

# Predictive Performance of Bayesian Stacking in Multilevel Education Data

Mingya Huang 

*University of Wisconsin-Madison*

David Kaplan

*University of Wisconsin-Madison*

*The issue of model uncertainty has been gaining interest in education and the social sciences community over the years, and the dominant methods for handling model uncertainty are based on Bayesian inference, particularly, Bayesian model averaging. However, Bayesian model averaging assumes that the true data-generating model is within the candidate model space over which averaging is taking place. Unlike Bayesian model averaging, the method of Bayesian stacking can account for model uncertainty without assuming that a true model exists. An issue with Bayesian stacking, however, is that it is an optimization technique that uses predictor-independent model weights and is, therefore, not fully Bayesian. Bayesian hierarchical stacking, proposed by Yao et al. further incorporates uncertainty by applying a hyperprior to the stacking weights. Considering the importance of multilevel models commonly applied in educational settings, this paper investigates via a simulation study and a real data example the predictive performance of original Bayesian stacking and Bayesian hierarchical stacking along with two other readily available weighting methods, pseudo-BMA and pseudo-BMA bootstrap (PBMA and PBMA+). Predictive performance is measured by the Kullback–Leibler divergence score. Although the differences in predictive performance among these four weighting methods in Bayesian stacking are small, we still find that Bayesian hierarchical stacking performs as well as conventional stacking, PBMA, and PBMA+ in settings where a true model is not assumed to exist.*

**Keywords:** *Bayesian statistics; Ensemble methods; Large-scale assessment; Prediction; Multilevel modeling.*

## *Issues of Model Uncertainty*

Model selection and model uncertainty have been a general challenge in statistical inference for decades. The problem has been summarized by, among others, Hoeting et al. (1999) who 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. (p. 382)

Similar sentiments have been expressed earlier by Leamer (1978) and Draper et al. (1987). Indeed, the problem of ignoring model uncertainty has been recognized by Breiman (1992).

In education studies, it is common to encounter data with a multilevel structure. For example, students are typically nested within schools, and thus there will be uncertainty in prediction at the individual level as well as the group level. To better capture the nesting effects, methods such as multilevel modeling (also referred to as linear mixed-effects models or variance components models) are typically employed (e.g., Raudenbush & Bryk, 2002; Goldstein, 2011). In addition, one can use modeling strategies such as the least absolute shrinkage and selection operator (LASSO) to select the appropriate variables and come up with the “best” model for prediction. However, despite the power and popularity of these approaches, they rely on the assumption that the selected model is the one that actually generated the data, thus ignoring the typical practice of searching for a best-fitting model and other “worse” models can also contain useful information for predicting the outcome variables. For instance, to predict students’ reading scores, the “best” model might omit *number of books at home* (HOMEBOOKS) as a covariate at the population level, but HOMEBOOKS itself might be significantly correlated with test scores. Furthermore, HOMEBOOKS could potentially be effective in forecasting reading scores for a particular student subgroup despite being omitted from the model deemed “best” for the overall population. In this case, if one simply selects the “best” model, one might lose information from HOMEBOOKS. This leads to a discussion of so-called  $\mathcal{M}$ -frameworks.

### *$\mathcal{M}$ -frameworks*

When considering the problem of model selection, three types of relationships between the true data-generating model (DGM) and substantive models need to be considered— $\mathcal{M}$ -closed,  $\mathcal{M}$ -complete, and  $\mathcal{M}$ -open, as introduced by Bernardo and Smith (1994):

- In the  $\mathcal{M}$ -closed setting, the true DGM is within the candidate model list. That is, the true DGM,  $M_t$ , is one of the  $M_k \in \mathcal{M}$ , where  $k = 1, 2, \dots, K$  denotes the number of models.

- In the  $\mathcal{M}$ -complete setting, the true DGM is assumed to exist, but it is not within the candidate model list, though the explicit form of  $p(\tilde{y}|y) = p(\tilde{y}|M_t, y)$  can be derived. Rather, a list of models is considered, with each model serving as a reasonable proxy to the true DGM.
- In the  $\mathcal{M}$ -open setting, not only is  $M_t$  not in  $\mathcal{M}$ , but also the explicit form of the true DGM cannot be specified.

As we see, in the  $\mathcal{M}$ -complete and  $\mathcal{M}$ -open settings, we cannot assume the true DGM is within the candidate model space. Therefore, instead of relying on a single best model that assumes  $\mathcal{M}$ -closed, averaging the information across different models could potentially be a superior option (Bernardo & Smith, 1994).

### *Limitations of Existing Methods*

There are numerous methods to evaluate multiple candidate models as well as handling model uncertainty from a Bayesian perspective (e.g., Geisser & Eddy, 1979a; Kass & Raftery, 1995). Through the joint efforts of many researchers (Clyde, 1999, 2003; Draper, 1995; Hoeting et al., 1999; Leamer, 1978; Raftery et al., 1997), *Bayesian model averaging* (BMA) was introduced to handle model uncertainty and obtain optimal predictive performance and has become the choice of methods. Generally speaking, BMA averages coefficients across a large space of models weighted by each model's marginal posterior probability. For a candidate model list  $\mathcal{M} = (M_1, \dots, M_K)$ , the posterior probability of the quantity of interest  $\theta$  (e.g., a predicted value, denoted as  $\tilde{y}$ ) can be expressed as

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

where  $y_1, y_2, \dots, y_n$  are the observed data. Each model is weighted by the posterior model probability,  $p(M_k|y)$ , where all weights sum to one and are all between 0 and 1.

Bayesian model averaging has shown good out-of-sample predictive performance across a variety of settings (Hoeting et al., 1999). In the context of education research, Kaplan and his colleagues (Kaplan, 2021; Kaplan & Chen, 2014; Kaplan & Huang, 2021; Kaplan & Lee, 2015, 2018; Kaplan & Yavuz, 2019), have discussed and extended Bayesian model averaging primarily to problems in large-scale educational assessments, such as the Program for International Student Assessment (PISA; OECD, 2002). The general problem of using BMA lies in its major assumption, namely, BMA assumes an  $\mathcal{M}$ -closed setting. Indeed, the posterior model probabilities for each candidate model is a measure

of the probability that model  $k$  is the true model.<sup>1</sup> Another issue with BMA is its sensitivity to the choices of priors on the model parameters. Different  $p(\theta_k|M_k)$  can lead to quite different results. For instance, Fernández et al. (2001) has shown that different choices of priors on the parameter  $\theta_k$  can yield disparate outcomes. Therefore, the accurate predictive performance of BMA requires the correct specification of the prior information.

To address these issues, Yao et al. (2018a) apply a particular method of ensemble modeling, referred to as *Bayesian stacking*, to combine predictions across candidate models, which has been shown to achieve better prediction than BMA under the  $\mathcal{M}$ -complete and  $\mathcal{M}$ -open settings. Instead of using the posterior model probability as the weighting scheme, Bayesian stacking employs an optimization function that selects weight to maximize the log predictive densities across all the models. In this way, Bayesian stacking is a viable approach to yield optimal predictions in the  $\mathcal{M}$ -complete and  $\mathcal{M}$ -open setting. For example, with  $p$  covariates to predict students' reading scores, there will be  $p^2$  models that can be considered as the candidate models. Researchers who are interested in the effects of demographic measures will only consider a subset of these models, for instance, the models that include gender. As such, the true DGM is possibly not within the candidate model sets ( $\mathcal{M}$ -open setting). Therefore, the prediction obtained by BMA, which uses the posterior model probability as the weighting scheme, is questionable since it only applies to the selected candidate model probabilities. Yao et al. (2022) further developed a more adaptive weighting method referred to as *Bayesian hierarchical stacking* (BHS), which has been shown to achieve optimal prediction and is to be described in the next section.

### *Our Contributions*

This paper aims to examine different weighting methods in Bayesian stacking using data with multilevel structures in the  $\mathcal{M}$ -closed,  $\mathcal{M}$ -complete, and  $\mathcal{M}$ -open framework. We compare Bayesian original stacking (BS), Bayesian Hierarchical Stacking (BHS), pseudo-BMA (BMA), and pseudo-BMA bootstrap (PBMA+). Our contribution to the literature is three-fold. First, BS, BHS, PBMA, and PBMA+ are relatively unknown methods in education research, and so our paper provides a systematic comparison of these weighting methods via a simulation study and a substantive example using data from a large-scale educational assessment. Second, this paper conducts the comparison of these different weighting schemes for Bayesian stacking under the three  $\mathcal{M}$ -frameworks: one empirical study to represent the  $\mathcal{M}$ -open setting and two simulation studies to construct the  $\mathcal{M}$ -closed and  $\mathcal{M}$ -complete setting. Besides examining the predictive capacity of each of these weighting methods, we will also show how these three  $\mathcal{M}$  frameworks lead to different results. Third, stacking, in general, is not well-known in the educational literature, where it is common to

implement multilevel models to study within and between-school predictors of academic and non-academic outcomes (Raudenbush & Bryk, 2002). Thus, questions such as improving the predictive performance of multilevel models should be of general interest to education and social science researchers.

The remainder of this paper is organized as follows. In the next section, we provide an overview of the stacking weights that will be used in this study. They include the original stacking weights proposed by Yao et al. (2018a) and the newly developed hierarchical stacking weights proposed by Yao et al. (2021). In addition, we include two other types of weights that have been proposed for stacking, including so-called *pseudo-BMA* (PBMA) and *pseudo-BMA bootstrapping* (PBMA+) weights. This is followed by an empirical study using United States data from the 2018 cycle of the Program on International Student Assessment (PISA; OECD, 2018). This is then followed by the details of our simulation design investigating the predictive performance of these four types of stacking weights in the context of multilevel models. The paper closes with conclusions and directions for further research.

### **Types of Bayesian Stacking Weights**

In this section, we briefly review the relevant background of different weighting methods in Bayesian stacking: original stacking weights, Bayesian hierarchical stacking, pseudo-BMA, and pseudo-BMA bootstrap.

#### *Original Stacking Weights*

Following a recent review by Kaplan (2021), stacking involves weighting the predictive distributions obtained from multiple candidate models comprising an ensemble based on different scoring rules such as the Kullback–Leibler divergence (KLD; Kullback, 1959, 1987) or log predictive densities (LPD; Good, 1952) to obtain an optimal prediction. In our empirical example later, the outcome of interest will be students’ scores on the reading literacy assessment from PISA 2018. First, we enumerate all the candidate models which can be denoted as  $f_k(x)$  with different covariates  $x$ :

$$y = f_k(x) + \epsilon \tag{2}$$

Note that the notation  $f_k(x)$  allows each model to have distinct functional forms. The optimal prediction is obtained from a weighted combination of the predictive densities from each  $f_k(x)$ . In particular,  $\hat{f}_k$  is used to estimate  $f_k$ , and  $\tilde{y}$  is the predictive distribution based on the data  $y$ . That is,

$$\hat{y} = \sum_{k=1}^K \hat{w}_k \hat{f}_k(x). \quad (3)$$

To minimize the loss function between the weighted combination of predictive distributions and the actual outcome distribution, the weight  $\hat{w}$  is computed based on an optimization function:

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \left( y_i - \sum_{k=1}^K w_k \hat{f}_{k,-i}(x_i) \right)^2, \quad (4)$$

where  $\hat{f}_{k,-i}(x_i)$  is an estimate of  $f_k$  based on  $n - 1$  observations, leaving the  $i^{\text{th}}$  observation out. In Bayesian stacking, leave-one-out cross-validation (LOO-CV) is used to compute  $\hat{f}_{k,-i}(x_i)$ . Similar to  $q$ -fold CV, which holds one fold out for validation data set, in LOO-CV, each observation serves as the validation set, with the remaining  $n - 1$  observations serving as the training set. To be precise, the expected log pointwise predictive density (ELPD) can be derived as

$$\text{ELPD} = \sum_{i=1}^n \int p_t(\tilde{y}_i) \log p(\tilde{y}_i | y) d\tilde{y}_i, \quad (5)$$

where  $p_t(\tilde{y}_i)$  represents true DGM process for the predicted values  $\tilde{y}_i$ . By leaving the  $i^{\text{th}}$  data point out one at a time, the Bayesian LOO estimates will be

$$\text{ELPD}_{\text{loo}} = \sum_{i=1}^n \log p(y_i | y_{-i}), \text{ where } p(y_i | y_{-i}) = \int p(y_i | \theta) p(\theta | y_{-i}) d\theta. \quad (6)$$

LOO-CV can be implemented by the R software program LOO (Vehtari et al., 2019).<sup>2</sup>

### *Hierarchical Stacking Weights*

An issue with Bayesian stacking is that it is an optimization technique that creates *predictor-independent* model weights for stacking—meaning that the weights do not vary as a function of the predictor variables. Specifically, in an ensemble of models, each member model receives one weight that is not dependent on predictors in the model but rather is determined in such a way as to optimize leave-one-out prediction. In this case, predictor-independent weights are not fully Bayesian, and it would be perhaps preferable to allow for *predictor-dependent weights*.

To provide for the flexibility of using predictor-dependent weights within the Bayesian framework, Yao et al. (2021) proposed the method of *Bayesian hierarchical stacking* (BHS). Taking our example using PISA data, the weight assigned to student  $i$  in school  $j$  for a specific model to predict the student's

reading score should be different from the weight assigned to student  $i'$  in school  $j'$  for the prediction of that student's reading score with the same model. Accounting for differences in the weights for each observation can explain their unique characteristics and thus lead to a more precise prediction for each student.

The main difference between original Bayesian stacking and BHS lies in the computation of stacking weights. Instead of using a single weight simplex in BMA and BS, Yao et al. (2021) suggested using a weight function  $\mathbf{w}(x) = (w_1(x), \dots, w_k(x))$ , where  $x$  generically denotes a set of predictors in each model. The weights themselves are then a function of an additive model written as

$$w_k^*(x) = \mu_k + \sum_{m=1}^M \alpha_{mk} f_m(x), \quad k \leq K-1, \quad w_K^*(x) = 0, \quad (7)$$

and where  $\mu_k$  is the average weight for model  $k$ ,  $\alpha_{mk}$  is a weight attached to each of the  $m$  predictors in model  $k$ , and  $f_m$  are  $m$  distinct predictors. The BHS algorithm that we use follows the approach of Yao et al. (2021) that separates predictors into those that are discrete and those that are continuous. The following priors are attached to the model parameters, namely

$$\alpha_{mk} | \sigma_k \sim \mathcal{N}(0, \sigma_k), \sigma_k \sim \mathcal{N}^+(0, \tau_\sigma), \mu_k \sim \mathcal{N}(\mu_0, \tau_\mu), \quad (8)$$

where  $\mathcal{N}^+$  is the half-normal distribution (see Yao et al., 2021, for a discussion of hyperprior choices). Following the suggestion given by Yao et al. (2021), for the example and simulation study below, we use the following weakly informative priors:

$$\mu_0 \sim \mathcal{N}(0, 1), \tau_\mu = 1, \tau_{\sigma_{k1}} = \tau_{disc} = 0.5, \tau_{\sigma_{k2}} = \tau_{cont} = 1, \quad (9)$$

where  $\tau_{\sigma_{k1}}$  and  $\tau_{\sigma_{k2}}$  are hyperpriors for  $\sigma_k$  depending on whether the predictors are discrete or continuous (see Yao et al., 2021, for more details).

From here, the pointwise predictive density can be written as

$$p(\hat{y} | \tilde{x}, \mathbf{w}(\cdot)) = \sum_{k=1}^K w_k(\tilde{x}) p(\hat{y} | \tilde{x}, M_k), \quad (10)$$

The posterior distribution of the stacking weights is obtained as

$$\log p(\mathbf{w}(\cdot) | \mathcal{D}) = \sum_{i=1}^n \log \left( \sum_{k=1}^K w_k(x_i) p_{k,-i} \right) + \log p^{prior}(\mathbf{w}) + \text{constant} \quad (11)$$

$$w_{1:K}(x) = \text{softmax}(w_{1:K}^*(x)), \quad (12)$$

where the term  $p^{prior}$  refers to the prior distributions in Equation 8, and where the softmax function converts a vector of real numbers into a vector of probabilities (Stan Development Team, 2021).

The connection between BHS and original Bayesian stacking can be seen through the concept of pooling as it pertains to the stacking framework for the weight functions, namely, *complete pooling*, *no-pooling*, and *partial pooling method* (Yao et al., 2021). Complete pooling is the same as the original stacking approach, that is, using a weight simplex where each predictor has the same weight  $w_k(x) = w_k$ . No-pooling stacking separately optimizes the objective function in Equation 11 for each  $x_i$  independently. The last method is *partial pooling* stacking, which requires an appropriate hierarchical prior  $p^{prior}(\cdot)$  so that the posterior distribution of the stacking weights can be obtained by solving Equation (11). Partial pooling stacking is the BHS method.

Finally, Yao et al. (2021) also discussed different choices of priors and recommended, as a general rule, using weakly informative priors, such as using a half-normal prior on the model scale parameters rather than half-Cauchy or inverse-gamma priors because the latter two will lead to larger dispersion. However, researchers can choose different priors based on the purpose of their research and the structure of the data.

Considerably more detail regarding hierarchical stacking weights can be found in Yao et al. (2021). Suffice it to say that a contribution of this paper is the evaluation of Bayesian hierarchical stacking in light of other existing stacking weights and in the context of multilevel models with a specific focus on predictive accuracy.

### *Pseudo-BMA Weights*

Pseudo-BMA weights were proposed by (Geisser & Eddy, 1979b; see also Gelfand, 1996; Yao et al., 2018b). The basic idea behind PBMA is as follows. First, as discussed in Yao et al. (2021), LOO-CV has connections to other types of weights that can be used for stacking. For example, in the case of maximum likelihood estimation, LOO-CV weights are asymptotically equivalent to Akaike information criterion (AIC) weights (Akaike, 1973) that are used in frequentist model averaging applications (Yao et al., 2018b; see also Burnham & Anderson, 2002; Fletcher, 2018). As a method of model selection, earlier work by Geisser and Eddy (1979b; see also Gelfand, 1996) criticized the underpinnings of Bayes factors and suggested substituting the marginal likelihood of the  $k^{th}$  model,  $p(y|M_k)$ , used in the calculation of Bayes factors with Bayesian leave-one-out cross-validation predictive densities, defined as  $\prod_{i=1}^n p(y_i|y_{-i}, M_k)$ . Yao et al. (2018b) refer to AIC weighting using LOO-CV predictive densities as PBMA.



*Pseudo-BMA + Weights*

The difficulty with PBMA weights is that they do not take into account uncertainty in the LOO estimation of the weights. To address this Yao et al. (2018b) proposed an approach that combines the Bayesian bootstrap (see Rubin, 1981) with the ELPD defined earlier. They refer to this approach as *pseudo-BMA+* (PBMA+). The essential idea behind PBMA+ is that the posterior distribution of the realizations of a random variable  $Z$ , that is,  $z_i, i = 1, \dots, n$ , follows a Dirichlet(1, ..., 1) distribution. Taking samples from this distribution yields Bayesian bootstrap samples from which parameters from this distribution can be calculated. Yao et al. (2018b) has noted that the ELPD based on LOO can be highly skewed and argues that the Bayesian bootstrap might be an improvement over the usual Gaussian approximation. The PBMA+ weighting follows essentially the same line of argument as the conventional Bayesian bootstrap. That is, define for each model  $k$ , we have

$\{z\}_{i=1}^k = \left\{ \widehat{\text{ELPD}}_{loo} \right\}_{k=1}^K$ . Then taking  $B$  bootstrap samples  $(\pi_{1,b}, \dots, \pi_{n,b})$ ,

$b = 1, \dots, B$  from  $\overbrace{\text{Dirichlet}(1, \dots, 1)}^n$  allows us to calculate the weighted means as  $\bar{z}_b^k = \sum_{i=1}^n \pi_{i,b} z_i^k$ . From here, a Bayesian bootstrap sample of the stacking weight for model  $k$  based on bootstrap samples of size  $B$  can be obtained as

$$w_{k,b} = \frac{\exp(n\bar{z}_b^k)}{\sum_{k=1}^K \exp(n\bar{z}_b^k)}, \quad b = 1, \dots, B, \quad (13)$$

leading to the final PBMA+ weight for model  $k$ ,

$$w_k = \frac{1}{B} \sum_{b=1}^B w_{k,b}, \quad (14)$$

Of importance to this paper, Yao et al. (2018b) showed that PBMA+ performs better than BMA and PBMA in  $\mathcal{M}$ -open settings but not as well as stacking using the log score. This paper adds to the existing literature by comparing original stacking and hierarchical stacking weights to PBMA and PBMA+ weights in the context of multilevel models applied to large-scale assessments.

### Empirical Study

This section examines the predictive performance of BS, BHS, PBMA, and PBMA+ using data from PISA 2018 under the  $\mathcal{M}$ -open setting. We considered

this empirical study as an  $\mathcal{M}$ -open setting due to two reasons: (1) We are not able to know if the true DGM is within the candidate model sets; (2) we also cannot derive the explicit form of the true DGM. We use open-source PISA data from the Organisation for Economic Cooperation and Development, which is a triennial international survey that aims to evaluate education systems across the world (79 countries) and measures 15-year-olds' ability to use their cognitive outcomes such as reading, mathematics, and science knowledge and skills to meet real-life challenges. There are 4,838 participants randomly selected for this study (an average of 30 students in 164 schools), and we specify four models to be our candidate models using nineteen covariates (see Table A1 for details) and the first plausible value of the reading assessment as the dependent variable.

For this empirical data, two sample sizes are examined: (a) a small sample of 500 students and (b) the full PISA 2018 sample size of 4,838 with 164 participating schools and approximately 30 students within the sampled schools. Considering the ratio between the number of students in each school and the number of schools in PISA data is approximately 1:5, we randomly selected schools ( $J=1, 2, \dots, 50$ ) and 10 randomly selected students in the selected schools for the small sample ( $I=1, 2, \dots, 10$ ). Based on previous literature, student's academic performance is influenced by several predictors: (1) students' level covariates such as demographic measures, motivations, attitudes, behaviors, etc. (e.g., Brozo et al., 2014; Caro et al., 2016; Michael & Kyriakides, 2023); (2) covariates such as ICT resources at the school level (e.g., Zhang & Liu, 2016). Following the empirical study conducted by Kaplan (2021) here extended to multilevel models, our candidate models for the ensemble are as follows:

- Model 1 includes the demographic measures (FEMALE, ESCS, HOMEPOS) and a random intercept and a random slope for ICTRES nested within schools.
- Model 2 investigates the effects of attitudes and behaviors on reading scores (JOYREAD, PISADIFF, SCREADCOMP, SCREADDIFF) with a random intercept for the school effects.
- Model 3 consists of predictors about academic mindset and students' general well-being (METASUM, GFOFAIL, MASTGOAL, SWBP, WORKMAST, ADAPTIVITY, COMPETE) and a random intercept accounting for school effects.
- Model 4 examines the effects of students' attitudes toward the school on reading scores (PERFEED, TEACHINT, BELONG) with a random intercept and a random slope of TEACHINT nested in schools.

To begin, Table 1 is an example of the differences in the model obtained by different Bayesian stacking methods for the first two students in the small

TABLE 1.  
*Model Weights for the First Two Students in the Small Sample (N = 500)*

Model weights	Student A				Student B			
	BS	BHS	PBMA	PBMA+	BS	BHS	PBMA	PBMA+
Model 1	0.042	<b>0.155</b>	0.000	0.014	0.042	<b>0.067</b>	0.000	0.014
Model 2	0.576	<b>0.648</b>	0.921	0.615	0.576	<b>0.555</b>	0.921	0.615
Model 3	0.382	<b>0.068</b>	0.079	0.367	0.382	<b>0.311</b>	0.079	0.367
Model 4	0.000	<b>0.130</b>	0.000	0.004	0.000	<b>0.068</b>	0.000	0.004

*Note.* BMA = Bayesian model averaging; PBMA = pseudo-BMA; PBMA+ = pseudo-BMA bootstrap. The bold values denote the weights obtained by BHS, indicating only BHS has obtained different weights for different students.

sample. While BS, PBMA, and PBMA+ assign the same model weights to different students, we see that BHS assigns different weights to different students, even for the same model. In Table 1, model 2 has less weight for student B (0.555) compared to student A (0.648) by BHS. This indicates for student A, model 2 plays a more important role compared to student B in terms of prediction. To wit, the effects of attitudes and behaviors on reading scores are more important for student A than student B regarding reading scores. Thus, we see that BHS can account for the randomness due to the school effects and assign different model weights to students in different schools.

The results for the average model weights are summarized in the upper panel of Table 2. Across both sample size conditions and weighting methods, Model 2 is preferred over the other models. However, we can see that PBMA and PBMA+ tend to put the majority of the weight on a single model. By contrast, BS and BHS have a more balanced weighting scheme on all the models without emphasizing one model. In addition, we also find that the average model weight obtained by BHS differs from the ones in Table 1. This shows that BHS not only provides a general summary of how each model performs at the sample level but also at the individual level.

The lower panel of Table 2 summarizes the results of the Kullback -Leibler divergence (KLD) for sample sizes of 500 and 4,838. Here we consider two distributions,  $p(y)$  and  $g(y|\theta)$ , where  $p(y)$  denotes the distribution of observed reading literacy scores and  $g(y|\theta)$  denotes the prediction of these reading scores based on a model. The KLD between these two distributions can be written as

$$\text{KLD}(p, g) = \int p(y) \log \left( \frac{p(y)}{g(y|\theta)} \right) dy, \quad (15)$$

where  $\text{KLD}(f, g)$  is the information lost when  $g$  is used to approximate  $f$ . For example, the actual reading outcome scores might be compared to the predicted

TABLE 2.  
Average Model Weights and Predictive Performance Comparisons for PISA 2018  
Example

Model weights	$N=500$				$N=4,838$			
	BS	BHS	PBMA	PBMA+	BS	BHS	PBMA	PBMA+
Model 1	0.042	0.189	0.000	0.014	0.000	0.113	0.000	0.000
Model 2	0.576	0.393	0.921	0.615	0.638	0.477	1.000	0.954
Model 3	0.382	0.277	0.079	0.367	0.362	0.322	0.000	0.046
Model 4	0.000	0.142	0.000	0.004	0.000	0.098	0.000	0.000
Predictive scores								
KLD	0.031	0.045	<b>0.018</b>	0.032	0.055	<b>0.039</b>	0.074	0.073

The bold values denotes the smallest KLD across all the methods.

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 the lowest.

Inspecting Table 2 we find that each method puts most of the weight on Model 2, but that BHS spreads the weights a bit more evenly across the models. In terms of KLD, we find very little difference between BS, PBMA, and PBMA+ for the small sample size case. BHS has the highest KLD and, therefore, poorer predictive performance in the small sample size case. We speculate that this result may be due to the sensitivity of BHS to the choice of weakly informative hyperpriors on the weight function (see Equation 8). For the large sample case, we find that the weight is mostly placed on Model 2 and that BHS shows the best performance in terms of KLD. Based on this real data example, one can conclude that BHS performs better than the other weighting methods for stacking in large sample sizes, such as those found in multilevel models applied to large-scale assessments such as PISA. In practice, and consistent with most Bayesian workflows (e.g., Gelman et al., 2020; Kaplan, 2023), it would be important to examine the sensitivity of the results to small changes to the hyperpriors of the weight function such as those we used in Equation 9.

### Design of Simulation Study

In the previous section, we found evidence that BHS shows better predictive performance, particularly in the large sample size case, whereas PBMA might be preferred in the small sample size case. To gain a better understanding of the predictive accuracy of these different stacking methods, we next implement a comprehensive simulation study to examine the predictive performance for all

Bayesian stacking methods under  $\mathcal{M}$ -closed and  $\mathcal{M}$ -complete settings under different sample size conditions and different intraclass correlations.

### *Data Generation Process*

For our simulation study, we generate data based on model 2 since it obtains the highest weight in the empirical study. Model 2 includes four covariates and one random intercept for the school. Let  $x_{1ij}$ ,  $x_{2ij}$ ,  $x_{3ij}$ , and  $x_{4ij}$  denote the covariates which are normally distributed with the same mean and variance as the corresponding attitude and behaviors related variables in the PISA data. To distinguish  $\mathcal{M}$ -closed and  $\mathcal{M}$ -complete settings, we add a scalar,  $\sigma$ , to the Gaussian noise  $\epsilon_{ij}$ . More specifically,

$$\begin{aligned} y_{ij} &= f(x) + \sigma\epsilon_{ij} \\ f(x) &= \beta_{00} + \beta_{01}x_{1ij} + \beta_{02}x_{2ij} + \beta_{03}x_{3ij} + \beta_{04}x_{4ij} + U_{0j}, \end{aligned} \tag{16}$$

where  $y_{ij}$  is the response variable for student  $i$  in school  $j$ ,  $x_{pij}$  denotes student  $i$  who is in school  $j$ , and has a value on variable  $p$  ( $p = 1, \dots, P$ ). The parameter  $\beta_{00}$  is the overall intercept,  $\beta_{0p}$  denote the regression coefficients for  $x_{pij}$   $p = 1, 2, 3, 4$ . The term  $U_{0j}$  denotes the random intercept for the school effects. We set  $\sigma$  to be 0 for the  $\mathcal{M}$ -closed setting, which indicates the data is generated exactly according to the true DGM without any noise. In the  $\mathcal{M}$ -complete setting, we set  $\sigma$  to be 5, which indicates that we know the explicit form of the true DGM, but the generated data does not completely depend on the true DGM. In this way, we can include the DGM,  $f(x)$  in the ensemble for both situations to examine the predictive performance in  $\mathcal{M}$ -closed and  $\mathcal{M}$ -complete settings.

As mentioned above, PISA 2018 data for the United States has approximately 30 students nested in 150 schools, which is a 1:5 ratio for within-group sample size to the between-group sample size. To mimic real data settings, we generate the data with a small sample of 500 where the number of schools  $J$  is set to be 50 and the number of students  $I$  in each school is 10. For the large sample, we set  $I$  to be 30 and  $J$  to be 150. In addition, Hedges and Hedberg (2007) have demonstrated that the intra-class correlation (ICC) in educational data generally falls in the range between .1 to .25. Therefore, we also study the impact of between-school variability by setting the ICC to .1, .2, and .3 when we generate the data. More specifically, the corresponding standard deviations are 1.667, 2.041, and 5 in the  $\mathcal{M}$ -complete setting. Noticing that ICC cannot be computed when the individual variation is zero ( $\sigma = 0$ ) in the  $\mathcal{M}$ -closed setting, we also set the same between-group standard deviation as in the  $\mathcal{M}$ -complete setting.

After generating the data, we fit the simulated data using the `rstanarm` packages (Goodrich, Gabry, Ali, & Brilleman, 2022) in the statistical software environment  $R$  (R Core Team, 2022) version 4.2.1. Given that we are interested

TABLE 3.  
*Summary of Covariates in Each Candidate Model.*

Models	Covariates	Random effects
M1	$x_{1ij}$	$U_{0j}$
M2	$x_{2ij}$	$U_{0j}$
M3	$x_{3ij}$	$U_{0j}$
M4	$x_{4ij}$	$U_{0j}$
M5	$x_{1ij} + x_{2ij}$	$U_{0j}$
M6	$x_{1ij} + x_{3ij}$	$U_{0j}$
M7	$x_{1ij} + x_{4ij}$	$U_{0j}$
M8	$x_{2ij} + x_{3ij}$	$U_{0j}$
M9	$x_{2ij} + x_{4ij}$	$U_{0j}$
M10	$x_{3ij} + x_{4ij}$	$U_{0j}$
M11	$x_{1ij} + x_{2ij} + x_{3ij}$	$U_{0j}$
M12	$x_{1ij} + x_{2ij} + x_{4ij}$	$U_{0j}$
M13	$x_{1ij} + x_{3ij} + x_{4ij}$	$U_{0j}$
M14	$x_{2ij} + x_{3ij} + x_{4ij}$	$U_{0j}$
M15	$x_{1ij} + x_{2ij} + x_{3ij} + x_{4ij}$	$U_{0j}$

in how the school effects bring randomness to the prediction, all of our candidate models include a random intercept for school. Therefore, there are 15 candidate models (denoted as M1–M15) in total with different choices of the combination of four covariates (i.e.,  $\binom{4}{1}$ ,  $\binom{4}{2}$ ,  $\binom{4}{3}$ ,  $\binom{4}{4}$ ). Table 3 summarizes the covariates in each candidate model and  $U_{0j}$  denotes the random school intercept. We used `stan_lmer` function in `rstanarm` to fit all the candidate models and extracted the weighted densities for each model using `loo` function in `R` package `loo` (Vehtari et al., 2022). For BHS, we used `rstan` (Stan Development Team, 2020) to specify the corresponding priors, hyperpriors, predictor-dependent weights, and log-likelihoods. All software code for the simulation study is available in a GitHub repository [https://github.com/mhuang233/BayesStacking\\_Edu\\_2024](https://github.com/mhuang233/BayesStacking_Edu_2024).

### Results of Simulation Study

This section presents the simulation study results. First, we discuss the convergence of the algorithms. This is followed by a discussion of the impact of different stacking methods across sample sizes and between-group variability conditions. Following, we discuss the predictive performance of the different stacking methods. Finally, we discuss the computational efficiency of running these different stacking approaches.

### *Convergence*

To begin, we examine the two MCMC convergence criteria for all 15 models. Conventional measures for convergence are the *potential scale reduction factor* (PSRF) and the *effective sample size* (ESS). PSRF, often denoted as *Rhat* or  $\hat{R}$  is based on an analysis of variance and is intended to assess convergence among several parallel chains with varying starting values and is measured by the ratio of the between-chain variance to the within-chain variance (Gelman, 1996). The idea is that if the ratio of these two sources of variance is equal to one, then this is evidence that the chains have converged. If  $\hat{R} > 1.01$ , this may be a cause for concern. The ESS is a measure of the number of independent MCMC draws, which is proportional to the number of iterations of the MCMC algorithm and the autocorrelation present in the samples. The closer the ESS is to the number of samples taken from the posterior distribution (accounting for warm-up samples and thinning), the better convergence the model has achieved.

With 100 replications used in this study, the relative prediction bias for all the fitted models was less than 10%, which indicates that 100 replications are adequate. With four chains and 10,000 iterations per chain, 5,000 iterations used for warm-up, and a thinning interval of 10, the total number of draws on which inferences are being made is 2,000. We find that all of the models managed to converge with the effective samples at around 2,000, and all  $\hat{R}$  were less than 1.01.

### *Model Weights*

In this section, we investigate the difference in model weights obtained by different weighting methods in Bayesian stacking. Since BHS uses predictor-dependent weights for each covariate, we take an average of the weights for each predictor and compare it with the other BS methods.

Figure 1 shows the model weights for both the small sample and the large sample across different levels of between-group variability in the  $\mathcal{M}$ -closed setting ( $\sigma = 0$ ). As expected, the true DGM (M15) yields the highest weights within the  $\mathcal{M}$ -closed setting. More specifically, PBMA and PBMA+ assign 100% weight to model 15, while BHS and BS distribute the weights across the other models. This becomes more apparent in the large sample ( $N=4,500$ ), where BHS and BS appear to put more weight in model 12, compared to the small sample case. This might be due to the existence of unstructured random noise created during the data generation process. Generally speaking, there is not much difference in model weights when the between-group standard deviation changes from 1.667 to 2.041 and 5. The variation in the model weights appears to be mainly due to the sample size for all the Bayesian stacking methods.

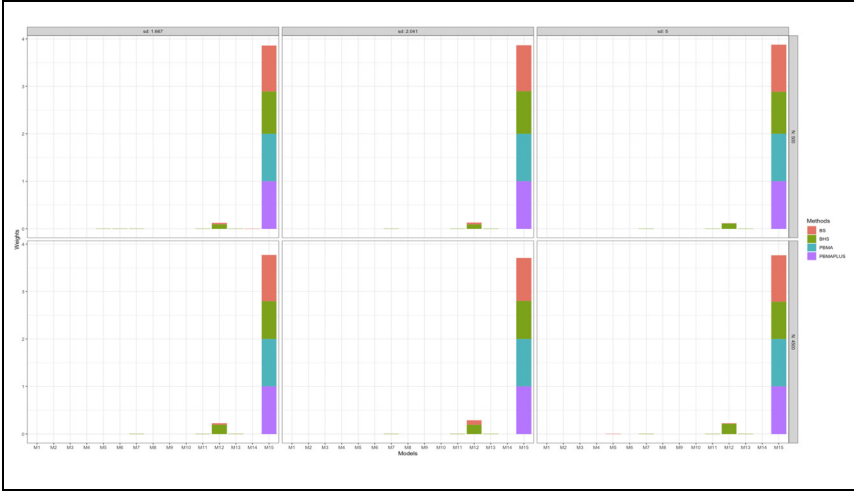


FIGURE 1. Model weights across different sample sizes with different levels of between-group variability in  $\mathcal{M}$ -closed settings.

Similarly, Figure 2 shows the model weights in the  $\mathcal{M}$ -complete setting ( $\sigma = 5$ ). Generally speaking, the weights are more “spread out” in the  $\mathcal{M}$ -complete setting. For instance, in both the small sample and large samples, model 12 stands out in terms of model weights, along with model 15. To wit, model 15 does not attain almost 100% weight as it does in the  $\mathcal{M}$ -closed setting. Different from BS and BHS, PBMA, and PBMA+ assign the vast majority of weights to model 15. For instance, the weights assigned to model 15 are not 100% in the small sample case when the between-group standard deviation=1.667, even though they are very close to 1.00. However, as the between-group variability increases, the weights assigned to model 15 by PBMA and PBMA+ are close to 100% again for both small samples and large samples. Therefore, we anticipate that if the sample size approaches infinity, the weights assigned to model 15 using these two methods will approach one. As for BHS, both model 12 and model 15 obtain noticeable weights when  $N=500$  and  $N=4,500$ . Unlike the other three methods, BS assigns different proportions of weights to each model. That is, all of the fifteen models obtain non-zero weights and there is no prominent model. Noticed that we set  $\sigma$  to be 5, which is not a large number. Therefore, we anticipate there will be larger variability in model weights if we increase the error variance.



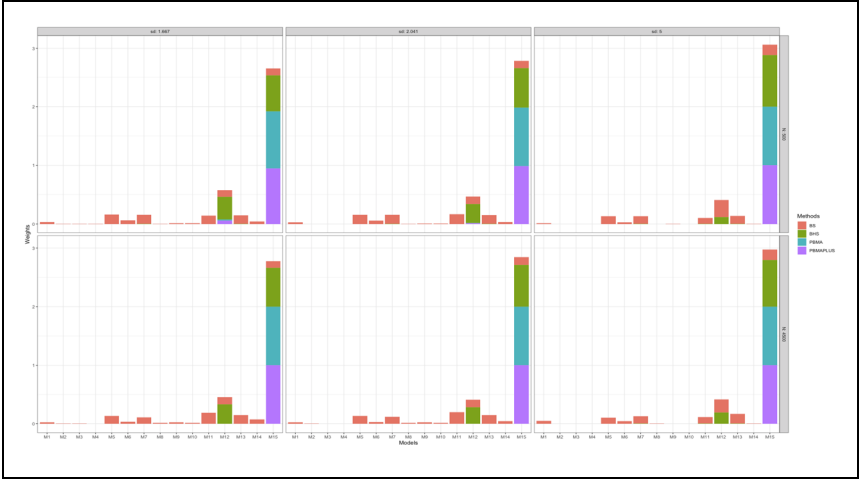


FIGURE 2. Model weights across different sample sizes with different levels of between-group variability in  $\mathcal{M}$ -complete setting.

### Predictive Performance

After examining the difference in model weights, in this section, we compare the predictive performance for different weighting methods in Bayesian stacking. Figures 3 and 4 are the boxplots that summarize the distribution of the KLD obtained from all four Bayesian stacking methods in  $\mathcal{M}$ -closed and  $\mathcal{M}$ -complete settings, respectively. In Figure 3, we can see that there is not much difference in terms of the average KLD for all the methods across sample sizes at different between-group variability levels. All of the KLD is close to zero with small variability, which indicates that BS, BHS, PBMA, and PBMA+ have almost the same but good prediction capacity in the  $\mathcal{M}$ -closed setting. While the KLD obtained from PBMA and PBMA+ have almost no variability, there are some extreme cases in the KLD obtained from BS and BHS. This is more obvious when the between-group standard deviation is equal to 1.667 in the small sample sizes. However, these extreme KLDs are still very small, which does not change our conclusion that the predictive performance of all these four Bayesian stacking methods is identical in  $\mathcal{M}$ -closed setting.

In  $\mathcal{M}$ -complete setting, as shown in Figure 4, BS yields the highest KLD compared to the others in all the samples with different levels of the between-group variability. The standard deviation of the KLD obtained from BS is also the largest. By contrast, BHS, PBMA, and PBMA+ are comparable in terms of predictive capacity. To be precise, as the between-group variability increases

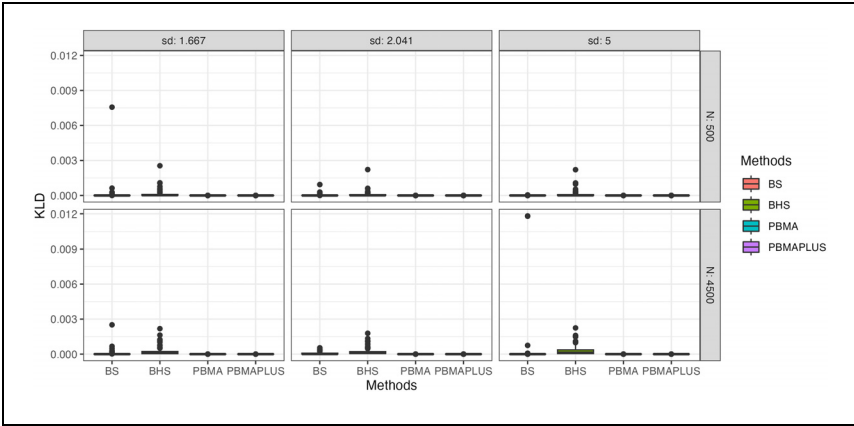


FIGURE 3. Boxplot of Kullback–Leibler divergence across different sample sizes with different levels of between-group variability in  $\mathcal{M}$ -closed setting.

from 1.667 to 2.041, and 5, the variability in KLD obtained by these three methods becomes smaller in both small samples and large samples. In addition, BHS, PBMA, and PBMA+ achieve lower KLD compared to BS. Therefore, we conclude that in the  $\mathcal{M}$ -complete setting, BHS, PBMA, and PBMA+ outperform BS in prediction. In addition, the variability for all the Bayesian stacking methods becomes smaller as the between-group variability increases. Furthermore, Figure 4 does not show a large difference in terms of KLD between the large and small sample sizes. However, this might be due to the small value we set for  $\sigma$ , which is only five in this simulation. Therefore, we anticipate that the variability of KLD will increase if we increase  $\sigma$  for the individual errors. Overall, in the  $\mathcal{M}$ -closed setting, there is not much difference in prediction using all four weighing methods in Bayesian stacking. However, in the  $\mathcal{M}$ -complete setting, BS has the poorest predictive capacity, while the other three are comparable. It is important to note that the calculation of the KLD was based on the average model weights at the sample level. To wit, we did not compare their predictive capacity at the individual level, which is left to be investigated in the future.

## Conclusion and Discussion

This paper examined different weighting schemes in Bayesian stacking for large-scale assessment data with a multilevel structure across three  $\mathcal{M}$  frameworks. Four methods were investigated, including original BS, BHS, PBMA, and PBMA+. Unlike the BS method, which is not fully Bayesian, BHS not

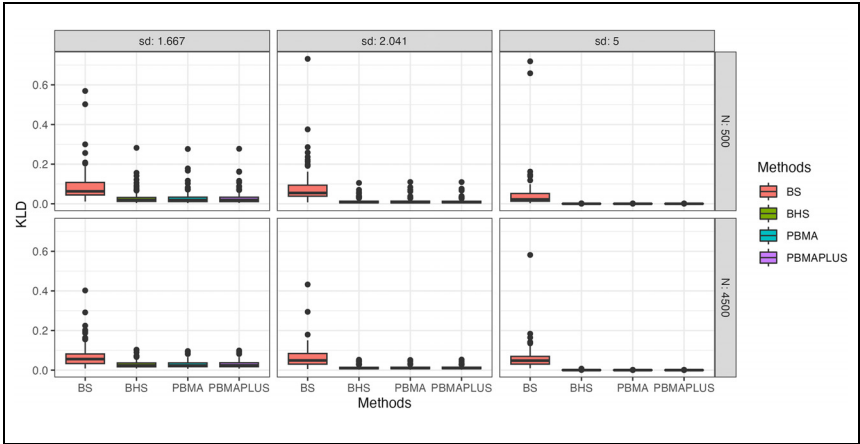


FIGURE 4. *Boxplot of Kullback–Leibler divergence across different sample sizes with different levels of between-group variability in  $\mathcal{M}$ -complete setting.*

only incorporates predictor-dependent weighting but also adds priors and hyper-priors to the weights, which allows for more flexibility in predicting outcomes of interest. For instance, we have shown in the empirical study that BHS assigns distinct model weights to different students while the other three methods assign the same model weights to the entire sample. Broadly speaking, although the predictive accuracy as measured by KLD among the different weighing schemes does not appear to be very large, we still find that their predictive performance varies depending on the underlying  $\mathcal{M}$ -framework (here assessed via the scalar assigned to the error term,  $\sigma$ ) and the between-group variability. Given that BHS, PBMA, and PBMA+ have lower KLD compared to BS, we would recommend one consider using the three methods and try to avoid using BS for the prediction problem in multilevel data.

To be precise, in the  $\mathcal{M}$ -open setting of our empirical example, PBMA obtains the lowest KLD, indicating the best predictive capacity. BHS does not provide optimal predictive performance in the small sample condition. We speculated that this may be due to a sensitivity to the choice of hyperpriors attached to the weight function in small samples and recommended sensitivity analyses. In contrast, BHS and PBMA seem to provide relatively equivalent performance in the large sample condition. Our simulation studies, in contrast, mimic the  $\mathcal{M}$ -closed and  $\mathcal{M}$ -complete setting. In the  $\mathcal{M}$ -closed setting, all of the methods obtain approximately zero KLD and assign more or less 100% weight on the true DGM, as expected, whereas in the  $\mathcal{M}$ -complete setting, the predictive performance of BHS, PBMA, and PBMA+ are comparable while

BS yields the highest KLD. Therefore, we anticipate that as the between-group standard deviation and  $\sigma$  for the Gaussian noise increase, BHS, PBMA, and PBMA+ will demonstrate even better predictive performance in comparison with the other methods. For instance, adding more layers in the multilevel structure, such as a three-level hierarchy, or more complex group relationships, such as crossed random effects, could increase the between-group variability. Another option would be varying  $\sigma$  in the simulation study to examine which weighting methods in Bayesian stacking will obtain the best predictive performance when there are different amounts of noise in the  $\mathcal{M}$ -complete setting. Considering the computation efficiency, in the empirical study, the time used to implement BS, PBMA, and PBMA+ are close to each other. BHS takes a little longer time than the others, which might be due to the large value we set for the iteration (10,000). However, the computation time of BHS is close to PBMA and PBMA+ in the simulation study. Noticing that we use High Throughput Computing (Center for High Throughput Computing, 2006) to implement the simulation study, which optimizes the computation efficiency in general, the algorithmic performance of these four methods will still be a question for future study.

There are several limitations in this study, and thus, there are different directions for future research. Of course, our simulation study is not exhaustive, but we believe that the models and conditions that we have chosen are in line with the applications of multilevel models to large-scale assessments. Nevertheless, extensions of our study may be useful. First, we applied the same hyperpriors that Yao et al. (2021) used in their paper. Therefore, it would be useful to explore how different priors and hyperpriors affect predictive performance using BHS. Second, the value we set to regularize the Gaussian noise for the  $\mathcal{M}$ -complete setting is small and a bit arbitrary. For example, a follow-up study could focus on examining how the predictive performance of these four methods change if we vary the in the future. Third, in this study, we used parametric regression methods to model the weights in BHS. In the future, it may be interesting to use non-parametric methods to compute the weights for different predictors might lead to more flexibility without relying on parametric assumptions. A fourth limitation derives from specific issues associated with the construction of large-scale assessments. Specifically, although the present paper investigated the performance of various stacking weights for multilevel models, the full utility of stacking for large-scale assessments will require incorporating the full complement of plausible values as well as sampling weights. Finally, the data structure we investigated in this study has fixed group memberships, which is not always the case in large-scale assessments. Thus, it will be interesting to explore stacking methods with more complicated data structures of relevance to education research, such as multiple membership models.

To conclude, a number of approaches now exist to address the problem of model uncertainty with applications to the social and behavioral sciences, and although Bayesian model averaging remains a prevalent and often adequate procedure for addressing model uncertainty, it rests on the  $\mathcal{M} - closed$  assumption that analysts may feel uncomfortable holding. Bayesian stacking methods relax this assumption, and a variety of choices for stacking weights are available and easily implemented through open-source software. For this study, we examined a variety of stacking methods for multilevel models applied to large-scale educational assessments. Overall, BHS, PBMA, and PBMA+ have advantages under different conditions and generally have better predictive capacity than BS. We also found Bayesian hierarchical stacking to be a promising approach for calculating stacking weights at the individual levels. It would be interesting to extend the current study, as described above, to investigate different aspects of Bayesian stacking in the future study.

## Appendix

TABLE A1.  
*PISA 2018 Predictors of Reading Scores.*

Variable name	Variable label
FEMALE	Sex (1 = Female)
ESCS	Index of economic, social, and cultural status
METASUM	Meta-cognition: summarizing
PERFEED	Perceived feedback
HOMEPOS	Home possessions
ADAPTIVE	Adaptive instruction
TEACHINT	Perceived teacher's interest
ICTRES	ICT resources
JOYREAD	Joy/Like reading
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
SCREADCOMP	Perception of reading competence
SCREADDIFF	Perception of reading difficulty
PISADIFF	Perception of difficulty of the PISA test
PV1READ	First plausible value reading score

### Acknowledgments

We are grateful to Dr. Sameer Deshpande, Dr. Jee-seom Kim, Ajinkya H. Kokandakar, and Dr. James Pustejovsky for their insightful discussion in this study. This research was supported by the Center For High Throughput Computing (CHTC) in the Department of Computer Science at UWMadison.


### 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) disclosed receipt of the following financial support for the research and/or authorship of this article: Support for M.H. was provided by the University of Wisconsin-Madison Graduate School with fellowship funding.

### ORCID iD

Mingya Huang  <https://orcid.org/0000-0002-0647-7390>

### Notes

1. Gelman and Rubin (1995) have also argued against the use of posterior model probabilities for model selection in the context of Bayes factors.
2. The *widely applicable information criterion* (WAIC) has also been advocated for model selection. Although the WAIC and LOO-CV are asymptotically equivalent (Watanabe, 2010), the implementation of LOO-CV in the loo package is more robust in finite samples with weak priors or influential observations (Vehtari et al., 2017).

### 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*. Akademiai Kiado.
- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. Wiley.
- Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419), 738. <https://doi.org/10.2307/2290212>
- Brozo, W. G., Sulkunen, S., Shiel, G., Garbe, C., Pandian, A., & Valtin, R. (2014). Reading, gender, and engagement: Lessons from five PISA countries. *Journal of Adolescent & Adult Literacy*, 57 (7), 584–593.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.

- Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Studies in Educational Evaluation*, 49, 30–41.
- Center for High Throughput Computing. (2006). *Center for High Throughput Computing*. <https://chtc.cs.wisc.edu/>; <https://doi.org/10.21231/GNT1-HW21>
- Clyde, M. A. (1999). Bayesian model averaging and model search strategies (with discussion). In J. M. Bernardo, A. P. Dawid, J. O. Berger, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 6, pp. 157–185). Oxford University Press.
- Clyde, M. (2003). Model Averaging. In S. J. Press (Ed.), *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, 57, 55–98.
- 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.). Rand Corporation. <https://www.rand.org/pubs/notes/N2683.html> (N-2683-RC)
- 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*. Springer.
- Geisser, S., & Eddy, W. F. (1979a). A predictive approach to model selection. *Journal of the American Statistical Association*, 74 (365), 153–160.
- Geisser, S., & Eddy, W. F. (1979b). A predictive approach to model selection. *Journal of the American Statistical Association*, 74, 153–160.
- Gelfand, A. (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). Chapman & Hall.
- Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131–143). Chapman & Hall.
- Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, 25, 165–173.
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian workflow. *arXiv*. <https://arxiv.org/abs/2011.01808>
- Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Wiley.
- 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. (2022). *rstanarm: Bayesian applied regression modeling via Stan*. <https://mc-stan.org/rstanarm/> (R package version 2.21.3)
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60–87.
- Hoeting, J. A., Madigan, D., Raftery, A., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382–417.
- Kaplan, D. (2021). On the quantification of model uncertainty: A Bayesian perspective. *Psychometrika*, 86(1), 215–238. <https://doi.org/10.1007/s11336-021-09754-5>
- Kaplan, D. (2023). *Bayesian statistics for the social sciences* (2nd ed.). Guilford Press.

- Kaplan, D., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research, 49*, 505–517.
- Kaplan, D., & Huang, M. (2021). Bayesian probabilistic forecasting with large-scale educational trend data: A case study using NAEP. *Large-scale Assessments in Education, 9*(1), 1–31.
- 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, 23*(3), 343–353. <https://doi.org/10.1080/10705511.2015.1092088>
- Kaplan, D., & Lee, C. (2018). Optimizing prediction using Bayesian model averaging: Examples using large-scale educational assessments. *Evaluation Review, 42*(4), 423–457. <https://doi.org/10.1177/0193841X18761421>
- Kaplan, D., & Yavuz, S. (2019). An approach to addressing multiple imputation model uncertainty using Bayesian model averaging. *Multivariate Behavioral Research, 55*(4), 553–567. <https://doi.org/10.1080/00273171.2019.1657790>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.
- Kullback, S. (1959). *Information theory and statistics*. John Wiley and Sons.
- Kullback, S. (1987). The Kullback-Leibler distance. *The American Statistician, 41*, 340–341.
- Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. Wiley.
- Michael, D., & Kyriakides, L. (2023). Mediating effects of motivation and socioeconomic status on reading achievement: A secondary analysis of PISA 2018. *Large-scale Assessments in Education, 11*(1), 31.
- OECD. (2002). *PISA 2000 technical report*. Organization for Economic Cooperation and Development.
- OECD. (2018). *PISA 2018 technical report*. OECD. <https://www.oecd.org/pisa/data/pisa2018technicalreport/>
- R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. <https://www.R-project.org/>
- Raftery, A., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*, 179–191.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications.
- Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics, 9*, 130–134.
- Stan Development Team. (2020). *RStan: The R interface to Stan*. <http://mc-stan.org/> (R package version 2.21.1)
- Stan Development Team. (2021). *Stan modelling language users guide and reference manual v. 2.27*. Stan Development Team Sydney, Australia.
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. <https://mc-stan.org/loo/> (R package version 2.5.1)
- Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. 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>
- 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.
- Yao, Y., Pirš, G., Vehtari, A., & Gelman, A. (2021). Bayesian hierarchical stacking: Some models are (somewhere) useful. *Bayesian Analysis*, 1(1), 1–29.
- Yao, Y., Pirš, G., Vehtari, A., & Gelman, A. (2022). Bayesian hierarchical stacking: Some models are (somewhere) useful. *Bayesian Analysis*, 17 (4), 1043–1071.
- Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018a). Using stacking to average Bayesian predictive distributions Bayesian Analysis. <https://doi.org/10.1214/17-BA1091SUPP>.
- Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018b). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13, 917–1007. <https://doi.org/10.1214/17-BA1091>
- Zhang, D., & Liu, L. (2016). How does ICT use influence students’ achievements in math and science over time? Evidence from PISA 2000 to 2012. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(9), 2431–2449.

### Authors

**MINGYA HUANG** is a Ph.D. Candidate in Educational Psychology (Quantitative Methods) at the University of Wisconsin — Madison. She is pursuing research in Bayesian statistics, machine learning, clustering, and causal inference. Currently, her work focuses on Bayesian nonparametrics and its application in large-scale clustered data.

**DAVID KAPLAN** PhD is Hilldale Professor and Patricia Busk Professor of Quantitative Methods in the Department of Educational Psychology at the University of Wisconsin — Madison. His interests are in Bayesian statistics with applications to large-scale educational assessments.

Manuscript received March 27, 2023

Revision received January 16, 2024

Accepted March 24, 2024