Article

Evaluation Review
1-35

DOI: 10.1177/0193841X18761421
journals.sagepub.com/home/erx

Optimizing Prediction Using Bayesian Model Averaging: Examples Using Large-Scale Educational Assessments

David Kaplan[1] **and Chansoon Lee**[1]

Abstract

This article provides a review of Bayesian model averaging as a means of optimizing the predictive performance of common statistical models applied to large-scale educational assessments. The Bayesian framework recognizes that in addition to parameter uncertainty, there is uncertainty in the choice of models themselves. A Bayesian approach to addressing the problem of model uncertainty is the method of Bayesian model averaging. Bayesian model averaging searches the space of possible models for a set of submodels that satisfy certain scientific principles and then averages the coefficients across these submodels weighted by each model's posterior model probability (PMP). Using the weighted coefficients for prediction has been shown to yield optimal predictive performance according to certain scoring rules. We demonstrate the utility of Bayesian model averaging for prediction in education research with three examples: Bayesian regression analysis, Bayesian logistic regression, and a recently developed approach for Bayesian structural equation modeling. In each case, the model-averaged

[1] University of Wisconsin, Madison, WI, USA

Corresponding Author:
David Kaplan, University of Wisconsin, 1025 West Johnson Street, Madison, WI 53706, USA.
Email: dkaplan@education.wisc.edu

estimates are shown to yield better prediction of the outcome of interest than any submodel based on predictive coverage and the log-score rule. Implications for the design of large-scale assessments when the goal is optimal prediction in a policy context are discussed.

Keywords
Bayesian model averaging, large-scale assessments, education

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 the Bayesian framework concerns background knowledge about model parameters. In the Bayesian framework, all unknowns are described by probability distributions designed to encode background knowledge about parameters; and because parameters are, by definition, unknown, Bayesian inference encodes background knowledge about parameters in the form of prior distributions.

As with frequentist model building, another goal of Bayesian statistical analysis is model choice. Two popular methods are the *Bayesian information criterion* (BIC; Kass & Raftery, 1995; Schwarz, 1978) and the *deviance information criterion* (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). In both cases, a set of models are compared, and the model with the lowest BIC or DIC value is chosen for summary and discussion.

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 among a set of competing models that could also have generated the data. 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)

In practice, model uncertainty often goes unnoticed, and the impact of this uncertainty can be quite profound. Although a number of methods exist in the Bayesian literature to aid in improving model prediction, including sensitivity analyses via posterior predictive checking (Gelman, Meng, &

Stern, 1996) and more recently the Bayesian "lasso" (Park & Casella, 2008), in the end, a single model is chosen for prediction purposes. As in the quote by Hoeting et al. (1999), we do not wish to settle on a single model but rather draw predictive strength through combining models. The current approach to addressing the problem of model uncertainty through combing models from a Bayesian point of view lies in the method of *Bayesian model averaging*.

Bayesian model averaging has had a long history of theoretical developments and practical applications. Early work by Leamer (1978) laid the foundation for Bayesian model averaging. Fundamental theoretical work on Bayesian model averaging was conducted in the mid-90s by Madigan and his colleagues (e.g., Hoeting, Madigan, Raftery, & Volinsky, 1999; Madigan & Raftery, 1994; Raftery, Madigan, & Hoeting, 1997). Additional theoretical work was conducted by Clyde (1999, 2003). Draper (1995) discussed how model uncertainty can arise even in the context of experimental designs, and Kass and Raftery (1995) provided a review of Bayesian model averaging and the costs of ignoring model uncertainty. A review of the general problem of model uncertainty can be found in Clyde and George (2004). Bayesian model averaging has been implemented in the R software programs "BMA" (Zeugner & Feldkircher, 2015) and "BAS" (Clyde, 2017). These packages are quite general, allowing Bayesian model averaging over linear models, generalized linear models, and survival models, with flexible handling of parameter priors.

Practical applications of Bayesian model averaging can be found across a wide variety of domains. A perusal of the extant literature shows applications of Bayesian model averaging to economics (e.g., Fernández, Ley, & Steele, 2001), political science (e.g., Montgomery & Nyhan, 2010), bioinformatics of gene express (e.g., Yeung, Bumbarner, & Raftery, 2005), weather forecasting (e.g., Sloughter, Gneiting, & Raftery, 2013), causal inference using propensity score analysis (Chen & Kaplan, 2015; Kaplan & Chen, 2012, 2014), and structural equation modeling (SEM; Kaplan & Lee, 2015) to name just a few.

Policy Significance

The subject-matter motivation for this article lies in the use of large-scale assessments for education policy analysis. Specifically, of critical importance to educational evaluation and policy analysis is the monitoring of trends in important educational outcomes. For example, the United Nations Sustainable Development Goals identified Goal 4 as focusing on quality

education for all. Many of the stated targets under Goal 4 focus on reducing the gender gap in quality education, and, in particular, Goal 4.6 focuses on achieving literacy and numeracy for men and women (http://www.un.org/sustainabledevelopment/).

Developing optimal predictive models would allow education researchers and policy makers to assess cross-country progress and forecasts toward Goal 4.6 using, among other data sources, large-scale cross-sectional and longitudinal educational data such as the Early Childhood Longitudinal Study (ECLS-K; National Center for Educational Statistics (NCES), 2001), the Program for International Student Assessment (PISA; e.g., Organization for Economic Cooperation and Development (OECD), 2010), or the Trends in International Mathematics and Science Study (TIMSS; e.g., Mullis, 2013). Large-scale assessments provide a unique lens on the antecedents, mediators, and outcomes of education policies and practices. Although applications of advanced statistical models to these types of data are, of course, not new, a review of the extant literature indicates that these models have not been applied to educational data with the explicit goal of obtaining models exhibiting optimal predictive performance for education research, evaluation, or policy purposes.

Purpose and Organization of this Article

The purpose of this article is to provide a general review of Bayesian model averaging with applications to common statistical models applied in educational research, evaluation, and policy analysis. Our focus is on traditional methods of Bayesian model averaging that make use of readily available software applied to real data in order to demonstrate the gain in predictive accuracy when applying Bayesian model averaging for optimizing predictive performance. It should be noted, however, that Bayesian model averaging is still an active field of methodological development.

The organization of this article is as follows. In the next section, we discuss the model choice problem. Next, we outline the method of Bayesian model averaging with an additional discussion of Occam's window and the MC^3 algorithm, following closely the work of Madigan and his colleagues (Hoeting et al., 1999; Madigan & Raftery, 1994; Raftery et al., 1997). We then demonstrate Bayesian model averaging with three examples: Bayesian linear regression, Bayesian logistic regression, and Bayesian structural equation modeling (BSEM). This article next discusses some additional technical considerations, a brief discussion of frequentist approaches to model averaging, and finally, implications for large-scale assessment

designs when the goal is to optimize prediction. The last section of this article concludes. All analyzes are conducted within the R programming environment (R Core Team, 2017), and all data and code are available at http://bise.wceruw.org/index.html

The Method of Bayesian Model Averaging

Following Madigan and Raftery (1994), consider a quantity of interest such as a future observation. We will 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 y can be written as a mixture distribution,

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

$p(M_k|y)$ is the posterior probability of model M_k written as:

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \tag{2}$$

the first term in the numerator on the right-hand side of Equation 2 is the probability of the data given model k, also referred to as the *integrated likelihood* written as:

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{3}$$

$p(\theta_k|M_k)$ is the prior distribution of the parameters θ_k under model M_k (Raftery et al., 1997). The PMPs can be considered mixing weights for the mixture distribution given in Equation 1 (Clyde & Iversen, 2015). The second term $p(M_k)$ on the right-hand side of Equation 2 is the prior model probability for model k, allowing each model to have a different prior probability based on past performance of that model or a belief regarding which of the models might be the true model. The denominator of Equation 2 ensures that $p(M_k|y)$ integrates to 1.0, as long as the true model is in the set of models under consideration. A review of this latter issue is reserved for the Discussion section of this article.

Connections to Bayes Factors (BFs)

An important feature of Equation 2 is that $p(M_k|y)$ captures the posterior (postdata) uncertainty in a given model and will likely vary across models.

Herein lies the problem of model selection; given the choice of a particular model, the analyst effectively ignores the uncertainty in other models that could have generated the data. Of course, Equation 2 could be used as a method for model selection, simply choosing the model with the largest PMP. However, to settle on a particular model still ignores the uncertainty inherent in the choice problem.

Yet another common approach for model selection is the BF which provides a way to quantify the odds that the data favor one hypothesis over another (Kass & Raftery, 1995). A key benefit of BFs is that models do not have to be nested. To motivate the BFs, consider two competing models, denoted as M_k and M_l, that could be nested within a larger space of alternative models. Let θ_k and θ_l be the two parameter vectors associated with these two models. These could be two regression models with a different number of variables or two SEMs specifying very different directions of mediating effects. The goal is to develop a quantity that expresses the extent to which the data support M_k over M_l. One quantity could be the posterior odds of M_k over M_l, expressed as:

$$\frac{p(M_k|y)}{p(M_l|y)} = \frac{p(y|M_k)}{p(y|M_l)} \times \left[\frac{p(M_k)}{p(M_l)}\right]. \tag{4}$$

The first term on the right-hand side of Equation 4 is the ratio of two integrated likelihoods. This ratio is referred to as the BF for M_k over M_l, denoted here as BF_{kl}. In words, our prior opinion regarding the odds of M_k over M_l, given by $p(M_k)/p(M_l)$, is weighted by our consideration of the data, given by $p(y|M_k)/p(y|M_l)$.

A connection between the BF in Equation 4 and the PMP in Equation 2 has been pointed by others (see, e.g., Clyde, 1999). Specifically, when examining more than two models, and assuming equal prior odds, then the BF for M_k over M_l can be written as:

$$\mathrm{BF}_{kl} = \frac{p(y|M_k)}{p(y|M_l)}. \tag{5}$$

Assuming that we fix the first model M_1 as the baseline model, Equation 2 can be reexpressed as:

$$p(M_k|y) = \frac{\mathrm{BF}_{k1}p(M_k)}{\sum_{l=1}^{K}\mathrm{BF}_{l1}p(M_l)}. \tag{6}$$

Computational Issues

As pointed out by Hoeting et al. (1999), Bayesian model averaging is difficult to implement. In particular, they note that the number of terms in Equation 1 can be quite large, the corresponding integrals are hard to compute (though possibly less so with the advent of Markov chain Monte Carlo [MCMC] sampling), the specification of $p(M_k)$ may not be straightforward, and choosing the class of models to average over is also challenging.

To address the problem of computing Equation 3, the Laplace method, which has been used productively for the computation of BFs (Kass & Raftery, 1995), can be used, and this will lead to a simple BIC approximation under certain circumstances (Raftery, 1996; Tierney & Kadane, 1986).[1]

The problem of reducing the overall number of models that one could incorporate in the summation of Equation 1 has led to two interesting solutions. One solution is based on the so-called *Occam's window* criterion (Madigan & Raftery, 1994) and the other is based on *MCMC model composition* (MC^3)

Occam's window. To motivate the idea behind Occam's window, consider the problem of finding the best subset of predictors in a linear regression model.[2] Following closely the discussion given in Raftery, Madigan, and Hoeting (1997), we could initially start with very large number of predictors, but perhaps the goal is to narrow down this initially large set of predictors to a small number of predictors that provide accurate predictions. 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 provides an approach to Bayesian model averaging that reduces the subset of models under consideration.

The algorithm proceeds in two steps (Raftery et al., 1997). In the first step, models are eliminated from Equation 1 if they predict the data much less well than the model that provides the best predictions based on a "caliper" value C chosen in advance by the analyst. The caliper C sets the "width" of Occam's window. Formally, consider again a set of models M_k, $k = 1, ..., K$. Then, the set A' is defined as:

$$A' = \left\{ M_k : \frac{max_l\{p(M_l|y)\}}{p(M_k|y)} \leq C \right\}. \tag{7}$$

In words, Equation 7 compares the model with the largest PMP, $max_l\{p(M_l|y)\}$, to a given model, $p(M_k|y)$. If the ratio in Equation 7 is greater than the chosen value C, then it is discarded from the set A' of models to be included in the model averaging. Notice that the set of models contained in A' is based on BF values.

The set A' now contains models to be considered for model averaging. In the second, optional, step, models are discarded from A' if they receive less support from the data than simpler submodels. Formally, models are further excluded from Equation 1 if they belong to the set:

$$B = \left\{ M_k : \exists M_l \in A', M_l \subset M_k, \frac{p(M_l|y)}{p(M_k|y)} > 1 \right\}. \tag{8}$$

Again, in words, Equation 8 states that there exists a model M_l within the set A' and where M_l is simpler than M_k. If a complex model receives less support from the data than a simpler submodel—again based on the BF—then it is excluded from 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 Bayesian model averaging is simplified by replacing Equation 1 with:

$$p(\Upsilon|y, A) = \sum_{M_k \in A} p(\Upsilon|M_k, y) p(M_k|y, A). \tag{9}$$

In other words, models under consideration for Bayesian model averaging are those that are in A' but not in B. Formally, $A = A' \backslash B$.

Madigan and Raftery (1994) then outline an approach to the choice between two models to be considered for Bayesian model averaging. To make the approach clear, consider the case of just two models M_1 and M_0, where M_0 is the simpler 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 is large and negative, then we reject M_0 in favor of M_1. Finally, if the log-posterior odds lies in between the preset criterion, then both models are retained.

MC^3. The goal of MC^3 is the same as that of Occam's window—namely—to reduce the space of possible models that can be explored in a Bayesian model averaging exercise. Following Hoeting et al. (1999), the MC ^{3} algorithm proceeds as follows. First, let M represent the space of models of interest; in the case of linear regression, this would be the space of all possible combinations of variables. Next, the theory behind MCMC allows

us to construct a Markov chain $\{M(t), t = 1, 2, \ldots, \}$ which converges to the posterior distribution of model k, that is, $p(M_k|y)$.

The manner in which models are retained under MC ^{3} is as follows. First, for any given model currently explored by the Markov chain, we can define a neighborhood for that model which includes one more variable and one less variable than the current model. So, for example, if our model has four predictors x_1, x_2, x_3, and x_4, and the Markov chain is currently examining the model with x_2 and x_3, then the neighborhood of this model would include $\{x_2\}$, $\{x_3\}$, $\{x_2, x_3, x_4\}$, and $\{x_1, x_2, x_3\}$. Now, a transition matrix is formed such that moving from the current model M to a new model M' has probability 0 if M' is not in the neighborhood of M and has a constant probability if M' is in the neighborhood of M. The model M' is then accepted for model averaging with probability,

$$\min\left\{1, \frac{pr(M'|y)}{pr(M|y)}\right\}; \tag{10}$$

otherwise, the chain stays in model M.

Gauging Predictive Performance in Bayesian Model Averaging

A key characteristic of statistics is to develop accurate predictive models (Dawid, 1984). Indeed, as pointed out by Bernardo and Smith (2000), all other things being equal, a given model is to be preferred over other competing models if it provides better predictions of what actually occurred. Thus, a critical component in the development of accurate predictive models is to decide on rules for gauging predictive accuracy—often termed *scoring rules*.

Scoring rules provide a measure of the accuracy of probabilistic forecasts, and a forecast can be said to be "well-calibrated" if the assigned probabilities of the outcome match the actual proportion of times that the outcome occurred. The development of accurate predictive models has, arguably, been overlooked in education where the goal has been instead an orientation toward finding well-fitting models, particularly in the context of SEM (e.g., Kaplan, 2009).

A number of scoring rules are discussed in the literature (see, e.g., Gneiting & Raftery, 2007; Jose, Nau, & Winkler, 2008; Merkle & Steyvers, 2013a; Winkler, 1996); however, for this article, we will primarily evaluate predictive performance using the 90% predictive coverage criterion (Hoeting et al., 1999) and the log of the percentage predictive coverage for continuous outcomes referred to as the *log score*. Predictive coverage is

used productively in frequentist and Bayesian settings and is assessed using the proportion of predicted observations that fall in the corresponding 90% prediction interval. For this article, the predictive coverage criterion is implemented via the R routine "predict" in the program "stats" (R Core Team, 2017). For the prediction of a dichotomous outcome, it is common to use the Brier (1950) score defined as:

$$\text{Brier} = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2, \tag{11}$$

where over each forecast instant t, f_t is the probabilistic forecast and o_t is the observed event (1 *if the forecasted event took place*, 0 *otherwise*). Both the log score and the Brier score are so-called *proper scoring rules* insofar as the score is maximized (or minimized in the case of the Brier score) when the reported forecast probability is the same as true probability. In both cases, the log score is a local and strictly proper scoring rule that assesses the quality of the prediction by providing a numerical score based on the accuracy of the match between the predictive distribution and the actual obtained values. The log score is strictly proper in the sense that it is unique (see, e.g., Gneiting & Raftery, 2007; Merkle & Steyvers, 2013b, for more detail).

Case Study 1: Bayesian Model Averaging for the Linear Regression Model Using PISA 2009

Data and Model

The first case study make uses of the same data set and linear regression model as used by Kaplan (2014) except here we focus on relative predictive performance. The data set was collected from PISA 2009–eligible students in the United States (OECD, 2009). The sample size was 4,924 after list-wise deletion which was then was split into a model averaging set (n = 2,462) and a predictive testing set (n = 2,462). For this regression example, the outcome, reading proficiency (READING), was regressed on a set of background, attitudinal, and reading strategy variables. Background variables included FEMALE (*male* = 0, *female* = 1), immigrant status (NATIVE), language that the students use (SLANG: coded 1 *if the test language is the same as language at home*, 0 *otherwise*), and a measure of economic, social, and cultural status of the students (ESCS). Variables measuring reading attitudes were enjoyment of reading (JOYREAD) and diversity in reading (DIVREAD). Measures of student reading strategies

were memorization strategies (MEMOR), elaboration strategies (ELAB), and control strategies (CSTRAT). The first plausible value of the PISA 2009 reading assessment was used as the dependent variable for the regression model (see von Davier, 2013, for a discussion of plausible values).[3] This model serves as the initial model for Bayesian model averaging and can be defined as:

$$\begin{aligned}\widehat{\text{READING}} = {} & \beta_0 + \beta_1(\text{FEMALE}) + \beta_2(\text{NATIVE}) + \beta_3(\text{SLANG}) \\ & + \beta_4(\text{ESCS}) + \beta_5(\text{JOYREAD}) + \beta_6(\text{DIVREAD}) \\ & + \beta_7(\text{MEMOR}) + \beta_8(\text{ELAB}) + \beta_9(\text{CSTRAT}). \qquad (12)\end{aligned}$$

It should be noted that neither the model in Equation 12 nor the models in Case Studies 2 and 3 below contain interaction terms. This was done for simplicity and ease of communicating the central purpose of BMA. However, BMA can incorporate interaction terms, but it is important to proceed with caution. We mention the issue of using interaction terms within BMA in the Discussion section.

Method

For Bayesian model averaging, we used the "bicreg" function within the R package BMA (Raftery, Hoeting, Volinsky, Painter, & Yeung, 2015) setting Occam's window to 20. The BMA program utilizes the Laplace approximation described above and is computationally quite fast even for fairly large regression models. For model parameters, we used the default in the BMA package—namely, the *unit information prior*. Following Raftery (1998, pp. 3–6), the unit information prior is a weakly informative prior that is diffused over the region of the likelihood where parameter values are considered mostly plausible but not overly spread out. This is accomplished by forming the prior based on the maximum likelihood estimate of the parameter mean, with variance equal to the expected information matrix for one observation. The prior on the model space is $1/M$, where M is the number of models.

The Bayesian regression model was estimated using the Gibbs sampler as implemented in the "MCMCregress" function within the "MCMCpack" package (A. D. Martin, Quinn, & Park, 2013). We used improper uniform priors for the regression coefficients and precisions. In addition, the variance of the disturbance term was set to have a noninformative inverse-gamma distribution with shape and scale of 0.001. This analysis used 100,000 iterations, with 5,000 burn-in iterations and a thinning interval of 10.

Table 1. Selected Models by Bayesian Model Averaging: Regression Model Using PISA 2009.

Predictor	Model 1	Model 2	Model 3	Model 4
GENDER				•
NATIVE				
SLANG			•	
ESCS	•	•	•	•
JOYREAD	•	•	•	•
DIVREAD		•		
MEMOR	•	•	•	•
ELAB	•	•	•	•
CSTRAT	•	•	•	•
BIC	−996.07	−992.77	−990.55	−990.29
PMP	.76	.15	.05	.04

Note. PMP = posterior model probability; PISA = Program for International Student Assessment; ESCS = measure of economic, social, and cultural status of the students; JOYREAD = enjoyment of reading; DIVREAD = diversity in reading; MEMOR = memorization strategies; ELAB = elaboration strategies; CSTRAT = control strategies; BIC = Bayesian information criterion.

Case Study 1 also compared the predictive performance of a Bayesian model averaging regression to Bayesian and frequentist regressions based on the initially specified regression model. For comparison of predictive performance, we used a measure of 90% prediction coverage, which is the percentage of the observations in the prediction set that fall in their corresponding 90% prediction interval (Hoeting et al., 1999).

Regression Results

Table 1 presents four regression models selected by BMA. Based on the full model in Equation 12, BMA selected four models (Models 1–4) after narrowing down the number of models via Occam's window. The four models are shown in descending order in term of the PMPs and accounts for 100% of the total PMP. Note that the best model (Model 1) accounts for only 76% of the total PMP indicating a fair amount of uncertainty remaining in the model selection. The black dots represent predictors that appear in the respective models. Of the nine predictors in the full model, only five predictors (ESCS, JOYREAD, MEMOR, ELAB, and CSTRAT) appear across all four models.

Table 2. Comparison of the Result of Bayesian Model Averaging to the Result of the Bayesian Regression.

	Bayesian Model Averaging			Bayes Regression		
Predictor	Mean $(\beta\|y)$	*SD* $(\beta\|y)$	$P(\beta \neq 0\|y)\%$	EAP	*SD*	95% PPI
INTERCEPT	495.16	2.35	100.00	486.97	4.79	[477.59, 496.32]
FEMALE	0.20	1.17	4.20	4.27	3.36	[−2.30, 10.88]
NATIVE	0.00	0.00	0.00	−2.16	5.51	[−13.08, 8.40]
SLANG	0.36	1.93	4.80	8.60	6.48	[−3.99, 21.33]
ESCS	28.73	1.74	100.00	28.35	1.85	[24.78, 32.06]
JOYREAD	29.64	1.68	100.00	30.10	1.78	[26.59, 33.63]
DIVREAD	−0.52	1.42	14.60	−3.42	1.68	[−6.70, −0.14]
MEMOR	−20.71	1.86	100.00	−21.02	1.87	[−24.62, −17.32]
ELAB	−16.44	1.78	100.00	−15.66	1.78	[−19.12, −12.18]
CSTRAT	28.37	2.11	100.00	28.56	2.13	[24.39, 32.75]

Note. $N = 2{,}462$. EAP = expected a posterior; *SD* = posterior standard deviation; PPI = posterior probability interval; ESCS = measure of economic, social, and cultural status of the students; JOYREAD = enjoyment of reading; DIVREAD = diversity in reading; MEMOR = memorization strategies; ELAB = elaboration strategies; CSTRAT = control strategies.

We compared results from Bayesian model averaging to results from a single Bayesian regression analysis of the initial model in Table 2 (labeled "Bayes Reg."). Some differences between two results are indicated. For example, the variable NATIVE under Bayesian model averaging has no effect on READING while having a negative effect under the Bayesian regression model. In addition, Bayesian model averaging shows weak evidence for DIVREAD's effect on READING with only 14.6% of the proportion of the nonzero posterior probability of the coefficient (D. Wang, Zhang, & Bakhai, 2004). Bayesian regression, however, indicated DIVREAD as an important predictor.

Prediction Results for the Regression Model

When comparing models in terms of 90% predictive coverage and the logscore rule, Bayesian model averaging yielded a more liberal predictive coverage (96%) compared to the best model from BMA, the Bayesian regression model, or the frequentist regression model as shown in Table 3. It is important to point out that although the predictive performance in this example is somewhat liberal, a small simulation based on this model with

Table 3. Comparison of the Predictive Performance.

Regression Model	Percentage of Predictive Coverage (%)	Log Score of Predictive Coverage
Bayesian model averaging	95.69	−.04
Best model from Bayesian model averaging	90.50	−.10
Bayesian model	90.50	−.10
Frequentist model	90.45	−.10

10,000 replications revealed the mean predictive coverage interval was 0.79–0.98 and centered at 0.90. This small simulation study highlights the importance of supplementing Bayesian prediction with a frequentist calibration. This fusing of Bayesian modeling and frequentist calibration has been discussed generally by Little (2006) and in the context of BMA by Draper (1999).

Case Study 2: Bayesian Model Averaging for the Logistic Regression Model Using ECLS-K 1998

Data and Model

For this case study, we use data from Kaplan and Chen (2012) who studied Bayesian model averaging in the context of propensity score analysis. The data set was randomly sampled from the Early Childhood Longitudinal Study Kindergarten cohort of 1998 (NCES, 2001, ECLS-K). For this example, we model whether full-day or part-day kindergarten attendance can be predicted by 14 variables based on Bayesian model averaging and Bayesian logistic regression. The sample size was 1,000 where 538 children were in full-day programs (FULLDAY = 1) and 462 children in part-day programs (FULLDAY = 0). The sample was evenly divided into a model averaging set ($n = 500$) and a predictive testing set ($n = 500$). Those predictors included GENDER, RACE, mother's employment status (MEMP), child's age at kindergarten entry (AGEENT), child's age at first nonparental care (AGEFRS), primary type of nonparental care (PRIMNW), both parent language to child (LANGUG), number of siblings (NUMSIB), family composition (FAMIL), mother's employment between child's birth and kindergarten (MEMPBK), number of nonparental care arrangement (NUMPRK), social economic status (SES), parent's expectation of

child's degree (EXPECT), and how often parent reads to child (READ). The full logistic regression model we were interested in was:

$$\begin{aligned}\text{logit}(\widehat{\text{FULLTIME}}) = {} & \beta_0 + \beta_1(\text{GENDER}) + \beta_2(\text{RACE}) + \beta_3(\text{MEMP}) \\ & + \beta_4(\text{AGEENT}) + \beta_5(\text{AGEFRS}) + \beta_6(\text{PRIMNW}) \\ & + \beta_7(\text{LANGUG}) + \beta_8(\text{NUMSIB}) + \beta_9(\text{FAMIL}) \\ & + \beta_{10}(\text{MEMPBK}) + \beta_{11}(\text{NUMPRK}) + \beta_{12}(\text{SES}) \\ & + \beta_{13}(\text{EXPECT}) + \beta_{14}(\text{READ}).\end{aligned} \tag{13}$$

Method

As with the linear regression example in Case Study 1, we used noninfomative priors for the logistic regression model for Case Study 2. For Bayesian model averaging, we utilize the "bicglm" function within the R program BMA (Raftery et al., 2015) with Occam's window set to 20. The Bayesian logistic regression model was estimated using the Gibbs sampler as implemented in the "MCMClogit" function within the R package "MCMCpack" package (A. D. Martin et al., 2013). We specify noninformative priors by setting the means for all regression coefficients to 0 and the precisions of all regression coefficients to 0.01 (variance = 100). This specification results in highly diffused priors for all coefficients. The variance of the disturbance term was set to have a noninformative inverse-gamma distribution with shape and scale of 0.001. The analysis used 100,000 iterations, with 5,000 burn-in iterations, a thinning interval of 50, and a Metropolis tuning of 0.25.

Logistic Regression Results

Based on the full logistic regression model in Equation 13, the BMA program selected 11 models, and the best 5 selected models are shown in Table 4. The five models account for 84% of the total PMP. Model 1 accounts for only 44% of the total PMP thus showing that a large amount of uncertainty still exists in the model selection process. In this case study, only two predictors, PRIMNW and FAMIL, appear in the best five models.

From Table 5, we see that the differences between the Bayesian model averaging model and Bayesian logistic model (labeled "Bayes Log. Reg.") are large. With the exception of PRIMNW and FAMIL, the remaining 12 predictors showed little or weak relationships to FULLDAY in the Bayesian

Table 4. The Best Five Selected Models by Bayesian Model Averaging: Logistic Regression Model.

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5
GENDER					
RACE					•
MEMP			•		
AGEENT					
AGEFRS		•			
PRIMNW	•	•	•	•	•
LANGUG					
NUMSIB					
FAMIL	•	•	•	•	•
MEMPBK					
NUMPRK					
SES					
EXPECT					
READ				•	
BIC	−2,426.70	−2,425.10	−2,423.43	−2,422.94	−2,421.99
PMP	0.44	0.20	0.09	0.07	0.04

Note. Cumulative PMP over the best five models = 0.84; PMP = posterior model probability; MEMP = mother's employment status; AGEENT = child's age at kindergarten entry; AGEFRS = child's age at first nonparental care; PRIMNW = primary type of nonparental care; LANGUG = both parent language to child; NUMSIB = number of siblings; FAMIL = family composition; MEMPBK = mother's employment between child's birth and kindergarten; NUMPRK = number of nonparental care arrangement; SES = social economic status; EXPECT = parent's expectation of child's degree; READ = how often parent reads to child; BIC = Bayesian information criterion.

model averaging model. The Bayesian logistic regression model, however, indicated two more important predictors: MEMP and MEMPBK.

Prediction Results for the Logistic Regression Model

We also assessed the predictive performance of the Bayesian model averaging logistic model and compared it to the predictive performance of the Bayesian and frequentist logistic models. To compare predictive performance based on the binary dependent variable, FULLDAY, the Brier (1950) score was adopted. For this study, the Brier score is defined in Equation 11. We see from Table 6 that although the differences are quite small, the Brier score is lowest for

Table 5. Comparison of the Result of Bayesian Model Averaging to the Result of the Bayesian Logistic Regression.

	Bayesian Model Averaging			Bayes Log. Regression		
Predictor	Mean $(\beta\|y)$	*SD* $(\beta\|y)$	*P* $(\beta \neq 0\|y)\%$	EAP	*SD*	95% PPI
INTERCEPT	−0.02	0.47	100.0	−.01	1.70	[−3.30, 3.36]
GENDER	.00	.00	0.0	−.04	0.19	[−0.40, 0.35]
RACE	.00	.02	4.1	−.13	0.10	[−0.33, 0.06]
MEMP	−.01	.05	8.5	−.19	0.09	[−0.36, −0.01]
AGEENT	.00	.00	2.7	.01	0.02	[−0.03, 0.06]
AGEFRS	.00	.01	27.0	−.01	0.01	[−0.03, 0.00]
PRIMNW	−.19	.05	100.0	−.25	0.05	[−0.35, −0.14]
LANGUG	.00	.00	0.0	.08	0.17	[−0.26, 0.42]
NUMSIB	.00	.00	0.0	.03	0.09	[−0.15, 0.20]
FAMIL	.39	.10	100.0	.42	0.11	[0.19, 0.63]
MEMPBK	.02	.10	6.5	.67	0.27	[0.14, 1.20]
NUMPRK	.00	.00	0.0	.00	0.11	[−0.22, 0.21]
SES	.00	.02	2.3	.05	0.14	[−0.21, 0.33]
EXPECT	.00	.01	2.3	−.04	0.10	[−0.22, 0.15]
READ	−.02	.07	10.1	−.21	0.13	[−0.47, 0.05]

Note. $N = 500$. EAP = expected a posterior; *SD* = posterior standard deviation; PPI = posterior probability interval; MEMP = mother's employment status; AGEENT = child's age at kindergarten entry; AGEFRS = child's age at first nonparental care; PRIMNW = primary type of nonparental care; LANGUG = both parent language to child; NUMSIB = number of siblings; FAMIL = family composition; MEMPBK = mother's employment between child's birth and kindergarten; NUMPRK = number of nonparental care arrangement; SES = social economic status; EXPECT = parent's expectation of child's degree; READ = how often parent reads to child.

Table 6. Comparison of the Predictive Performance.

Logistic Regression Model	Brier Score
Bayesian model averaging	.24566
Best model from Bayesian model averaging	.24644
Bayesian model	.25786
Frequentist model	.25674

the Bayesian model averaging logistic regression model indicating better predictive performance compared to the best model selected by BMA, the single Bayesian logistic regression model, or the frequentist regression model.

Case Study 3: Bayesian Model Averaging for SEM Using PISA 2009

Brief Definition and History of SEM

Following Kaplan (2009), SEM can perhaps best be defined as a class of methodologies that seeks to represent hypotheses about summary statistics derived from empirical measurements in terms of a smaller number of "structural" parameters defined by a hypothesized underlying model. The history of SEM can be roughly divided into two generations. The *first generation* of SEM began with the initial merging of confirmatory factor analysis and simultaneous equation modeling (see, e.g., Jöreskog, 1973). In addition to these founding concepts, the first generation of SEM witnessed important methodological developments in handling nonstandard conditions of the data. These developments included methods for dealing with nonnormal data, missing data, and sample size sensitivity problems (see, e.g., Kaplan, 2009). The *second generation* of SEM could be broadly characterized by another merger: This time, combining models for continuous latent variables developed in the first generation with models for categorical latent variables (see B. Muthén, 2001). The integration of continuous and categorical latent variables into a general modeling framework was due to the extension of finite mixture modeling to the SEM framework. This extension has provided an elegant theory, resulting in a marked increase in important applications. These applications include, but are not limited to, methods for handling the evaluation of interventions with noncompliance (Jo & Muthen, 2001), discrete-time mixture survival models (B. Muthén & Masyn, 2005), and models for examining unique trajectories of growth in academic outcomes (Kaplan, 2003).

A parallel development to first and second generation SEM has been the expansion of Bayesian methods for complex statistical models, including SEMs. Early papers include J. K. Martin and McDonald (1975), Lee (1981), and Scheines, Hoijtink, and Boomsma (1999). Lee (2007) provides a review and extensions of BSEM. The increased use of Bayesian tools for statistical modeling has come about primarily as a result of progress in computational algorithms based on MCMC sampling. The MCMC algorithm is implemented in software programs such as WinBugs (Lunn, Thomas, Best, & Spiegelhalter, 2000), Mplus (L. K. Muthén & Muthén, 1998–2010), and various packages within the R archive (R Core Team, 2017), such as "rjags" and "rstan."

Method

We focus our attention on SEMs among observed variables. Following the notation by Kaplan and Lee (2015, see also; Kaplan & Depaoli, 2012; Kaplan, 2009), a SEM can be specified as follows. Let

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}, \tag{14}$$

where $\mathbf{y}$ is a vector of manifest endogenous variables and $\mathbf{x}$ be a vector of observed exogenous variables with covariance matrix Φ. Further, let $\boldsymbol{\alpha}$ is 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 Ψ assumed to be diagonal.

Conjugate priors for SEM parameters. To specify prior distributions on all model parameters, we follow the notation of Kaplan and Depaoli (2012) and arrange the model parameters as sets of common conjugate prior distributions. Parameters with the subscript norm follow a normal distribution, while those with the subscript IW follow an inverse Wishart distribution. Let $\boldsymbol{\theta}_{\text{norm}} = \{\boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\Gamma}\}$ be the vector of free model parameters that are assumed to follow a normal distribution, and let $\boldsymbol{\theta}_{\text{IW}} = \{\Phi, \Psi\}$ be the vector of free model parameters that are assumed to follow the inverse Wishart distribution. Formally, we write:

$$\boldsymbol{\theta}_{\text{norm}} \sim N(\mu, \Omega), \tag{15}$$

where μ and Ω are the mean and variance hyperparameters, respectively, of the normal prior. For blocks of variances and covariances in Ξ and Ψ, we assume that the prior distribution is inverse Wishart, that is,[4]

$$\boldsymbol{\theta}_{\text{IW}} \sim \text{IW}(\mathbf{R}, \delta), \tag{16}$$

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.

The "BMASEM" algorithm. Following closely the recent paper by Kaplan and Lee (2015), our approach to Bayesian model averaging for SEMs draws on the fact that path diagrams within the SEM tradition can be seen as special cases of so-called *directed acyclic graphs* (DAGs), the latter having been developed by Pearl (2009). Bayesian model averaging over DAGs has also

been discussed in Madigan and Raftery (1994); however, a review of the extant literature indicates that Bayesian model averaging over DAGS has not been utilized in education and not fully developed for SEM.

In this section, we describe the full algorithm used to conduct Bayesian model averaging for SEMs for Case Study 3. The general steps of the algorithm are as follows: (a) specify an initial model of interest recognizing that this may not be the model that generated the data; (b) starting with the initial model represented as a DAG, implement a search over the DAG to reduce the space of models to a reasonable size while maintaining the distinction between exogenous, mediating, and endogenous variables; (c) obtain the PMPs for each model; (d) obtain the weighted average of structural parameters over each model, weighted by the PMPs; and (e) compare predictive performance of the Bayesian model averaging SEM to the initially specified BSEM by computing the reduced form of the models and calculating the log score and/or the predictive coverage (Kaplan & Lee, 2015).

Model selection via the up and down algorithm. We apply the search algorithm first suggested by Madigan and Raftery (1994) and implemented by Kaplan and Lee (2015) that they refer to as the "up and down algorithm." Starting with an initial model, the algorithm first executes the "down" algorithm, wherein each model in the set is compared with its submodels. If there is a model with no submodel in the down algorithm, then 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. Occam's window and Occam's razor are employed in the up and down algorithm to select the submodels which predict as well as the best model. The details of the algorithm are given in Kaplan and Lee (2015).

Averaging over SEM parameters. A set of K possible SEMs $(k = 1, 2, ..., K)$ in the set A are chosen through the up and down algorithm. With this set, the PMPs are obtained using a BIC approximation as:

$$p(M_k|D) = \frac{\exp(-.5 \times \Delta\text{BIC})}{\sum_{l=1}^{K} \exp(-.5 \times \Delta\text{BIC})}, \tag{17}$$

where ΔBIC is the difference of BIC of M_k and the maximum of BICs of all the models in the set. The PMPs 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

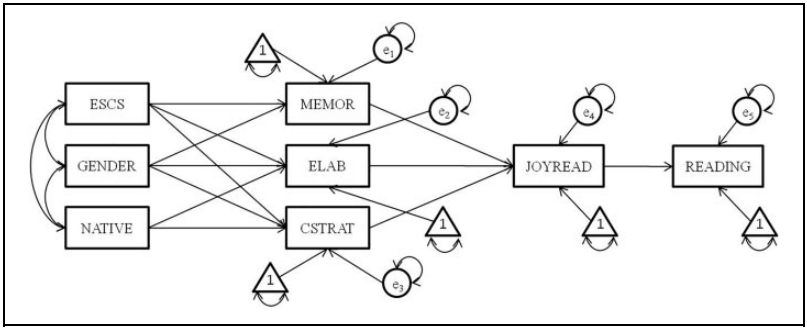

Figure 1. Original path model based on the Program for International Student Assessment data for Case Study 3.

posterior distributions for the set of selected models, weighted by their PMPs. The posterior mean for a parameter θ under model M_k, given the data D, can be written as (see Kaplan & Lee, 2015):

$$E(\theta|D, M_k) = \sum_{M_k \in A} \hat{\theta} p(M_k|D). \tag{18}$$

Data and Model

Case Study 3 used the same data set as used for Case Study 1 but selected 8 of the 10 variables including ESCS, GENDER, NATIVE, MEMOR, ELAB, CSTRAT, JOYREAD, and READING for SEM. The model of interest is illustrated in Figure 1. The first three background variables were exogenous variables and the rest five variables were endogenous variables. Of those endogenous variables, three reading strategy variables were indicators for JOYREAD which was an indicator for READING. The sample size was 4,979 after the list-wise deletion. The sample was then split into a model averaging set ($n = 2{,}489$) and a predictive testing set ($n = 2{,}490$).

For SEM Bayesian model averaging, we used the R package BMASEM which is available at http://bise.wceruw.org/publications.html. In the BMA-SEM program, the value of 100 was chosen for Occam's window, and the model in Figure 1 was used as a starting model for the model comparison in the down algorithm. A BSEM based on Figure 1 was estimated with non-informative conjugate priors for all model parameters was conducted using "rjags" (Plummer, 2016), "coda" (Plummer, Best, Cowles, & Vines, 2006),

and "MCMCpack" R packages. The MCMC algorithm for this case study was set to 500,000 iterations, with 5,000 burn-in iterations, a thinning interval of 50 from two chains starting at different locations in the posterior distribution.

Predictive performance for Bayesian model averaging SEM. We compared the predictive performance of Bayesian model averaging under SEM to the predictive performance based on the initially specified BSEM. Kaplan and Lee (2015) estimated the predictive performance of the model under Bayesian model averaging by transforming the structural form of the model into the reduced form where the endogenous variables are on the left side of the equation and the exogenous variables on the right side. The structural form was given in Equation 14 and can be rewritten as:

$$(\mathbf{I} - \mathbf{B})\mathbf{y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}. \tag{19}$$

Assuming that $(\mathbf{I} - \mathbf{B})$ is nonsingluar,

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1}\alpha + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{x} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta} \tag{20}$$

$$\mathbf{y} = \Pi_0 + \Pi_1\mathbf{x} + \boldsymbol{\zeta}^*, \tag{21}$$

where Π_0 is the vector of reduced form intercepts, Π_1 the vector of reduced form slopes, and $\boldsymbol{\zeta}^*$ the vector of reduced form disturbances with variance matrix, Ψ^*. With the structural form of the model transformed into the reduced form, we can obtain and compare predicted values in a manner similar to Case Study 1. Specifically, the comparison procedure is as follows (Kaplan & Lee, 2015):

1. Randomly divide the data set into a model-averaging set and a predictive testing set.
2. Fit a single BSEM and Bayesian model averaging 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 BSEM and Bayesian model averaging.
5. Compare their predictive performance based on 90% of predictive coverage.

Table 7. Selected Model by Bayesian Model Averaging: Structural Equation Modeling Using Program for International Student Assessment 2009.

Predictor	Model 1	Model 2	Model 3	Model 4
MEMOR ~ ESCS	•	•	•	•
MEMOR ~ GENDER	•	•	•	•
MEMOR ~ NATIVE	•	•	•	
ELAB ~ ESCS	•	•	•	•
ELAB ~ GENDER				
ELAB ~ NATIVE	•	•		•
CSTRAT ~ ESCS	•	•	•	•
CSTRAT ~ GENDER	•	•	•	•
CSTRAT ~ NATIVE	•		•	•
JOYREAD ~ ESCS	•	•	•	•
JOYREAD ~ GENDER	•	•	•	•
JOYREAD ~ NATIVE				
JOYREAD ~ MEMOR	•	•	•	•
JOYREAD ~ ELAB	•	•	•	•
JOYREAD ~ CSTRAT	•	•	•	•
READING ~ JOYRREAD	•	•	•	•
BIC	48,263.14	48,266.63	48,266.98	48,267.26
PMP	0.69	0.12	0.10	0.09

Note. PMP = posterior model probability. ESCS = measure of economic, social, and cultural status of the students; JOYREAD = enjoyment of reading; DIVREAD = diversity in reading; MEMOR = memorization strategies; ELAB = elaboration strategies; CSTRAT = control strategies; BIC = Bayesian information criterion; NATIVE = immigrant status.

SEM Results

Based on the model in Figure 1, the algorithm used in the BMASEM package selected four models for $C = 100$, resulting four models which accounted for 100% of the total PMP. Note again, that the best model in terms of the BIC has a PMP of 0.69, again indicating a relatively high degree of posterior model uncertainty. Table 7 displays the selected model for this example. Table 8 presents the results from the Bayesian model averaging SEM and the BSEM. There were four regressions set to 0 in the initial model including MEMOR on NATIVE and JOYREAD on ESCS, GENDER, and NATIVE. With the exception of JOYREAD on NATIVE, the remaining three regressions appeared in best model based on the BIC. On the contrary, the regression of ELAB on GENDER which was in the original model does not appear in the best model.

Table 8. Comparison of the Result of Bayesian Model Averaging to the Result of the Bayesian SEM.

	Bayesian Model Averaging			BSEM		
Predictor	Mean $(\beta\|y)$	*SD* $(\beta\|y)$	*P* $(\beta \neq 0\|y)\%$	EAP	*SD*	95% PPI
MEMOR ~ ESCS	0.09	.03	100.00	0.07	.02	[0.02, 0.12]
MEMOR ~ GENDER	0.18	.04	100.00	0.18	.04	[0.09, 0.27]
MEMOR ~ NATIVE	−0.19	.08	91.21	—	—	—
ELAB ~ ESCS	0.16	.03	100.00	0.16	.03	[0.11, 0.21]
ELAB ~ GENDER	0.00	.00	0.00	−0.05	.04	[−0.14, 0.04]
ELAB ~ NATIVE	−0.18	.08	89.90	−0.20	.06	[−0.32, −0.09]
CSTRAT ~ ESCS	0.29	.03	100.00	0.29	.02	[0.25, 0.34]
CSTRAT ~ GENDER	0.25	.04	100.00	0.25	.04	[0.18, 0.33]
CSTRAT ~ NATIVE	−0.17	.08	87.98	−0.20	.05	[−0.30, −0.09]
JOYREAD ~ ESCS	0.13	.02	100.00	—	—	—
JOYREAD ~ GENDER	0.64	.04	100.00	—	—	—
JOYREAD ~ NATIVE	0.00	.00	0.00	—	—	—
JOYREAD ~ MEMOR	−0.11	.02	100.00	−0.11	.02	[−0.16, −0.06]
JOYREAD ~ ELAB	0.08	.02	100.00	0.04	.02	[−0.01, 0.08]
JOYREAD ~ CSTRAT	0.28	.02	100.00	0.36	.02	[0.31, 0.41]
READING ~ JOYRREAD	0.34	.02	100.00	0.34	.02	[0.31, 0.38]
MEMOR ~ 1	0.01	.07	100.00	−0.14	.03	[−0.20, −0.08]
ELAB ~ 1	0.02	.07	100.00	0.06	.06	[−0.05, 0.18]
CSTRAT ~ 1	−0.09	.07	100.00	−0.08	.05	[−0.17, 0.02]
JOYREAD ~ 1	−0.36	.03	100.00	−0.02	.02	[−0.06, 0.02]
READING ~ 1	5.00	.02	100.00	5.00	.02	[4.96, 5.03]
MEMOR ~ ~ MEMOR	1.19	.03	100.00	1.20	.02	[1.14, 1.27]
ELAB ~ ~ ELAB	1.23	.03	100.00	1.23	.02	[1.17, 1.30]
CSTRAT ~ ~ CSTRAT	1.18	.03	100.00	0.98	.02	[0.95, 1.01]
JOYREAD ~ ~ JOYREAD	0.89	.03	100.00	0.98	.02	[0.95, 1.01]
READING ~ ~ READING	0.76	.02	100.00	0.98	.02	[0.95, 1.01]

Note. $N = 2{,}490$. Symbols ~ refers to regression of left-hand variable onto right-hand variable, ~ 1 refers to intercept, and ~ ~ refers to variance. EAP = expected a posteriori; *SD* = posterior standard deviation; PPI = posterior probability interval; ESCS = measure of economic, social, and cultural status of the students; JOYREAD = enjoyment of reading; DIVREAD = diversity in reading; MEMOR = memorization strategies; ELAB = elaboration strategies; CSTRAT = control strategies; NATIVE = immigrant status; READING = reading proficiency.

Table 9 presents the prediction results for the BSEM example. As predicted by the theory of Bayesian model averaging, we observe modestly better predictive performance in terms of 90% predictive coverage and the

Table 9. Comparison of the Predictive Performance.

Structural Equation Modeling	90% Coverage	Log Score
Bayesian model averaging	.88	−.13
Bayesian model	.87	−.14
Frequentist model	.87	−.14

log-score rule under Bayesian model averaging compared to Bayesian and frequentist approaches.

Discussion

We divide our Discussion section into four parts. First, we discuss some remaining important technicalities. Second, we discuss some limitations of BMA. Third, we discuss a frequentist approach to model averaging. Last, we discuss the importance of model averaging for large-scale assessment design.

Some Remaining Technicalities

Throughout this article, a subtle assumption was invoked but not discussed; namely, that the true model, say, M_T was one of the models in the set of models M_k, $k = 1, 2, ..., K$. This assumption is referred to as the M-closed framework discussed in Bernardo and Smith (2000) and Clyde and Iversen (2015). The M-closed framework can be contrasted with the M-completed framework and the M-open framework. In the M-closed framework, it makes sense to assign prior probabilities that M_T is in the space of models. In fact, this is the framework that underlies the standard approach to BMA discussed in this article; prior probabilities are assigned to the set of models (typical the indifference prior $1/M$) encoding ones belief that each model is equally likely to be the true model. The application of the indifference prior is the conventional default in the BMA software program used in this article (Raftery et al., 2015). In the M-completed and M-open frameworks, M_T is not in the set of models M_k, which are simply considered proxies to be compared. As such, the assignment of prior probabilities makes less sense and the question comes down to how models are to be chosen and averaged if the true model does not exist within the set of possible models.

Following Bernardo and Smith (2000, p. 385), in the *M*-completed framework, the analyst has entertained a true model M_T, but that this true model lies outside the set of models M_k, $k = 1, 2, ..., K$ being considered. The set of models M_k is enumerated for purposes of scientific communication and is evaluated in light of the true model M_T. By contrast, in the *M*-open framework, the analyst does not even entertain the existence of M_T and so here again, it does not make sense to assign prior probabilities. The range of models M_k is enumerated for comparison purposes but not with reference to the existence of a true model per se. The distinction among these modeling frameworks is quite important, and indeed, recent work by Clyde and Iversen (2015) have used a decision-theoretic framework that allows BMA within the *M*-open framework. Future research should be directed toward examining Clyde and Iversen's framework for BMA in the context of large-scale educational survey research.

Practical Limitations of Bayesian Model Averaging

Bayesian model averaging is not without a set of limitations that must be considered if it is to be employed in practical applications. First, BMA is sensitive to problems of collinearity. Following Draper's (1999) commentary on Hoeting et al. (1999), a problem with the indifference prior suggested by Hoeting et al. (1999) arises when two models are essentially identical in terms of their predictions. This can occur in models with nearly collinear sets of variables. In this case, as Hoeting et al. point out, assuming an indifference prior across the models results in placing twice as much weight on a single model, of which there are two slightly different versions. Draper (1999) shows this issue with the following example. Consider a model M_1 of some outcome variable y and predictors x_1, x_2, and a predictor x_3 which is collinear with x_2. Then, being forced to ignore x_3, a researcher using an indifference prior on the model space would weight each model by 1/4—namely, $M_1 = \{\text{no predictors}, x_1, x_2, (x_1, x_2)\}$. However, if another researcher includes x_3, this would lead to weights 1/8 for each model $M_2 = \{\text{no predictors}, x_1, x_2, x_3, (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_1, x_2, x_3)\}$. The problem is that the last two models in M_2 will fail to be estimated because of the collinear variable x_3, and this variable will have to be dropped. This will lead to a third model M_3 with weights 1/6—namely, $M_3 = \{\text{no predictors}, x_1, x_2, x_3, (x_1, x_2), (x_1, x_3\}$, which will then result in putting weights $\{1/6, 1/6, 1/3, 1/3\}$ on M_1. As a result of this issue, some thought needs to be given regarding the use of collinear variables in BMA with respect to model weights. Although to our knowledge, the issue of collinearity has not

been studied directly within BMA, it seems that an initial examination of collinearity diagnostics (e.g., Belsley, Kuh, & Welsch, 2005) prior to employing BMA is warranted.

A second practical issue with BMA is the inclusion of interaction terms. It was noted earlier that the case studies used in this article did not employ interaction terms, although interaction terms can be used in BMA. Specifically, the problem with interaction terms in BMA, as pointed out by Montgomery and Nyhan (2010) and Raftery, Hoeting, Volinsky, Painter, and Yeung (2015), is that averaging over interaction terms is problematic if a model is included in which one of the main effects involved in the interaction is dropped. This assumes that the main effect was 0, and if this assumption is false, then the interaction term will be incorrectly estimated. In the case where models are enumerated within BMA for theoretical purposes (as opposed to simply traversing the space of possible models), then Montgomery and Nyhan (2010) advocate averaging over the subset of models that include the main effects and interactions because this will lead to posterior distributions that are theoretically consistent and correctly estimated. In addition, Clyde (2003) pointed out that when models contain interaction terms, specifying independent priors across the model space may not be appropriate.

Non-Bayesian Approaches to Model Averaging

It should be noted that issues of model averaging and predictive performance are not restricted to the domain of Bayesian statistics. A considerable amount of theoretical and practical work has focused on frequentist approaches and data mining approaches to model averaging and predictive performance (see, e.g., Hjort & Claeskens, 2003; Strobl, 2013; Strobl, Malley, & Tutz, 2009; H. Wang, Zhang, & Zou, 2009). One approach to frequentist model averaging that bears a strong resemblance to BMA is based on the use of Akaike weights. Akaike weights have been discussed in Wagenmakers and Farrell (2004) and H. Wang, Zhang, and Zou (2009). In essence, Akaike (1973, 1985) weights are transformations of the Akaike information criterion (AIC), such that they can be interpreted as conditional probabilities for each model under consideration. Akaike weights have the advantage of allowing a more nuanced choice among models when the difference between an AIC value for one model versus the AIC for the best model (lowest AIC) is small. A clear difference between the use of Akaike weights and BMA is the absence of priors on the model space or parameter space in the frequentist case.

Nevertheless, it is crucial that future research provides detailed comparisons of these methods in terms of predictive performance.

Implications for Large-Scale Assessment Design

The subject-matter motivation for this article concerned using BMA for prediction with large-scale educational assessments. Although our examples in this article and elsewhere (Chen & Kaplan, 2015; Kaplan & Chen, 2012, 2014) demonstrate the potential of using BMA for educational policy analysis, additional research and development are required before BMA can be fully implemented for building predictive models with large-scale educational assessments. Specifically, an important feature of large-scale assessments such as TIMSS, PISA, or ECLS concerns the complexities of the sampling design. For example, following Kaplan and Kuger (2016), the nature of the sampling design for PISA ensures that the sample of students for a given country is chosen in such a way as to accurately represent the national population of 15-year-olds for that country. However, within countries, the selection probabilities to attain national representativeness might be different and so survey weights along with Bayesian hierarchical modeling need to be employed to ensure that each sampled student represents the appropriate number of students in the PISA-eligible population within a particular country. If the goal is to develop optimal predictive models in the context of education policy using large-scale assessment data, it is crucial that the nuances of the survey design be addressed. Because model averaging is, by definition and practice, a model-based methodology, future research will require focusing on model-based rather than designed-based inference (see, e.g., Little, 2004).

In addition to addressing the complex sampling design of large-scale educational assessments, it is necessary to address the design and implementation of the assessment instruments themselves. This is particularly important when the goal is to develop optimally predictive models for achievement outcomes because the implementation of the achievement tests in assessments such as PISA and TIMSS uses so-called matrix sampling designs. Matrix sampling designs yield "designed missing data" insofar as no student receives all of the test items. To obtain broad content coverage, a student will receive a subset of items not all of which are in common with the subset of items received by another student. As a result, precision is sacrificed for content coverage and individual scores cannot be reported. Instead of providing a single score for each student, a set of plausible values for each student are provided that represent the set of

plausible scores from a distribution of scores derived from a so-called conditioning model (see von Davier, 2013). It is essential that these scores are analyzed correctly in terms of obtaining point estimates and standard errors, and rules have been developed by Rubin (1987) to properly analyze plausible values. For Bayesian model averaging to be correctly utilized in large-scale educational assessments, the methodology must be extended to handle plausible value methodology.

Conclusion

The typical practice of statistical modeling for educational research and policy analysis has been to specify, estimate, and test a specific model of interest; examining the fit of the model to the data; and examining the statistical significance of parameters of interest. The development of predictive statistical models in the domain of education research has, arguably, received somewhat less attention. The question of using a model for some purpose beyond assessing model fit leads to a consideration of the accuracy of a model's predictions and this focus on predictive accuracy is a central feature of Bayesian statistics—arguably more central than the traditional ideas of goodness of fit. Indeed, the BF, BIC, and the DIC focus our attention on choosing models based on considering posterior predictive accuracy. If the goal of model building is one of predictive accuracy, then attachment to one's specific model is of less importance. Thus, we are less concerned about the fit of a theoretical model and more concerned about finding a model that will predict well.

In the Bayesian framework, Bayesian model averaging 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 PMPs—yet all models contain some useful information. By combining models while at the same time accounting for model uncertainty, we obtain a "stronger" model in terms of predictive accuracy.

To conclude, the purpose of this article was to demonstrate a well-known and useful approach to model averaging in the Bayesian domain with the goal of improving the predictive accuracy of common statistical models. We concur with the famous quote by Box and Draper (1987), "All models are wrong, but some are useful" (p. 424), but we add that even though models capture only a small portion of the data generating process, each model retains a degree of useful information. In particular, beyond capturing the data generating process, a model's usefulness lies in its predictive

capacity. Here, then, Bayesian model averaging provides an approach to optimizing the predictive utility of a large number of otherwise wrong models. Nevertheless, for this didactic article, we show that the theory of Bayesian model averaging works as expected, yielding models with better predictive performance than any given submodel including the initial model of interest. As always, the full benefit Bayesian model averaging will rest on its application to practical problems where prediction is of high priority.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Chansoon Lee http://orcid.org/0000-0002-3669-3019

Notes

1. The Laplace method of integrals is based on a Taylor expansion of a function $f(u)$ of a q-dimensional vector u. The approximation is $\int e^{f(u)}du \simeq 2(\pi)^{q/2}|A|^{1/2}\exp\{f(u^*)\}$, where u^* is the value of u at which f attains its maximum, and A is minus the inverse of the Hessian of f evaluated at u^*. Following Raftery (1996, p. 253), when the Laplace method is applied to Equation 3, we obtain the approximation $p(y|M_k) \simeq (2\pi)^{q_k}|A_k|^{1/2}p(y|\tilde{\theta}_k, M_k)p(\tilde{\theta}_k, M_k)$, where q_k is the dimension of θ_k, $\tilde{\theta}_k$ is the posterior model of θ_k, and A_k is minus the inverse of the Hessian of $\log\{p(y|\theta_k, M_k)p(\theta_k|M_k)\}$, evaluated at the posterior mode $\tilde{\theta}_k$.
2. 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 does not present a problem.
3. Note that it is not technically proper to use only one plausible value in a statistical analysis. A direction of future research will require developing Bayesian model averaging when analyzing multiple plausible values.
4. Note that in the case where there is only one element in the block, the prior distribution is assumed to be inverse-gamma, that is, $\boldsymbol{\theta}_{\mathrm{IW}} \sim \mathrm{IG}(a, b)$.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1985). Prediction and entropy. In A. C. Atkinson & S. E. Feinberg (Eds.), *A celebration of statistics* (pp. 1–24). New York, NY: Springer-Verlag.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: Wiley.

Bernardo, J., & Smith, A. F. M. (2000). *Bayesian theory*. New York, NY: Wiley.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model building and response surfaces*. New York, NY: John Wiley & Sons.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Chen, J., & Kaplan, D. (2015). Covariate balance in a two-step Bayesian propensity score approach for observational studies. *Journal of Research on Education Effectiveness*, *8*, 280–302.

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

Clyde, M. A. (2003). Model averaging. In S. James Press (Ed.), *Subjective and objective Bayesian statistics* (pp. 320–335). Hoboken, NJ: Wiley-Interscience.

Clyde, M. A. (2017). *BAS: Bayesian adaptive sampling for Bayesian model averaging* (R package version 1.4.7) [Computer software manual].

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

Clyde, M. A., & Iversen, E. S. (2015). Bayesian model averaging in the M-open framework. In P. Damien, P. Dellaportas, N. G. Polson, & D. A. Stephens (Eds.), *Bayesian theory and applications* (pp. 483–498). Oxford, England: Oxford University Press.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, *147*, 278–202.

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

Draper, D. (1999). Model uncertainty, yes, discrete model averaging, maybe. Comment on Hoeting, Madigan, Raftery and Volinsky. *Statistical Science*, *14*, 405–409.

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., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies: With commentary. *Statistical Science*, *6*, 733–807.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*, 879–899.

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

Jo, B., & Muthen, B. (2001). Modeling of intervention effects with noncompliance: A latent variable modeling approach for randomized trials. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 57–87). Mahwah, NJ: Lawrence Erlbaum Associates.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York, NY: Academic Press.

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research*, *56*, 1146–1157.

Kaplan, D. (2003). Methodological advances in the analysis of individual growth with relevance to education policy. *Peabody Journal of Education*, *77*, 189–215.

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 Press.

Kaplan, D., & Chen, J. (2012). A two-step Bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, *77*, 581–609. doi:10.1007/s11336-012-9262-8

Kaplan, D., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research*, *49*, 505–517.

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.

Kaplan, D., & Kuger, S. (2016). The methodology of PISA: Past, present, and future. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning—An international perspective* (pp. 53–73). Dordrecht, the Netherlands: Springer.

Kaplan, D., & Lee, C. (2015). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling*. doi:10.1080/10705511.2015.1092088

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

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York, NY: Wiley.

Lee, S.-Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika*, *46*, 153–160.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. New York, NY: Wiley.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, *99*, 546–556.

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, *60*, 213–223.

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

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.

Martin, A. D., Quinn, K. M., & Park, J. H. (2013, May 13). *Markov chain Monte Carlo (MCMC) package*. Retrieved from http://mcmcpack.wustl.edu/

Martin, J. K., & McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika*, *40*, 505–517.

Merkle, E. C., & Steyvers, M. (2013a). Choosing a strictly proper scoring rule. *Decision Analysis*, *10*, 292–304.

Merkle, E. C., & Steyvers, M. (2013b). Choosing a strictly proper scoring rule. *Decision Analysis*, *10*, 292–304.

Montgomery, J. M., & Nyhan, B. (2010). Bayesian model averaging: Theoretical developments and practical applications. *Political Analysis*, *18*, 245–270.

Mullis, I. V. S. (2013). Introduction. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 3–9). Boston, MA: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In A. Sayer & L. Collins (Eds.), *New methods for the analysis of change* (pp. 291–322). Washington, DC: APA.

Muthén, B., & Masyn, K. (2005). Mixture discrete-time survival analysis. *Journal of Educational and Behavioral Statistics*, *30*, 27–58.

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

National Center for Educational Statistics. (2001). *Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use data files user's manual* (Tech. Rep. No. NCES 2001-029). Washington, DC: U.S. Government Printing Office.

Organization for Economic Cooperation and Development. (2009). *PISA 2009 assessment framework—Key competencies in reading, mathematics and science*. Paris, France: Author.

Organization for Economic Cooperation and Development. (2010). *PISA 2009 results* (Vol. I–VI). Paris, France: Author.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.

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

Plummer, M. (2016). rjags: Bayesian graphical models using MCMC (R package version 4-6) [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rjags

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

R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, *83*, 251–266.

Raftery, A. E. (1998). *Bayes factors and the BIC: Comment on Weakliem* (Tech. Rep. No. 347). Seattle: University of Washington, Department of Statistics.

Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. Y. (2015, June 22). *Bayesian model averaging (BMA)* (version 3.12). Retrieved from http://www2.research.att.com/volinsky/bma.html

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.

Rubin, D. B. (1987). *Multiple imputation in nonresponse surveys*. Hoboken, NJ: Wiley.

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

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

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.

Strobl, C. (2013). Data mining. In T. Little (Ed.), *The oxford handbook of quantitative methods in psychology* (pp. 678–700). New York, NY: Oxford University Press.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*, 323–348.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175–202). Boca Raton, FL: Chapman Hall/CRC.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.

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.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*, 1–60.

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

Zeugner, S., & Feldkircher, M. (2015). Bayesian model averaging employing fixed and flexible priors: The BMS package for R. *Journal of Statistical Software*, *68*, 1–37. doi:10.18637/jss.v068.i04

Author Biographies

David Kaplan is the Patricia Busk Professor of Quantitative Methods in the Department of Educational Psychology at the University of Wisconsin – Madison. His research interests are in Bayesian statistics, missing data problems, and large-scale educational assessment design.

Chansoon Lee is a psychometrician at the National Council of State Boards of Nursing. Her research interests include Bayesian model averaging, machine learning, decision tree models, and predictive modeling.