMO~ESCS	0.07	0.04	0.87	0.06	0.02	0.01	0.10
ELAB~ESCS	0.10	0.02	1.00	0.14	0.03	0.09	0.19
CSTRAT~ESCS	0.18	0.02	1.00	0.28	0.02	0.23	0.32
Reading~ESCS	0.36	0.02	1.00	—			
MEMO~Gender	0.28	0.04	1.00	0.27	0.04	0.19	0.36
ELAB~Gender	–0.15	0.04	1.00	–0.02	0.04	–0.11	0.07
CSTRAT~Gender	0.18	0.03	1.00	0.30	0.04	0.21	0.38
Reading~Gender	0.24	0.03	1.00	—			
MEMO~Immigr	0.20	0.06	1.00	—			
ELAB~Immigr	0.00	0.00	0.00	0.14	0.06	0.03	0.26
CSTRAT~Immigr	0.01	0.04	0.15	0.23	0.06	0.12	0.34
Reading~Immigr	0.00	0.00	0.00	—			
ELAB~MEMO	0.46	0.02	1.00	—			
CSTRAT~MEMO	0.44	0.02	1.00	—			
Reading~MEMO	–0.22	0.02	1.00	–0.25	0.02	–0.29	–0.21
CSTRAT~ELAB	0.37	0.02	1.00	—			
Reading~ELAB	–0.13	0.02	1.00	–0.15	0.02	–0.19	–0.11
Reading~CSTRAT	0.34	0.02	1.00	0.44	0.02	0.40	0.48

Note. N = 2,526; BMA = Bayesian model averaging; EAP = expected a posteriori; *SD* = posterior standard deviation; PPI = posterior probability interval.

frequentist model averaging was beyond the scope of this article (see, e.g., H. Wang, Zhang, & Zou, 2009, for a review; see also Claeskens & Hjort, 2008). However, we believe that future research should formally compare frequentist-based model averaging to BMA in terms of their relative advantages with respect to predictive performance.

This article considered BMA in the context of structural models among observed variables (sometimes referred to as *path analysis*). It is important to examine BMA for structural models that include latent variables as well.

Incorporating latent variables into a BMA for structural equation models is not trivial insofar as a decision must be made regarding how latent variable loadings should be treated in the BMA process. That is, assuming that the measurement part of the model is specified a priori, consideration then needs to be given to whether factor loadings should be removed or added as cross-loadings in the up and down algorithm. Such changes would, in principle, change the meaning of the latent variables and lead to a final measurement model that for predictive purposes might bear little resemblance to the initial measurement model. Clearly, this issue deserves further investigation.

BMA is not without theoretical difficulties. Perhaps the most important issues lies in the consideration of priors assigned to models in the averaging process. As pointed out by Claeskens and Hjort (2008, p. 218), uniform priors on model parameters can be improper and the usual normalizing constants would need to be adjusted across models. Moreover, with regard to nested models, the uniform prior might not be the best way to represent "non-informativeness." Indeed, Jeffreys (1961) suggested that for nested models, one use $p_j = 1/(j+1)$ for $j = 0, 1, \ldots$ as the prior probability, thus giving low prior probabilities to high-dimensional models. These issues deserve careful future study in the context of BMA for SEM.

Still another issue that deserves future research is the comparison of the up and down algorithm that we use to implement Occam's window and other specification search algorithms with SEM, such as Tetrad (Spirtes, Glymour, & Scheines, 2000), ant-colony optimization (Marcoulides & Drezner, 2003), and Tabu (Marcoulides, Drezner, & Schumacker, 1998), to name a few.

To conclude, we show that BMA can be successfully applied to structural equation models as a means of obtaining models with good predictive properties. Our simulation study is, in essence, a frequentist evaluation of our BMA-SEM approach and shows that BMA-SEM performs as expected when the true model is known. The case study demonstrates that when there exists model uncertainty, as would certainly be the case in a real research setting, our BMA-SEM approach obtains a model that yields better predictions than any given submodel. As always, the full benefit of our BMA-SEM approach will rest on its application to practical problems where prediction is of high priority.

ACKNOWLEDGMENTS

We would like to thank Fabrizzio Sanchez for early contributions to the R code used in this project.

FUNDING

The research reported in this article was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110001 to the University of Wisconsin–Madison. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education.

REFERENCES

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, UK: Cambridge University Press.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In J. M. Bemardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 6* (pp. 157–185. Oxford, UK: Oxford University Press.

Clyde, M. A., & George, E. I. (2004). Model uncertainty. *Statistical Science*, *19*, 81–94.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, *57*, 55–98.

Fernández, C., Ley, E., & Steele, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, *16*, 563–576.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). London, UK: Chapman & Hall.

Gilks, W. R, Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London, UK: Chapman & Hall.

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* (3rd ed.). New York, NY: Oxford University Press.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Newbury Park, CA: Sage.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York, NY: Guilford.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Learner, E. E. (1978). *Specification searches: Ad hoc inference with non-experimental data*. New York, NY: Wiley.

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

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorthms. *Structural Equation Modeling*, *10*, 154–164.

Marcoulides, G. A., Drezner, Z., & Schumacker R.E. (1998). Model specification searches in structural equation modeling using Tabu search. *Structural Equation Modeling*, *5*, 365–376.

Muthén, L. K., & Muthén, B. (1998–2010). *Mplus*: *Statistical analysis with latent variables*. Los Angeles, CA: Muthén & Muthén.

Organization for Economic Cooperation and Development. (2010). *PISA 2009 Results* (Vol. I-VI). Paris, France: OECD.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

Plummer, M. (2014). *rjags: Bayesian graphical models using mcmc* [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=rjags (R package version 3–13)

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from http://CRAN.R-project.org/doc/Rnews/

R Core Team. (2014). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Statistical Computing Group. Retrieved from http://www.R-project.org/

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. V. Marsden (Ed.), *Sociological methodology* (Vol. 25, pp. 111–196). New York, NY: Blackwell.

Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. (2014). *BMA: Bayesian model averaging* [Computer software manual]. Retrieved from http://CRAN. R-project. org/package=BMA (R package version 3.18.1)

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.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Sloughter, J. M., Gneiting, T., & Raftery, A. E. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review*, *141*, 2107–2119.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*, 583–639.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: The MIT Press.

Wang, D., Zhang, W., & Bakhai, A. (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine*, *22*, 3451–3467.

Wang, H., Zhang, X., & Zou, G. (2009). Frequentist model averaging estimation: A review. *Journal of System Science & Complexity*, *22*, 732–748.

Yeung, K. Y., Bumbarner, R. E., & Raftery, A. E. (2005). Bayesian model averaging: Development of an improved multiclass, gene selection, and classification tool for microarray data. *Bioinformatics*, *21*, 2394–2402.