PSYCHOMETRIKA—VOL. 86, NO. 1, 215–238
MARCH 2021
https://doi.org/10.1007/s11336-021-09754-5

ON THE QUANTIFICATION OF MODEL UNCERTAINTY: A BAYESIAN PERSPECTIVE

DAVID KAPLAN
UNIVERSITY OF WISCONSIN–MADISON

Issues of model selection have dominated the theoretical and applied statistical literature for decades. Model selection methods such as *ridge regression*, *the lasso*, and the *elastic net* have replaced ad hoc methods such as stepwise regression as a means of model selection. In the end, however, these methods lead to a single final model that is often taken to be the model considered ahead of time, thus ignoring the uncertainty inherent in the search for a final model. One method that has enjoyed a long history of theoretical developments and substantive applications, and that accounts directly for uncertainty in model selection, is *Bayesian model averaging* (BMA). BMA addresses the problem of model selection by not selecting a final model, but rather by averaging over a space of possible models that could have generated the data. The purpose of this paper is to provide a detailed and up-to-date review of BMA with a focus on its foundations in Bayesian decision theory and Bayesian predictive modeling. We consider the selection of parameter and model priors as well as methods for evaluating predictions based on BMA. We also consider important assumptions regarding BMA and extensions of model averaging methods to address these assumptions, particularly the method of *Bayesian stacking*. Simple empirical examples are provided and directions for future research relevant to psychometrics are discussed.

Key words: Bayesian model averaging, Bayesian stacking, prediction.

Issues surrounding the specification of statistical models for prediction and explanation have dominated the theoretical and applied statistical literature for decades. The underlying issue has been one of the well-known bias-variance trade-off problem—namely that under-specified models can lead to parameter bias, and over-specified models can lead to poor predictions in future samples. Model selection methods such as *ridge regression* (Hoerl and Kennard 1970; Hoerl 1985), *the lasso* (Tibshirani 1996), the *elastic net* (Zou and Hastie 2005), and their Bayesian counterparts (Hsiang 1975; Park and Casella 2008; Li and Lin 2010) among others, have been advocated as a means of regularizing models to provide a balance between bias and variance. In the end, however, these methods lead to a single final model that is often taken to be the model considered ahead of time. The problem is that, in practice, model uncertainty often goes unnoticed, and the impact of this uncertainty can be quite serious. Hoeting et al. (1999) wrote

> 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. (pg. 382)

Similar observations were made earlier by Leamer (1978, pg. 91) who noted that

> ...ambiguity over a model should dilute information about the regression coefficients, since part of the evidence is spent to specify the model

and by Draper et al. (1987, pg. iii) who stated

Correspondence should be made to David Kaplan, University of Wisconsin–Madison, Madison, USA. Email: dkaplan@education.wisc.edu

> This [model selection] tends to underestimate Your actual uncertainty, with the result that Your actions both inferentially in science and predictively in decision-making, are not sufficiently conservative. [Capitalization authors.]

The Bayesian framework recognizes that model selection is conducted under pervasive uncertainty insofar as a particular model is typically chosen among a set of competing models that could also have generated the data. Although a number of methods exist in the Bayesian literature to aid in improving model prediction, including sensitivity analyses via posterior predictive checking (Gelman et al. 1996) as well as the Bayesian regularization methods cited earlier, in the end, a single model is chosen for prediction and/or explanatory purposes. As the quotes by Hoeting et al. (1999), Leamer (1978), and Draper et al. (1987) suggest, it is risky to settle on a single model. Rather, it may be prudent to draw predictive strength through combining models. Arguably, the most popular approach for addressing the problem of model uncertainty from a Bayesian perspective lies in the method of *Bayesian model averaging* (BMA).

The purpose of this paper is to provide a detailed review of the methodology of BMA. A review of the extant literature on BMA reveals interesting connections to Bayesian decision theory and Bayesian predictive modeling. As such, we will draw on a number of seminal sources addressing the problem of model uncertainty quantification from a Bayesian perspective. In particular, we will draw heavily from Bernardo and Smith (2000), Vehtari and Ojanen (2012), Piironen and Vehtari (2017), and Draper (2013) with respect to the idea of belief models and Bayesian decision theory; Bernardo and Smith (2000), Clyde and Iversen (2013), and Clarke and Clarke (2018) with respect to so-called $\mathcal{M}$*-frameworks*; Raftery and his colleagues (Madigan and Raftery 1994; Hoeting et al. 1999; Raftery et al. 1997) with respect to historical developments in BMA and algorithms; and Clyde and Iversen (2013), Vehtari et al. (2019), and Yao et al. (2018) with respect to Bayesian stacking as means of addressing certain assumptions underlying BMA.

The organization of this paper is as follows. In the next section, we discuss Bayesian predictive modeling as embedded in Bayesian decision theory. Here we discuss the concepts of expected utility and expected loss, and frame these ideas within the use of information-theoretic methods for judging decisions. We show that the action which optimizes the expected utility is the BMA solution. Next, we discuss the statistical elements of BMA including connections to Bayes factors, computation considerations, and the problem parameter and model priors. This is followed by a simple example of BMA in linear regression modeling and a comparison of results based on different parameter and model prior settings. This is then followed by a presentation of methods for evaluating the quality of a solution based on BMA, including the use of scoring rules and how they tie back to the information-theoretic concepts discussed earlier in the paper. We then discuss the main problem associated with conventional BMA—namely that BMA assumes that the true data generating model is contained in the set of models that are being averaged and demonstrate the method of Bayesian stacking that directly deals with this assumption. A simple example of Bayesian stacking is provided. The paper concludes with a discussion of challenges and opportunities associated with considering model uncertainty in psychometrics and the social and behavioral sciences more generally.

1. Elements of Predictive Modeling

This overview of BMA is situated in the Bayesian predictivist framework discussed in Bernardo and Smith (2000). An excellent review of Bayesian predictive modeling can be also found in Vehtari and Ojanen (2012), and we will borrow notation from their paper.

Arguably, the overarching goal of statistics is prediction. In other words, a key characteristic of statistics is to develop accurate predictive models, and 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 (Dawid 1984). Indeed, it is hard feel confident about inferences drawn from a model that does a poor job of predicting the extant data. The problem, however, is how to develop accurate predictive models, and, importantly, how to evaluate their accuracy. The approach to developing accurate predictive models discussed in this review is BMA, and the evaluation of BMA-based analyses is best situated in the context of Bayesian decision theory.

Bayesian decision theory (see, e.g., Good 1952; Lindley 1991; Berger 2013) provides a natural and intuitive approach to evaluating Bayesian predictive models generally, and BMA in particular. Specifically, as will be expanded on below, Bayesian decision theory casts the problem of predictive evaluation in the context of maximizing the *expected utility* of a model—that is, the benefit that is accrued from using a particular model to predict future observations. The greater the expected utility, the better the model is at predictive performance in comparison to other models.

1.1. Fixing Notation and Concepts

Let $D = \{y_i, x_i\}_{i=1}^n$ be a set of data assumed to be fixed in the Bayesian sense, where y_i is an outcome of interest and x_i is a (possibly vector-valued) set of predictors. Further, let $(\tilde{y}, \tilde{x})$ be future observations of the outcome and the set of predictors, respectively. Further, let $\mathcal{M} = \{M_k\}_{k=1}^K$ represent a set of models specified to provide a prediction of the outcome $\tilde{y}$, and let M_k represent a specific chosen model.

The elements of Bayesian decision theory that we adopt in this paper have been described by Bernardo and Smith (2000) and Vehtari and Ojanen (2012) among many others. These elements consist of (a) an unknown state of the world denoted as $\omega \in \Omega$, (b) an action $a \in \mathcal{A}$, where $\mathcal{A}$ is the action space, (c) a utility function $u(a, \omega) : \mathcal{A} \times \Omega \rightarrow \mathbb{R}$ that rewards an action a when the state of the world is realized as ω, and (d) $p(\omega|D)$ representing one's current belief about the state of world conditional on observing the data, D.

To provide a context for these ideas, and in anticipation of our empirical example, consider the problem of predicting reading performance measured on 15-year-old students in the USA using data from the OECD Program on International Student Assessment (PISA) (OECD 2017). In line with Bernardo and Smith (2000), Lindley (1991), Vehtari and Ojanen (2012) and Berger (2013) and the notation given previously, (a) the states of the world correspond to the future observations of reading literacy $\tilde{y} \in \mathcal{Y}$, (b) the action $a \in \mathcal{A}$ is the actual prediction of those future observations, (c) the utility function $u(a, \tilde{y})$ defines the reward attached to the prediction, and (d) $p(\tilde{y}|D, M_*)$ is a posterior predictive distribution that encodes our belief about the future reading literacy observations conditional on the data, D, and a belief model, M_* (described later).

1.2. Utility Functions for Evaluating Predictions

The goal of predictive modeling is to optimize the utility of taking an action a. A number of utility functions exist, but common utility functions rest on the negative *quadratic loss* function

$$u(a, \tilde{y}) = -(\tilde{y} - a)^2. \tag{1}$$

The optimal action a^* is the one that maximizes the *posterior expected utility*, written as (see Clyde and Iversen 2013)

$$a^* = \arg\sup_{a \in \mathcal{A}} \int_{\Omega} u(\omega, a)\, p(\omega|D) \mathrm{d}\omega \tag{2}$$

The idea here is to take an action a that maximizes the utility u when the future observation is $\tilde{y}$. Clyde and Iversen (2013) show that the optimal decision obtains when $a^* = E(\tilde{y}|D)$, which

is the posterior predictive mean of $\tilde{y}$ given the data D. Under the assumption that the true model exists and is among the set of models under consideration, this can be expressed as

$$E(\tilde{y}|D) = \sum_{k=1}^{K} E(\tilde{y}|M_k, D)p(M_k|D) = \sum_{k=1}^{K} p(M_j|D)\hat{\tilde{y}}_{M_k} \tag{3}$$

where $\hat{\tilde{y}}_{M_k}$ is the posterior predictive mean of $\tilde{y}$ under M_k. The expression in (3) is Bayesian model averaging.

It is important to note that when considering the selection of a single model, one might be tempted to choose the model with the highest posterior model probability (PMP) $p(M_k|D)$. In the case of only two models, the model with the largest PMP will be the closest to the BMA solution. However, for more than two models, Clyde and Iversen (2013) point out that the model closest to the BMA solution might not be the one with the largest PMP.

2. Bayesian Model Averaging

In a complex real-world setting such as the one we find ourselves in when trying to develop a predictive model of reading literacy, substantive discussions would suggest that many different models could be entertained as reasonable models of reading literacy. In this case, we agree with Clyde and Iversen (2013) that to maximize our utility it would be best to average over the space of possible models via BMA.

Bayesian model averaging has had a long history of theoretical developments and practical applications. Early work by Leamer (1978) laid the foundation for BMA. Fundamental theoretical work on BMA was conducted in the mid-1990s by Madigan and his colleagues (e.g., Madigan and Raftery 1994; Raftery et al. 1997; Hoeting et al. 1999). 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. Reviews of the general problem of model uncertainty can be found in Clyde and George (2004) and more recently in Steel (2020) with a focus on economics. A review of Bayesian model averaging with a focus on psychology can be found in Hinne et al. (2020).

Practical applications and developments 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 et al. 2001b), political science (e.g., Montgomery and Nyhan 2010), bioinformatics of gene express (e.g., Yeung et al. 2005), weather forecasting (e.g., Raftery et al. 2005; Sloughter et al. 2013), propensity score analysis (Kaplan and Chen 2014), structural equation modeling (Kaplan and Lee 2015), missing data (Kaplan and Yavuz 2019), probabilistic forecasting with large-scale assessment data (Kaplan & Huang, under review).

The popularity of BMA across many different disciplines is due to the fact that BMA is known to provide better out-of-sample predictive performance than any other model under consideration as measured by the logarithmic scoring rule (Raftery and Zheng 2003). In addition, Bayesian model averaging has been implemented in the R software programs BMA (Raftery et al. 2015), BMS (Zeugner and 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 and model priors.

2.1. Statistical Specification of BMA

Following Madigan and Raftery (1994), consider a quantity of interest such as a future observation. Again, denoting this quantity as $\tilde{y}$, our goal is to obtain an optimal prediction of $\tilde{y}$ the sense that the utility of predicting $\tilde{y}$ is maximized. Next, consider a set of competing models $\mathcal{M} = \{M_k\}_{k=1}^K$ that are not necessarily nested. The posterior distribution of $\tilde{y}$ given data D can be written as a mixture distribution,

$$p(\tilde{y}|D) = \sum_{k=1}^{K} p(\tilde{y}|M_k)p(M_k|D), \tag{4}$$

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

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{l=1}^{K} p(D|M_l)p(M_l)}, \tag{5}$$

where the first term in the numerator on the right-hand side of (5) is the probability of the data given model k, also referred to as the *integrated likelihood* written as

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)\mathrm{d}\theta_k, \tag{6}$$

where $p(\theta_k|M_k)$ is the prior distribution of the parameters θ_k under model M_k (Raftery et al. 1997). The posterior model probabilities can be considered mixing weights associated with the mixture distribution given in (4) (Clyde and Iversen 2013). The second term $p(M_k)$ on the right-hand side of (5) 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 (5) ensures that $p(M_k|y)$ integrates to 1.0, as long as the true model is in the set of models under consideration. We will defer the discussion of a true model until later and then explicitly deal with the case where the true model is not in the set of models under consideration.

2.2. Connections to Bayes Factors

An important feature of (5) is that $p(M_k|y)$ captures the posterior uncertainty that a given model is the true model and this uncertainty 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, (5) could be used as a method for model selection by simply choosing the model with the largest posterior model probability. However, model uncertainty is still being ignored because, often in practice, the posterior probability of this model will not be 1.0 which is assumed if one were to select a single model and act as though that was the model the analyst had in mind all along.

Yet another common approach for model selection is the Bayes factor which provides a way to quantify the odds that the data favor one model over another (Kass and Raftery 1995). A key benefit of Bayes factors is that models do not have to be nested. To motivate the Bayes factors, consider two competing models, denoted as M_k and M_l, which 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 structural equation models 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

$$\frac{p(M_k|D)}{p(M_l|D)} = \frac{p(D|M_k)}{p(D|M_l)} \times \left[\frac{p(M_k)}{p(M_l)}\right]. \tag{7}$$

The first term on the right-hand side of (7) is the ratio of two integrated likelihoods. This ratio is referred to as the *Bayes factor* 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 $p(D|M_k)/p(D|M_l)$.

A connection between the Bayes factor in (7) and the posterior model probability in (5) has been pointed by others (see, e.g., Clyde 1999). Specifically, when examining more than two models, and assuming equal prior odds, then the Bayes factor for M_k over M_l can be written as

$$BF_{kl} = \frac{p(M_k|D)}{p(M_l|D)}. \tag{8}$$

Assuming that we fix the first model M_l as the baseline model, (5) can be re-expressed as

$$p(M_k|D) = \frac{BF_{k1}\,p(M_k)}{\sum_{l=1}^{K} BF_{l1}\,p(M_l)} \tag{9}$$

2.3. Computational Considerations

As pointed out by Hoeting et al. (1999), BMA can be difficult to implement. In particular, they note that the number of terms in (4) can be quite large, the corresponding integrals can be hard to compute, 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 (6) the Laplace method, which has been used productively for the computation of Bayes factors (Kass and Raftery 1995), can be used and this will lead to a simple BIC approximation under certain circumstances (see Tierney and Kadane 1986; Raftery 1996). The problem of reducing the overall number of models that one could incorporate in the summation of (4) has led to two interesting solutions. One solution is based on the so-called *Occam's window* criterion (Madigan and Raftery 1994) and the other is based on a Metropolis sampler referred to as *Markov chain Monte Carlo Model composition* (MC^3) (Madigan and York 1995)

2.3.1. 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. Following closely the discussion given in Raftery et al. (1997) we would initially start with very large number of predictors, but the goal would be to pare this to a smaller 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 BMA that reduces the subset of models under consideration, but instead of settling on a final "best" model, we instead integrate over the parameters of the smaller set with weights reflecting the posterior uncertainty in each model.

The algorithm proceeds as follows (Raftery et al. 1997). In the initial step, the space of possible models is initially reduced by implementing the so-called "leaps and bounds" algorithm developed by Furnival and Wilson (1974) in the context of best subsets regression (see also Raftery 1995). This initial step can substantially reduce the number of models, after which Occam's window

can then be employed. The general idea is that models are eliminated from (4) if they predict the data 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 $\mathcal{A}'$ is defined as

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

In words, (10) compares the model with the largest posterior model probability, $max_l\{p(M_l|y)\}$, to a given model $p(M_k|y)$. If the ratio in (10) is greater than the chosen value C, then it is discarded from the set $\mathcal{A}'$ of models to be included in the model averaging. Notice that the set of models contained in $\mathcal{A}'$ is based on Bayes factor values.

The set $\mathcal{A}'$ now contains models to be considered for model averaging. In the second, optional, step, models are discarded from $\mathcal{A}'$ if they receive less support from the data than simpler sub-models. Formally, models are further excluded from (4) if they belong to the set

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

Again, in words (11) states that there exists a model M_l within the set $\mathcal{A}'$ and where M_l is simpler than M_k. If a complex model receives less support from the data than a simpler sub-model—again based on the Bayes factor—then it is excluded from $\mathcal{B}$. Notice that the second step corresponds to the principle of Occam's razor (Madigan and Raftery 1994).

With step 1 and step 2, the problem of reducing the size of the model space for BMA is simplified by replacing (4) with

$$p(\tilde{y}|y, \mathcal{A}) = \sum_{M_k \in \mathcal{A}} p(\tilde{y}|M_k, y) p(M_k|y, \mathcal{A}), \tag{12}$$

In other words, models under consideration for BMA are those that are in $\mathcal{A}'$ but not in $\mathcal{B}$. Formally, $\mathcal{A} = \mathcal{A}' \setminus \mathcal{B}$.

Madigan and Raftery (1994) 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 posterior odds, if the odds are positive, indicating support for M_0, then we reject M_1. If the posterior odds is large and negative, then we reject M_0 in favor of M_1. Finally, if the posterior odds lies in between the pre-set criterion, then both models are retained. For linear regression models, the leaps and bounds algorithm combined with Occam's window is available in the BICREG option in the R program BMA (Raftery et al. 2015).

2.3.2. Markov Chain Monte Carlo Model Composition Markov chain Monte Carlo model composition (MC^3) is based on the Metropolis–Hastings algorithm (see, e.g., Gilks et al. 1996) and is also designed to reduce the space of possible models that can be explored via Bayesian model averaging. Following Hoeting et al. (1999), the MC^3 algorithm proceeds as follows. First, let $\mathcal{M}$ 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 zero 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

$$\min\left\{1, \frac{pr(M'|y)}{pr(M|y)}\right\}, \tag{13}$$

otherwise, the chain stays in model M.

2.4. Model and Parameter Priors

One of the key steps when implementing BMA is to choose priors for both the parameters of the model and the model space itself. Discussions of the choice of parameter and model priors can be found in Fernández et al. (2001a); Liang et al. (2008); Eicher et al. (2011); Feldkircher and Zeugner (2009), with applications found in Fernández et al. (2001a) and Kaplan and Huang (under review). A large number of choices for model and parameter priors are implemented in the R software program BMS (Zeugner and Feldkircher 2015). This section discusses the extant choices of parameter and model priors, following closely the discussion given in Kaplan and Huang (under review).

2.4.1. Parameter Priors The choice of parameter priors available in the extant BMA software rests on variations of Zellner's g-prior (Zellner 1986). Specifically, Zellner introduced a natural-conjugate Normal-Gamma g-prior for regression coefficients β under the normal linear regression model, written as,

$$y_i = x_i'\beta + \varepsilon, \tag{14}$$

where ε is i.i.d. $N(0, \sigma^2)$. For a give model, say M_k, Zellner's g-prior can be written as

$$\beta_k|\sigma^2, M_k, g \sim N\left(0, \sigma^2 g\left(x_k'x_k\right)^{-1}\right). \tag{15}$$

Feldkircher and Zeugner (2009) have argued for using the g-prior for two reasons: its consistency in asymptotically uncovering the true model, and its role as a penalty term for model size.

The g-prior has been the subject of some criticism. In particular, Feldkircher and Zeugner (2009) have pointed out that the particular choice of g can have a very large impact on posterior inferences drawn from BMA. In particular, small values of g can yield a posterior mass that is spread out across many models while large values of g can yield a posterior mass that is concentrated on fewer models. Feldkircher and Zeugner (2009) use the term *supermodel effect* to describe how values of g impact the posterior statistics including posterior model probabilities (PMPs) and posterior inclusion probabilities (PIPs).

To account for the supermodel effect researchers (Fernández et al. 2001a; Liang et al. 2008; Eicher et al. 2011; Feldkircher and Zeugner 2009) have proposed alternative priors based on extensions of the work of Zellner (1986). Generally speaking, these alternatives can be divided into two categories: *fixed priors* and *flexible priors*. Fernández et al. (2001a) recommended using

benchmark priors which belong to the class of fixed priors when sample sizes are large. Liang et al. (2008) introduced mixtures of g-priors to address the inconsistency when using fixed priors and showed its advantages compared to other default priors. Instead of only employing model-dependent priors, Feldkircher and Zeugner (2009) proposed a hyper-g-prior that "let the data choose," thus reducing the sensitivity of the prior choice of the g-prior on the posterior mass. In a detailed study, Eicher et al. (2011) compared twelve candidate default priors and concluded that the unit information prior (UIP) combined with the uniform model prior outperformed the other choices.

2.4.2. Fixed Parameter Priors The set of fixed priors that are available in BMS are (see Zeugner and Feldkircher 2015):

- *Unit Information Prior*: $g = N$. This is a typical default prior. Liang et al. (2008) suggested using UIP in combination with the uniform model prior to yield the best predictive performance.
- *Risk Inflation Criterion Prior* (RIC): $g = Q^2$, where Q is the number of predictors. Foster and George (1994) showed that the selection of the model with the highest PMP is equivalent to selecting the model with the highest RIC as long as $g = Q^2$.
- *Benchmark risk inflation criterion* (BRIC): $g = \max(N, Q^2)$. This is a combination of the UIP and RIC. When $N \leq Q^2$, Fernández et al. (2001a) recommend using $g = Q^2$; When $N > Q^2$, use $g = N$ in the variable selection context.
- *Hannan and Quinn Priors*: $g = \log(N)^3$: This prior is based on the Hannan–Quinn criterion for model selection. Hannan and Quinn (1979) advocated to use HQ criteria = 3 for large N.

2.4.3. Flexible Parameter Priors The set of flexible priors are (see Zeugner & Feldkircher, 2015):

- *Local Empirical Bayes*: $g_k = \arg\max(0, F_k - 1)$, where $F_k = \frac{R_k^2(N-Q_k-1)}{(1-R_k^2)Q_k}$; F_k is the F-statistic and R_k^2 is the regression coefficient of determination for model M_k. This approach estimates g separately for each model with maximum likelihood methods based on the observed data (George and Foster 2000; Liang et al. 2008; Hansen and Yu 2001).
- *Hyper-g priors*: This family of priors was proposed for data-dependent shrinkage. Following Feldkircher and Zeugner (2009), the hyper-g prior is a Beta prior on the shrinkage factor $\frac{g}{1+g}$, that is $p\left(\frac{g}{1+g}\right) \sim Beta(1, \frac{\alpha}{2} - 1)$, with $E\left(\frac{g}{1+g}\right) = \frac{2}{\alpha}$. Instead of eliciting g directly, the hyper-g prior requires the elicitation of the hyperparameter $\alpha \in (2, \infty)$. As α approaches 2, the prior distribution on the shrinkage factor $\frac{g}{1+g}$ will be close to 1; while for $\alpha = 4$, the prior distribution on the shrinkage factor will be uniform distributed. In the context of noisy data, the hyper-g prior will distribute posterior model probabilities more uniformly across the model space. In the case of low noise in the data, the hyper-g prior will be concentrated on a few models, and perhaps even more concentrated than in the fixed prior case with large g (Feldkircher and Zeugner 2009).

2.4.4. Model Priors Here we discuss three model priors that are available in the BMS program: (a) the uniform model prior, (b) the binomial model prior, and (c) the Beta-binomial model prior.

- *Uniform model prior*: The uniform model prior is a common default prior for Bayesian model averaging. Specifically, if there are Q predictors, then the prior on the space of models is 2^{-Q}. The difficulty with the uniform model prior was pointed out by Zeugner and Feldkircher (2015) who noted that the uniform model prior implies that the expected

model size is $\sum_{q=0}^{Q}\binom{Q}{q}q2^{-Q} = Q/2$. However, the distribution of model sizes is not even—there are more models of size 2 or 5, then there are of size 1 or 6. The result is that the uniform model prior actually places more mass on intermediate size models. A demonstration of the impact of this problem is given in Zeugner and Feldkircher (2015).

- *Binomial model prior*: To address the problem with the uniform model prior, Zeugner and Feldkircher (2015) proposed placing a fixed inclusion probability on each predictor in the model, denoted as θ. Then, for model k, the prior probability for a model of size q is $p(M_k) = \theta^{q_k}(1-\theta)^{Q-q_m}$. Notice that the expected model size, say $\bar{m}$, is $Q\theta$, and thus the analysts prior expected model size is $\bar{m}$. Moreover, if $\theta = .5$, then the binomial model prior reduces to the uniform model prior. In practice, this suggests that choosing $\theta < .5$ weights the posterior mass toward smaller models, and visa versa (Zeugner and Feldkircher 2015).
- *Beta-binomial model prior*: The binomial prior discussed above suffers from the fact that the inclusion probability θ is fixed. Following Ley and Steel (2009), greater flexibility is provided by treating θ as random. A logical choice for the probability distribution of θ is the Beta distribution with hyperparameters $a, b > 0$, viz., $\theta \sim \text{Beta}(a, b)$. Under the Beta-binomial prior the first and second moments of the model size $\bar{m}$ are,

$$E(\bar{m}) = \frac{a}{a+b}Q, \tag{16}$$

$$\text{var}(\bar{m}) = \frac{ab(a+b+Q)}{(a+b)^2(a+b+1)}Q. \tag{17}$$

3. Evaluating Bayesian Model Averaging

With such a wide variety of choices available for parameter and model priors, it is important to have a method for evaluating the impact of these choices when applying BMA to substantive problems. Given that the utility of BMA lies in its optimal predictive performance, a reasonable method for evaluation should be based on measures that assess predictive performance—referred to as *scoring rules*.

3.1. Strictly Proper Scoring Rules

Scoring rules provide a measure of the accuracy of probabilistic predictions (or synonymously, forecasts), and a prediction can be said to be "well-calibrated" if the assigned probability of the outcome match the actual proportion of times that the outcome occurred (Dawid 1982).[1]

A scoring rule is a utility function (Gneiting and Raftery 2007), and the goal of the forecaster is to be honest and provide a forecast that will maximize the utility. One can consider scoring rules from a subjectivist Bayesian perspective. Here, Winkler (1996) quotes Finetti (1962, pg. 359)

> The scoring rule is constructed according to the basic idea that the resulting device should oblige each participant to express his true feelings, because any departure from his own personal probability results in a diminution of his own average score as he sees it.

Because scoring rules only require the stated probabilities and realized outcomes, they can be developed for ex-post or ex-ante probability evaluations. However, as suggested by Winkler (1996), the ex-ante perspective of probability evaluation should lead us to consider *strictly proper*

[1] Scoring rules should be distinguished from scoring functions. The former describe rules for predictive distributions while the later describe rules for point predictions, and include variants of squared error.

scoring rules because these rules are maximized if and only if the forecaster is honest in reporting their scores.

Following the discussion and notation given in Winkler (1996, see also; Jose et al. 2008), let $\mathbf{p}$ represent the forecaster's subjective probability distribution of an outcome of interest, let $\mathbf{r}$ represent the forecaster's reported forecast probability, and let $\mathbf{e}_i$ represent the probability distribution that assigns probability one if the event i occurs and probability zero for all other events. Then, a scoring rule, denoted as $S(\mathbf{r}, \mathbf{p})$, provides a score $S(\mathbf{r}, \mathbf{e}_i)$ if the event occurs. The expected score obtained when the forecaster reports $\mathbf{r}$ when their true distribution is $\mathbf{p}$ is

$$S(\mathbf{r}, \mathbf{p}) = \sum_i p_i S(\mathbf{r}, \mathbf{e}_i) \tag{18}$$

The scoring rule is strictly proper if $S(\mathbf{p}, \mathbf{p}) \geq S(\mathbf{r}, \mathbf{p})$ for every $\mathbf{r}$ and $\mathbf{p}$ with equality when $\mathbf{r} = \mathbf{p}$ (Jose et al. 2008, pg. 1147).

A large number of scoring rules have been reviewed in the literature (see, e.g., Winkler 1996; Bernardo and Smith 2000; Jose et al. 2008; Merkle and Steyvers 2013; Gneiting and Raftery 2007). Here, however, we highlight two related strictly proper scoring rules that are commonly used to evaluate predictions arising from Bayesian model averaging: the log predictive density score, and the Kullback–Leibler divergence score (see, e.g., Fernández et al. 2001b; Hoeting et al. 1999; Kaplan and Yavuz 2019; Kaplan and Huang, under review, for examples).

3.1.1. The Log Predictive Density Score The log predictive density score (LPS) (Good 1952; Bernardo and Smith 2000) can be written as

$$-\sum_i \log\left[p(\tilde{y}_i|x, y, \tilde{x}_i)\right] \tag{19}$$

where, for example, $\tilde{y}_i$ is the predictive density for ith person, x and y represent the model information for the remaining individuals, and $\tilde{x}_i$ is the information on the predictors for individual i. The model with the lowest log predictive score is deemed best in terms of long-run predictive performance.

3.1.2. Kullback–Leibler Divergence Score Closely related to the log-predictive score is the Kullback–Leibler Divergence (KLD) score (also referred to as *relative entropy* (Kullback and Leibler 1951; Kullback 1959, 1987). Here we consider two distributions, $p(y)$ and $g(y|\theta)$, where $p(y)$ could be the distribution of observed reading literacy scores, and $g(y|\theta)$ could be the prediction of these reading scores based on a model. The KLD between these two distributions can be written as

$$\text{KLD}(f, g) = \int p(y)\log\left(\frac{p(y)}{g(y|\theta)}\right)\text{d}y \tag{20}$$

where $\text{KLD}(f, g)$ is the information lost when $g(y|\theta)$ is used to approximate $p(y)$. For example, the actual reading outcome scores might be compared to the predicted outcome using Bayesian model averaging along with different choices of model and parameter priors. The model with the lowest KLD measure is deemed best in the sense that the information lost when approximating the actual reading outcome distribution with the distribution predicted on the basis of the model is lowest.

The LPS and KLD scoring rules are applicable to continuous outcomes and we will focus on these two in our example below. However, it should be noted that BMA can be applied to binary

outcomes, and here, a popular scoring rule that is based on quadratic loss is the Brier score (Brier 1950). Following the notation above, the Brier score can be defined as

$$-(\|\mathbf{e}_i - \mathbf{p}\|_2)^2 \tag{21}$$

where the forecast $\mathbf{p}$ estimates an indicator vector of an event $\mathbf{e}$. For example, $\mathbf{p}$ may represent the forecast probability of rain on a given day, and $\mathbf{e}_i$ represents the realization of the event scored 1/0 should rain occur on that day or not. The Brier score penalizes the forecaster in proportion to the squared Euclidean distance between the forecast and the event (Jose et al. 2008, p. 1148)

4. A Simple Example of BMA

This example will draw on reading literacy data obtained from the Program for International Student Assessment (PISA). Launched in 2000 by the Organization for Economic Cooperation and Development, PISA is a triennial international survey that aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. In 2018, 600,000 students, statistically representative of 32 million 15-year-old students in 79 countries and economies, took an internationally agreed-upon two-hour test. Students were assessed in science, mathematics, reading, collaborative problem solving, and financial literacy. PISA is arguably the most important policy-relevant international survey that is currently operating (OECD 2002).

The benefit of developing optimally predictive models for large-scale assessments such as PISA is to recognize that these assessments can be used to monitor progress toward internationally agreed-upon educational goals such as the United Nations (UN) adopted the Sustainable Development Goals (SDGs). Also, as educational systems around the world face new challenges due to the COVID-19 pandemic, developing optimally predictive models may help identify the long-run impact of this unprecedented health crisis on global education.

Following the overview in Kaplan and Kuger (2016), the sampling framework for PISA follows a two-stage stratified sample design. Each country/economy provides a list of all "PISA-eligible" schools, and this list constitutes the sampling frame. Schools are then sampled from this frame with sampling probabilities that are proportional to the size of the school, with the size being a function of the estimated number of PISA-eligible students in the school. The second stage of the design requires sampling students within the sampled schools. A target cluster size of 35 students within schools was desired, though for some countries, this target cluster size was negotiable.

We will focus on the reading literacy results from PISA 2018. The method of assessment for PISA follows closely the spiraling design and plausible value methodologies originally developed for National Assessment of Educational Progress. (see, e.g., OECD 2017) In addition to these so-called "cognitive outcomes," policymakers and researchers alike have begun to focus increasing attention on the non-academic contextual aspects of schooling. Context questionnaires provide important variables for models predicting cognitive outcomes and these variables have become important outcomes in their own right—often referred to as "non-cognitive outcomes" (see, e.g., Heckman and Kautz 2012). PISA also assesses these non-cognitive outcomes via a one-half hour internationally agreed-upon context questionnaire (see Kuger et al. 2016). The list of variables used in this example are given in Table 1.

For this example, we use the software package BMS (Zeugner and Feldkircher 2015) which implements the so-called *Birth/Death* algorithm as a default for conducting MC^3. It is beyond the scope of this paper to describe the BD algorithm or other choices of algorithms in the BMS program. See Zeugner and Feldkircher (2015) for more detail.

The analysis steps for this example are as follows:

TABLE 1.
PISA 2018 predictors of reading literacy.

Variable name	Variable label
FEMALE	Sex (1 = Female)
ESCS	Index of economic, social, and cultural status
METASUM	Meta-cognition: summarising
PERFEED	Perceived feedback
HOMEPOS	Home possessions (WLE)[a]
ADAPTIVE	Adaptive instruction (WLE)
TEACHINT	Perceived teacher's interest
ICTRES	ICT resources (WLE)
JOYREAD	Joy/Like reading
ATTLNACT	Attitude towards school: learning activities
COMPETE	Competitiveness
WORKMAST	Work mastery
GFOFAIL	General fear of failure
SWBP	Subjective well-being: positive affect
MASTGOAL	Mastery goal orientation
BELONG	Subjective well-being: sense of belonging to school (WLE)
SCREADCOMP	Perception of reading competence
SCREADDIFF	Perception of reading difficulty (WLE)
PISADIFF	Perception of difficulty of the PISA test
PV1READ	First plausible value reading score (outcome variable)[b]

[a]Weighted likelihood estimates. See OECD (2018) for more details.
[b]Plausible values. See OECD (2018) for more details.

1. We begin by implementing BMA with default unit information priors for the model parameters and the uniform prior on the model space. We will outline the major components of the results including the posterior model probabilities and the posterior inclusion probabilities.
2. We next examine the results under different combinations of parameter and model priors available in BMS and compare results using the LPS and KLD.

4.1. BMA Results

The Bayesian model averaging results under unit information priors for model parameters and the uniform prior for the model space are shown in Tables 2 and 3. It should be noted that the results under different choices of parameter and model priors demonstrate some sensitivity to the choice of parameter and model priors and these results are available on the author's website (http://bmer.wceruw.org/index.html). We note that there are 19 predictors and thus $2^{19} = 524288$ models in the full space of models to be visited. Table 2 presents a summary of the birth/death algorithm used to implement MC^3 in BMS. We find that the algorithm only visited 471 models (0.09%) out of the total model space; however, these models accounted for 100% of the posterior model mass.[2] The column labeled "Avg # predictors" shows that across all of the models explored by the algorithm, the average number of predictors was 11.8 out of 19.

In the second row of Table 2 we present the posterior model probabilities (PMPs) associated with the top five models out of the 471 models explored by the algorithm. It is important to note

[2]This percentage is obtained by summing over the PMPs for all models explored by the algorithm and dividing by the total number of those models.

TABLE 2.
Summary of birth/death algorithm and top posterior model probabilities.

Algorithm summary	Modelspace 2^K	Models visited	% visited	% Topmodels	Avg. # predictors
	524288	471	0.09	100	11.8
PMPs for top five models	Model 1	Model 2	Model 3	Model 4	Model 5
	0.35	0.20	0.06	0.04	0.03

TABLE 3.
Summary of BMA with unit information parameter priors and uniform model priors.

Predictor	PIP	Post. Coef.	Post. SD	Cond. Pos. Sign
ESCS	1.00	18.97	1.30	1.00
METASUM	1.00	27.99	1.29	1.00
TEACHINT	1.00	12.53	1.51	1.00
JOYREAD	1.00	10.33	1.32	1.00
GFOFAIL	1.00	11.06	1.21	1.00
MASTGOAL	1.00	− 13.34	1.50	0.00
SCREADCOMP	1.00	10.13	1.49	1.00
PISADIFF	1.00	− 29.71	1.46	0.00
PERFEED	0.98	− 5.00	1.55	0.00
SWBP	0.87	− 3.95	1.92	0.00
WORKMAST	0.86	4.28	2.16	1.00
FEMALE	0.64	5.14	4.37	1.00
ADAPTIVE	0.12	0.45	1.31	1.00
SCREADDIFF	0.08	− 0.19	0.78	0.00
COMPETE	0.07	0.19	0.79	1.00
BELONG	0.05	− 0.15	0.72	0.00
HOMEPOS	0.02	0.02	0.29	1.00
ICTRES	0.01	− 0.02	0.23	0.00
ATTLNACT	0.00	0.00	0.09	1.00

that Model 1 would also be associated with the lowest Bayesian information criterion. Hence, on the basis of the low PMP for Model 1 (0.35) we can see that selecting Model 1 and acting as though this is the model we considered ahead of time, considerably underestimates the uncertainty in our model choice. Moreover, as Clyde and Iversen (2013) remind us, this model might not be the one closest to the BMA solution.

The results of the BMA are shown in Table 3. The column labeled "PIP" shows the posterior inclusion probabilities for each variable, referring to the proportion of times the variable appeared in the models visited by the algorithm. For example, the PIP for ESCS is 1.00, meaning that across all the models selected by the algorithm, ESCS appeared in 100% of the models. In contrast, ATTLNACT only appears in 0.09% of the models visited by the algorithm. The PIP thus provides a different perspective on variable importance. The columns labeled "Post Mean" and "Post SD" are the posterior estimates of the regression coefficients and their posterior standard deviations, respectively. The column labeled "Cond. Pos. Sign" refers to the probability that the sign of the respective regression coefficient is positive conditional on its inclusion in the model. We find, for example, that the sign of ESCS is positive in 100% of the models in which ESCS appears. By

contrast, the probability that the sign of the PISADIFF effect positive is zero, meaning that in 100% of the models visited by the algorithm, the sign of PISADIFF is negative.[3]

We find that the first 12 predictors (ESCS thru GENDER) have relatively high PIPs. The majority of these predictors have PIPs of 1.0 indicating their importance. It is also interesting to note that these predictors contain a mix of demographic measures (e.g., ESCS, GENDER), attitudes/perceptions (e.g., TEACHINT, JOYREAD, SCREADCOMP) and cognitive strategies involved in reading (e.g., METASUM).

4.2. Comparison of Parameter and Model Priors

Table 4 presents the results based on comparing LPS and KLD values for different parameter priors under the fixed prior setting (upper half) and flexible prior setting (lower half), respectively. Owing to the large-sample size, the findings show relative robustness to the choice of parameter and model priors.

5. True Models, Belief Models, and $\mathcal{M}$-Frameworks

Earlier, our discussion made mention of belief models and true models, and a perusal of the extant literature on model averaging reveals a distinction between a so-called *actual belief model* (sometimes referred to as a *Bayesian belief model*), M_*, and a true model, denoted as M_T. Unfortunately, there does not appear to be a consensus about the meaning of belief models or true models, and in some cases, they are viewed as roughly the same thing. For example, Bernardo and Smith (2000) introduced the idea of the actual belief model but seem to be describing it as the true model, and in fact labeled the actual belief model as M_T, suggesting that M_* is the true but unknown data generating model This position appears to be held by Clyde and Iversen (2013) who adopt a similar notation and seem to use the terms "belief model" and "true model" interchangeably.

In contrast, Vehtari and Ojanen (2012) suggest that M_* is different from M_T and is something that we have access to insofar as it derived from what we have learned from our encounter with data. For our example, M_* would be the result of what has been learned from the construction and criticism of a substantive model of reading literacy. That is, if (a) one has specified a reasonable probability model for the reading literacy outcome, (b) one has access to a rich enough set of policy-relevant predictors of reading literacy, (c) all of the important prior uncertainties have been captured, and (d) the model has withstood criticisms, say in the form of posterior predictive checks (Gelman et al. 1996), then this model is M_*. Regarding M_T, Vehtari and Ojanen (2012, p. 155) suggest that "the properties of the true model are specified by the modeller a priori, and they are not learned from the data properties." Vehtari and Ojanen (2012) view the specification of M_T in very general terms, such as an $i.i.d$ assumption regarding the outcome of interest.

With respect to model averaging, the distinction between M_* and M_T is important in practice. First, BMA assumes that $M_T \in \mathcal{M}$. If that assumption does not hold, then conventional BMA does not make sense because the priors on the model space are elicited to reflect the analyst's belief about the existence of the true model within the full set of models under consideration. Second, as noted by Vehtari and Ojanen (2012), the process of model averaging begins with assuming the existence of a true model that must be approximated from the data, and this approximation is often based on an actual belief model that the researcher implicitly holds. Moreover, when using BMA software such as BMA or BMS, an intital model must be specified to initiate the search through the model space. The question that arises is whether this initial model is M_* or M_T. It

[3]The probabilities listed under the Cond. Pos. Sign results will often range from zero to one, but for these results, it appears that all 471 models show clarity with respect to the sign of the posterior coefficients.

TABLE 4.
KLD and LPS values for fixed and flexible priors priors.

	KLD				LPS			
	UIP	RIC	BRIC	HQ	UIP	RIC	BRIC	HQ
Uniform	0.015	0.015	0.015	0.015	5.843	5.842	5.843	5.842
Binomial ($m = 2$)	0.015	0.015	0.015	0.015	5.844	5.844	5.844	5.844
Binomial ($m = 4$)	0.015	0.015	0.015	0.015	5.844	5.843	5.844	5.843
Beta-binomial	0.015	0.015	0.015	0.015	5.843	5.842	5.843	5.842

	KLD						LPS					
	EBL	HQ-3	HQ-4	HG-UIP	HG-RIC	HG-BRIC	EBL	HQ-3	HQ-4	HG-UIP	HG-RIC	HG-BRIC
Uniform	0.015	0.015	0.015	0.015	0.015	0.015	5.841	5.841	5.841	5.841	5.841	5.841
Binomial ($m = 2$)	0.015	0.015	0.015	0.015	0.015	0.015	5.844	5.844	5.844	5.844	5.844	5.844
Binomial ($m = 4$)	0.015	0.015	0.015	0.015	0.015	0.015	5.843	5.843	5.843	5.843	5.843	5.843
Beta-binomial	0.015	0.015	0.015	0.015	0.015	0.015	5.841	5.841	5.841	5.841	5.841	5.841

These results are based on a sample size of 2500 respondents due to a sample size limitation in the BMS software.

does not seem reasonable that this model is M_T, if by M_T we mean some general probability model for the outcome, not conditional on any covariates. Rather, this initial model is much more akin to what is meant by M_*, having perhaps resulted from giving careful consideration to an initial model based on past research, expert opinion, and so forth.

To clarify terminology, we prefer the following set of distinctions. First, there is a true model M_T that we do not fully know, except in the case of computer simulation studies (Clarke and Clarke 2018), and it may be a highly complex process associated with many covariates in perhaps complicated non-linear and structural relationships. Second, M_* is our best, or most convenient, approximation to M_T and forms the empirical launching point for model averaging. Finally, the goal of model averaging is to start with M_* and locate a model M_k that is as close to M_T as possible with "closeness" defined in terms of an index such as the Kullback–Leibler divergence (Kullback and Leibler 1951; Kullback 1959, 1987).

Our discussion of belief models and true models is necessary in order to address a critical problem when conducting BMA in practice—namely whether it is reasonable to assume that $M_T \in \mathcal{M}$. If we assume that M_T is in the space of models under consideration, this is referred to as the $\mathcal{M}$-*closed* framework, introduced by Bernardo and Smith (2000) and further discussed in Clyde and Iversen (2013).

As implied earlier, the $\mathcal{M}$-*closed* framework for BMA may be especially difficult to warrant in the social and behavioral sciences. Nevertheless, as pointed out by Bernardo and Smith (2000), there may be cases in which it is reasonable to act as though there is a true model, keeping in mind that Bernardo and Smith (2000) seem to suggest that M_T is what we are referring to here as M_*. For example, we may wish to act as though M-closed holds when a model has demonstrated good predictive capabilities under a wide variety of situations, but that under a new situation, new uncertainties arise. Taking our example of the prediction of reading literacy, an analyst typically would specify a set of variables that cover the range of what theory and past analyses have suggested are important predictors of reading competencies—predictors such (a) as measures of socio-demographics, (b) measures of teacher practices in support of reading literacy, (c) perceptions of classroom and school environments, including instructional resources, and (d) student attitudes and dispositions toward reading. In this example, Bayesian analyses from prior relevant studies such as PISA 2009 (the last reading cycle of PISA) (OECD 2009), might serve to provide informative priors for the analysis of reading data from PISA 2018. Given policy and research papers that derive from the analyses of PISA 2018, it may be reasonable to specify an initial belief model, M_*. However, now the analyst might recognize that there is uncertainty in that choice of M_* and wish to address that uncertainty using BMA to optimize the prediction of reading competencies for future cycles of PISA. Such uncertainties may be due to applying a model estimated on one PISA country to another. Or, changes in educational policies and practices due to the COVID-19 pandemic may render much greater uncertainty to a model that may have worked well in the past. As long as the analyst is comfortable assigning model priors, then the $\mathcal{M}$-*closed* framework can be adopted. Nevertheless, the truth or falsity of the $\mathcal{M}$-*closed* framework notwithstanding, it is important to reiterate that conventional BMA takes place under the $\mathcal{M}$-*closed* framework and, indeed, readily available BMA software typically employ a non-informative prior to the space of models as a default, with the idea that the true model lies in the model space.

5.1. Model Averaging in the $\mathcal{M}$-Complete Framework

With the M-closed assumption unlikely to hold in practice, we are faced with the problem of how to obtain the benefits of model averaging with respect to predictive accuracy. One approach would be to create a list of simpler "proxy" linear models, $\{M_k\}_{k=1}^K$ specified for clarity of communication and ease of analysis (Bernardo and Smith 2000). Each of these models would be

evaluated in light of the true model. This is referred to as the $\mathcal{M}$-$complete$ framework (Bernardo and Smith 2000). Under $\mathcal{M}-complete$, BMA would not, in principle, be conducted as it does not make sense to place a discrete prior on the model space when one does not believe that $M_T \in \mathcal{M}$. Instead, as suggested by Clyde and Iversen (2013), Yao et al. (2018), and Vehtari and Ojanen (2012) one simply selects the model $M_k \in \mathcal{M}$ that maximizes expected utility with respect to predictive distributions. However, this suggests that a single model is being used for predictive purposes with the result that model uncertainty is again not being addressed.

More recently, Clarke and Clarke (2018) discussed the idea that $\mathcal{M}$-$complete$ could constitute a range of inaccessibility to M_T, and that methods such as BMA could be justified under $\mathcal{M} - complete$ insofar as the model priors would encode ones belief as to how good an approximation a given model is to M_T under an $\mathcal{M} - closed$ problem. In any event, modeling under the $\mathcal{M} - complete$ framework does not provide an approach to directly addressing the problem of model uncertainty, when $\mathcal{M} - closed$ is hard to maintain.

5.2. *Model Averaging in the $\mathcal{M}$-Open Framework*

If it is difficult to warrant model priors as required under M-closed, and if selecting a single model under M-complete that maximizes expected utility is not satisfactory, then we need an approach that allows for model averaging without the need to assume that $M_T \in \mathcal{M}$. This is referred to as the $\mathcal{M}$-$open$ framework (Bernardo and Smith 2000). An example of an $\mathcal{M}$-$open$ problem is in specifying a set of regression models with different choices of predictors. These different regression models would represent reasonable alternative belief models. Then, rather than using posterior model probabilities as weights, each model would yield a separate score without presuming the existence of a true model underlying any of the separate models. These models would be combined using their scores as weights, and the resulting predictive distribution would be obtained. This type of model averaging in the M-open framework describes the methodology of *Bayesian stacking* which we consider next.

6. Bayesian Stacking

The method of *stacking* was originally developed in the machine learning literature by Wolpert (1992) and Breiman (1996) and brought into the Bayesian paradigm by Iversen (2013). The basic idea behind stacking is to enumerate a set of K $(k = 1, 2, \ldots K)$ models and then create a weighted combination of their predictions. Returning to our reading literacy example, we can specify a set of candidate (belief) models of reading literacy as

$$y = f_k(x) + \epsilon \tag{22}$$

were f_k are different models of the reading literacy outcome— e.g., some models may include only demographic predictors, others may include various combinations of attitudes and behaviors related to reading literacy. Predictions from these models are then combined (stacked) as (see Le and Clarke 2017)

$$\tilde{y} = \sum_{k=1}^{K} \hat{w}_k \hat{f}_k(x), \tag{23}$$

where $\hat{f}_k$ estimates f_k. The weights, $\hat{w}_k$ $(\hat{w}_1, \hat{w}_2, \ldots \hat{w}_K)$ are obtained as

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} w_k \hat{f}_{k,-i}(x_i) \right)^2 \tag{24}$$

where $\hat{f}_{k,-i}(x_i)$ is an estimate of f_k based on $n-1$ observations, leaving the ith observation out.

6.1. Leave-One-Out Cross-Validation

We see from (24) that a method is needed to estimate f_k based on $n-1$ observations leaving the i^{th} observation out, and the most common approach is referred to as *leave-one-out cross validation*. Leave-one-out-cross-validation (LOOCV) is a special case of k-fold cross-validation (k-fold CV) when $k = n$. In k-fold CV, a sample is split into k groups (folds) and each fold is taken to be the validation set with the remaining $k-1$ folds serving as the training set. For LOOCV, each observation serves as the validation set with the remaining $n-1$ observations serving as the training set. Leave-one-out cross-validation is available in the R software program loo (Vehtari et al. 2019).[4]

Following Vehtari et al. (2017), let y_i $(i = 1, \ldots, n)$ be an n-dimensional vector of data following a distribution conditional on parameters θ - viz., $p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$. Given a prior distribution on the parameters, $p(\theta)$, we can obtain the posterior distribution, $p(\theta|y)$ as well as a posterior predictive distribution of predicted values $\tilde{y}$ written as $p(\tilde{y}|y) = \int p(\tilde{y}_i|\theta)p(\theta|y)\mathrm{d}\theta$. The Bayesian LOOCV rests on the derivation of the *expected log point-wise predictive density* (elpd) for new data defined as

$$\text{elpd} = \sum_{i=1}^{n} \int p_t(\tilde{y}_i) \log\, p(\tilde{y}_i|y)\mathrm{d}\tilde{y}_i, \tag{25}$$

where $p_t(\tilde{y}_i)$ represents the distribution of the true but unknown data-generating process for the predicted values $\tilde{y}_i$ and where (25) is approximated by cross-validation procedures. The elpd provides a measure of predictive accuracy for the n data points taken one at a time (Vehtari et al. 2017). From here, the Bayesian LOO estimate can be written as

$$\text{elpd}_{loo} = \sum_{i=1}^{n} \log\, p(y_i|y_{-i}), \tag{26}$$

where

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})\mathrm{d}\theta, \tag{27}$$

which is the leave-one-out predictive distribution using the log predictive score to assess predictive accuracy.

It is useful to note that an information criterion based on LOO (*LOOIC*) can be easily derived as

$$\text{LOOIC} = -2\widehat{\text{elpd}}_{loo} \tag{28}$$

which places the LOOIC on the "deviance scale" (see Vehtari et al. 2017 for more details on the implementation of the LOOIC in loo). Among a set of competing models, the one with the smallest LOOIC is considered best from an out-of-sample point-wise predictive point of view.

As pointed out by Vehtari et al. (2017), it can be time-consuming to calculate exact LOOCV values and this may be a reason why LOOCV is not widely adopted. To remedy this, Vehtari et al. (2017) developed a fast and stable approach to obtaining LOOCV referred to as *Pareto-smoothed importance sampling* (PSIS-LOO) (see Vehtari et al. 2017, for more details). The PSIS approach is implemented in loo (Vehtari et al. 2019).

[4]The *widely applicable information criterion* (WAIC) has also been advocated for model selection. Although the WAIC and LOOCV are asymptotically equivalent (Watanabe 2010), the implementation of LOOCV in the loo package is more robust in finite samples with weak priors or influential observations (Vehtari et al. 2017).

6.2. Other Types of Weights

It is interesting to note that LOOCV has connections to other types of weights that can be used for stacking. For example, in the case of maximum likelihood estimation, LOOCV weights are asymptotically equivalent to Akaike information criterion (AIC) weights (Akaike 1973) that are used in frequentist model averaging applications (Yao et al. 2018, see also; Burnham and Anderson 2002; Fletcher 2018). In addition, so-called *pseudo-BMA* weights were proposed by Geisser and Eddy (1979, see also; Gelfand 1996). This approach replaces marginal likelihoods with Bayesian LOOCV predictive densites. The difficulty with pseudo-BMA weights is that they do not take into account uncertainty in future data distributions. To address this Yao et al. (2018) proposed an approach that combines the Bayesian bootstrap (see Rubin 1981) with the ELPD defined earlier. They refer to this approach as *pseudo-BMA+* and show that it performs better than BMA and pseudo-BMA in M-open settings, but not as well as stacking using the log-score.

7. An Example of Bayesian Stacking

For this paper, we demonstrate Bayesian stacking using the software program loo with the same PISA 2018 data set used to demonstrate BMA. The analysis steps for this demonstration are as follows:

1. Specify four models of reading literacy. From Table 1, Model 1 includes only demographic measures (FEMALE, ESCS, HOMEPOS, ICTRES); Model 2 includes only attitudes and behaviors specifically directed toward reading (JOYREAD, PISADIFF, SCREADCOMP, SCREADDIF); Model 3 includes predictors related to academic mindset as well as general well-being; (METASUM, GFOFAIL, MASTGOAL, SWBP, WORKMAST, ADAPTIVITY, COMPETE); and Model 4 includes attitudes toward school (PERFEED, TEACHINT, BELONG).
2. Obtain results from log-score stacking weights, pseudo-BMA weights, and pseudo-BMA+ weights.
3. Obtain posterior predictive distributions using the R software program rstanarm (Goodrich et al. 2020).
4. Obtain KLD measures comparing the predicted distribution of reading scores to the observed distribution. All code and data for this example are available on the authors website (http://bmer.wceruw.org/index.html).

7.1. Bayesian Stacking Results

Table 5 presents the results for Bayesian stacking. We find that Model 2, which includes predictors-related attitudes and behaviors directed toward reading, has the highest weight regardless of how the weights were calculated. We find that pseudo-BMA and pseudo-BMA+ places almost all of the weight on Model 2 whereas the stacking weights based on the log predictive score are somewhat more spread out, with model 3 having the next highest weight. We also find that the Model 2 has the lowest LOOIC value.

The bottom row of Table 5 presents the KLD measures obtained from comparing the distribution of predicted reading scores to the observed reading scores for each method of obtaining weights. Here we find the that lowest KLD value obtained under the log-score stacking weights. Overall, we find that stacking using loo weights provides overall the best predictive performance. It may be interesting to note that the KLD values for the BMA results in Table 4 are uniformly lower compared to the KLD values in Table 5 although it needs to be reiterated that BMA assumes an M-closed framework.

TABLE 5.
Log-score stacking, PseudoBMA, and PseudoBMA+ weights along with LOOIC and Kullback–Leibler divergence.

	Stacking	PseudoBMA	PseudoBMA+	LOOIC
Model 1	0.001	0.000	0.000	48277.66
Model 2	0.594	1.000	0.995	47644.65
Model 3	0.405	0.000	0.005	47860.30
Model 4	0.000	0.000	0.000	48757.57
KLD	0.016	0.018	0.018	

8. Conclusions

Although the orientation of this paper was focused on Bayesian methods for quantifying model uncertainty, it should be pointed out that issues of model uncertainty and model averaging have been addressed within the frequentist domain. The topic of frequentist model averaging (FMA) has been covered extensively in Hjort and Claeskens (2003), Claeskens and Hjort (2008) and Fletcher (2018). Our focus on Bayesian model averaging is based on some important advantages over FMA. As noted by Steel (2020), (a) BMA is optimal (under $\mathcal{M}$-closed) in terms of prediction as measured by the log predictive density score; (b) BMA is easier to implement in situations where the model space is large due to very fast algorithms such as MC^3; (c) BMA naturally leads to substantively valuable interpretations of posterior model probabilities and posterior inclusion probabilities; and (d) in the majority of content domains wherein model averaging is required, BMA is more frequently used than FMA.

As the history of model uncertainty quantification illustrates, the problem has received considerable attention in statistical theory and practice—particularly in the fields of economics and weather forecasting. However, relatively less attention has been paid to the problem of model uncertainty in psychometrics. Given the pervasive nature of model uncertainty and the consequence of ignoring the problem with respect to predictive accuracy, certain challenges and opportunities for research present themselves. The first challenge relates to shifting our research orientation away from developing models for explanation and toward developing models for prediction. Of course, these orientations are not mutually exclusive, but we argue that prediction should take precedent to explanation simply due to the fact that it is hard to warrant explanations of psychological phenomena derived from a psychometric model if the model has difficulty predicting what has actually occurred (see Dawid 1982). Posterior predictive checking (Gelman et al. 1996) in the context of model selection or model averaging should become standard practice across the social and behavioral sciences. The second challenge derives from listing those uniquely psychometric methods, such as item response theory and factor analysis, and then identifying the elements of model uncertainty that might arise when such methods are applied to real problems. For example, an interesting practical question that might arise concerns the extent of model uncertainty in the estimation of plausible values developed for large-scale assessments such as NAEP and PISA (Mislevy 1991; Mislevy et al. 1992). The estimation of plausible values is essentially a missing data problem, and although Kaplan and Yavuz (2019) showed how BMA could be incorporated into the multiple imputation setting, they did not extend their method to the full machinery of plausible value methodology. Indeed, the problem of model uncertainty as it pertains to the estimation of latent variables generally, either through IRT or factor analysis, would be very interesting to explore and obviously quite relevant to psychometric theory and practice. Promising research on this topic has begun with Rights et al. (2018), but much more work remains.

In conclusion, this review highlighted the ubiquitous problem of model uncertainty and the availability of Bayesian methods to address this problem. Given that model uncertainty raises the risk of "...over-confident inferences and decisions that are more risky than one thinks they are" (Hoeting et al. 1999, pg. 382), future research should continue to be directed to the problem of model uncertainty in the social and behavioral sciences.

Acknowledgments

The author would like to thank Mingya Huang for valuable contributions to the empirical examples in this paper, and to Betrand Clarke, David Draper, Jonah Gabry, Aki Vehtari, and Yuling Yao for valuable discussions on the problem of model uncertainty. Any errors of commission or omission in this paper are solely the responsibility of the author.

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*. Budapest: Akademiai Kiado.

Berger, J. (2013). *Statistical decision theory and Bayesian analysis*. New York: Springer.

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

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*, 49–64.

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

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

Clarke, B. S., & Clarke, J. L. (2018). *Predictive statistics: Analysis and inference beyond models*. Cambridge: Cambridge University Press.

Clyde, M. A. (1999). *Bayesian model averaging and model search strategies. Bayesian statistics* (Vol. 6, pp. 157–185). Oxford: Oxford University Press.

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

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

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

Clyde, M. A., & Iversen, E. S. (2013). *Bayesian model averaging in the M-open framework. Bayesian theory and applications* (pp. 483–498). Oxford: Oxford University Press.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, *77*, 605–610.

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

de Finetti, B. (1962). Does it make sense to speak of good probability appraisers. In I. J. Good (Ed.), *The scientist speculates—A anthology of partly-baked ideas* (pp. 357–364). London: Heinemann.

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

Draper, D. (2013). *Bayesian model specification: Heuristics and examples. Bayesian theory and applications* (pp. 483–498). Oxford: Oxford University Press.

Draper, D., Hodges, J. S., Leamer, E. E., Morris, C. N., & Rubin, D. B. (1987). *A Research Agenda for Assessment and Propagation of Model Uncertainty* (Tech. Rep.). Santa Monica, CA: Rand Corporation. Retrieved from https://www.rand.org/pubs/notes/N2683.html (N-2683-RC).

Eicher, T. S., Papageorgiou, C., & Raftery, A. E. (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, *26*(1), 30–55.

Feldkircher, M. & Zeugner, S. (2009). Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in Bayesian model averaging (No. 9-202). International Monetary Fund.

Fernández, C., Ley, E., & Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, *100*, 381–427.

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

Fletcher, D. (2018). *Model averaging*. Berlin: Springer.

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, *22*, 1947–1975.

Furnival, G. M., & Wilson, R. W, Jr. (1974). Regressions by leaps and bounds. *Technometrics*, *16*, 499–511.
Geisser, S., & Eddy, W. F. (1979). *Journal of the American Statistical Association*, *74*, 153–160.
Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 145–161). Boca Raton: Chapman & Hall.
Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies: With commentary. *Statistical Science*, *6*, 733–807.
George, E., & Foster, D. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, *1*, 87. https://doi.org/10.1093/biomet/87.4.731.
Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov Chain Monte Carlo in practice*. London: Chapman and Hall.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.
Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Series B (Methodological)*, *14*, 107–114.
Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. Retrieved from https://mc-stan.org/rstanarm (R package version 2.21.1)
Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B (Methodological)*, *41*(2), 190–195.
Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, *96*, 746–774.
Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*, 451–464.
Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*, 200–215.
Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*, 879–899.
Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
Hoerl, R. W. (1985). Ridge analysis 25 years later. *The American Statistician*, *39*(3), 186–192.
Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.
Hsiang, T. C. (1975). A Bayesian View on Ridge Regression. *Journal of the Royal Statistical Society, D (The Statistician), 24*, 267–268.
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., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research*, *49*, 505–517.
Kaplan, D., & Huang, M. (under review). Bayesian probabilistic forecasting with state NAEP data.
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 world-wide—extended context assessment frameworks*. Dordrecht: 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*,. https://doi.org/10.1080/10705511.2015.1092088.
Kaplan, D., & Yavuz, S. (2019). An approach to addressing multiple imputation model uncertainty using Bayesian model averaging. *Multivariate Behavioral Research*, *1*, 21. https://doi.org/10.1080/00273171.2019.1657790.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (2016). *Assessing contexts of learning: An international perspective*. Dordrecht: Springer.
Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.
Kullback, S. (1987). The Kullback–Leibler distance. *The American Statistician*, *41*, 340–341.
Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.
Le, T., & Clarke, B. (2017). A Bayes interpretation of stacking for $\mathcal{M}$-complete and $\mathcal{M}$-open settings. *Bayesian Analysis*, *12*, 807–829.
Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: Wiley.
Ley, E., & Steel, M. F. J. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, *24*, 651–674.
Li, Q., & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*, 151–170. https://doi.org/10.1214/10-BA506.
Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. (2008). Mixtures of *g*-priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.
Lindley, D. (1991). *Making Decisions*. London: 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.
Madigan, D., & York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, *63*, 215–232.
Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis*, *10*, 292–304.
Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.
Montgomery, J. M., & Nyhan, B. (2010). Bayesian model averaging: Theoretical developments and practical applications. *Political Analysis*, *18*, 245–270.
OECD. (2002). *PISA 2000 Technical Report*. Paris: Organization for Economic Cooperation and Development.
OECD. (2009). *Pisa 2009 assessment framework-key competencies in reading, mathematics and science*. Paris: Organization for Economic Cooperation and Development.
OECD. (2017). *PISA 2015 Technical Report*Paris: OECD.
OECD. (2018). *Equity in Education: Breaking Down Barriers to Social Mobility* (Tech. Rep.). Paris. https://doi.org/10.1787/9789264073234-en.
Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.
Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian prediction methods for model selection. *Statistics and Computing*, *27*, 711–735.
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: Blackwell.
Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, *83*, 251–266.
Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*, 1155–1174.
Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. (2015). 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.
Raftery, A. E., & Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association*, *98*, 931–938.
Rights, J., Sterba, S., Cho, S.-J., & Preacher, K. (2018). Addressing model uncertainty in item response theory person scores through model averaging. *Behaviormetrika*, *45*, 495–503. https://doi.org/10.1007/s41237-018-0052-1.
Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*, 130–134.
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.
Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*, 644–719.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, *58*, 267–288.
Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.
Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.* Retrieved from https://CRAN.R-project.org/package=loo (R package version 2.1.0).
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.
Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. https://doi.org/10.1214/12-SS102.
Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.
Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*, 1–60.
Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259.
Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*, 917–1007. https://doi.org/10.1214/17-BA1091.
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.
Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g* prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics* (pp. 233–243). New York: Elsevier.
Zeugner, S., & Feldkircher, M. (2015). Bayesian model averaging employing fixed and flexible priors: The BMS package for R. *Journal of Statistical Software*, *68*(4), 1–37. https://doi.org/10.18637/jss.v068.i04.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Manuscript Received: 11 OCT 2020
Final Version Received: 15 FEB 2021
Published Online Date: 15 MAR 2021